Dipartimento di Economia e Finanza

METODI E ANALISI STATISTICHE

2017

Università degli Studi di Bari Aldo Moro

Dipartimento di Economia e Finanza

METODI E ANALISI STATISTICHE

2017

Università degli Studi di Bari Aldo Moro

Prima edizione: dicembre 2017

ISBN 978-88-6629-013-1

Gli articoli qui presentati sono stati oggetto, oltre che di valutazione interna, anche di revisione anonima (in "doppio cieco").

Editing finale: F. D. d'Ovidio

Sommario

Presentazione

Il Dipartimento di Economia e Finanza (già di Scienze Economiche e Metodi Matematici) cura da molti anni alcune collane di "working paper", sia su temi economici, (SERIES - *Southern Europe Research in Economic Studies*), e sia su altri temi di interesse delle diverse sezioni dipartimentali: la collana *Geografia Economica*, i *Saggi di Storia Economica* e infine gli *Studi di Diritto Pubblico*. Il portale del Dipartimento (http://www.uniba.it/ricerca/dipartimenti/dse/ricerca) riporta i principali riferimenti bibliografici di tali collane, e spesso l'intero testo in PDF per libera consultazione.

A dette collane, si è aggiunta ultimamente una raccolta di studi matematico-statistici curata dalla sezione di Statistica, costola del precedente Dipartimento di Scienze Statistiche "Carlo Cecchi": *Metodi e applicazioni statistiche*, che rappresenta il proseguimento ideale della precedente collana degli *Annali dell'Istituto di Statistica*, creata nel 1927 e ben rinomata nel settore. Tale pubblicazione, per ora esclusivamente in formato PDF *open access*, è ora giunta al terzo anno (comprendendo anche il volume *Studi in ricordo di Carlo Cecchi* pubblicato nel 2015); al momento, il suo scopo è soprattutto quello di diffondere i primi risultati di ricerche in corso, sia di studiosi interni che esterni al Dipartimento, ma in essa trovano luogo anche risultati scientifici molto validi e articolati, a volte troppo ponderosi per trovare altri spazi editoriali.

Questo volume presenta alcuni articoli di matrice chiaramente statistico-metodo–logica (Girone e Nannavecchia, Girone *et al.*, Campobasso *et al.*), altri di matematica applicata a temi economici e finanziari (Diomede e Taglialatela, Bisceglia), altri ancora in cui la metodologia statistica è approfondita ma orientata a uno scopo applicativo (Bilancia e Novembre, d'Ovidio *et al.*), articoli sulla gestione e l'utilizzo di determinate fonti statistiche, anche per la costruzione di indicatori (Marini e Nicolardi, Mancarella e Marastoni, Carbonara). Completano il quadro scientifico del volume studi di statistica economica e aziendale (Carucci *et al.,* Summo), su temi socio-demografici (Carella *et al.*), sulle previsioni finanziarie con reti neurali artificiali (Firza e Monaco), un articolo di statistica storica su basi economiche (Ricci), un articolato studio sui dati INVALSI (Mariella e Tarantino) e una nota di farmaco-economia (d'Ovidio e Viola).

Alcuni di questi articoli possono essere considerati un lavoro compiuto e rifinito, altri sono "work in progress", ma tutti affrontano con competenza temi di interesse. Sperando che il volume incontri il favore dei lettori, colgo l'occasione per ringraziare tutti coloro che vi hanno contribuito, a partire dagli Autori ma senza dimenticare i *referees* esterni che hanno fortemente contribuito alla qualità degli scritti qui proposti.

Bari, 13/12/2017

Il Direttore del Dipartimento di Economia e Finanza
Ernesto Toma

UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO
Metodi e analisi statistiche 2017
DIPARTIMENTO DI ECONOMIA E FINANZA
ISBN: 978-88-6629-013-1
pp. 7-20

Influence function of the Gini's index of cograduation

Giovanni Girone[1], Antonella Nannavecchia[2]*

[1]*Università degli Studi di Bari "Aldo Moro",*
[2]*Università Lum Jean Monnet, Casamassima, Bari*

Abstract: The robustness of an estimator, meaning to be resistant to contaminations of the model distribution, can be evaluated studying its influence function; from influence function it is possible to evaluate other aspects of the estimator as its gross-error sensitivity, which is the maximal influence an observation may have, and its asymptotic variance. The robustness of the main statistical measures of correlation were studied by Croux and Dehon (2010). The indexes they considered were Pearson's correlation, Quadrant correlation, Kendall's correlation and Spearman's rank correlation. In this paper we study the robustness of the Gini's index of cograduation which presents great difficulties in order to obtain similar results because of the complexity of its expression. We derive the influence functions and the gross-error sensitivities of the above index. The Gini's index of cograduation has results similar to those of Quadrant, Kendall and Spearman correlation indexes overall and a better performance for some aspects.

Keywords: Gini's index of cograduation.; robustness; influence function; gross-error sensitivity.

1. Introduction

Croux and Dehon (2010) studied the robustness of the main statistical measures of correlation. The indexes they considered were Pearson's correlation, Spearman's rank correlation (Spearman 1904), Kendall's rank correlation (Kendall 1938) and

* Correspondent author: nannavecchia.a@gmail.com .
The paper is the result of a joint research but paragraphs 1, 2, 3 and 4 are due to G. Girone, while paragraphs 5, 6, 7, 8, and 9 are due to A. Nannavecchia.

Quadrant correlation (Mosteller 1946). For each of those indexes they found the functional representation of the estimators, the Fisher consistent version of the functional at the bivariate normal distribution with correlation coefficient ρ, the Hampel influence function (Hampel et al. 1986), the gross-error sensitivity index (*GES*), which is the maximal influence an observation may have, and the asymptotic variance. Croux and Dehon did not consider the Gini's index of cograduation G (Gini 1916 and 1954). The reason for this lack may be due to the low use of this index by non-Italian statisticians and probably to the great difficulty to obtain similar results because of the complexity of the G index expression.

The aim of this note is to achieve for the G index results similar to those obtained for the above mentioned indexes, considering the important features of the G index studied by many authors and, particularly, by Salvemini (1951) and Amato (1954). Some good properties of Gini's index have been recently investigateded by Genest and al (2010).

2. The Gini's index of cograduation: definition and functional

Let (x_i,y_i), $i = 1,2,\ldots,n$, be a bivariate sample according to two given criteria. Let r_i and s_i, $i = 1,2,\ldots,n$, be the sequences of ranks of the sample units of the two criteria. A measure of rank correlation is given by Gini's cograduation index

$$r_G = \frac{\sum_{i=1}^{n}|n + 1 - r_i - s_i| - \sum_{i=1}^{n}|r_i - s_i|}{K}$$

where the constant K equals $n^2/2$ when n is even, and $(n^2 - 1)/2$ when n is odd. The index value is 1 if the rankings based on the two criteria are perfectly concordant and -1 if the rankings are perfectly discordant. Gini's cograduation index is a particular case of a large class of rank statistics proposed by Cifarelli, Conti and Regazzini (1996).

Let $(X,Y) \sim H$, where H is a generic bivariate distribution having second moments and $F(t) = P_H(X \leq t)$, $G(t) = P_H(Y \leq t)$ are the marginal cumulative distribution functions of X and Y. The functional associated with Gini's cograduation index for the population is given by

$$R_G(H) = 2E_H[|1 - F(X) - G(Y)| - |F(X) - G(Y)|]. \qquad (1)$$

An alternative expression, obtained by considering the signs of the terms in absolute value, is given by

$$R_G(H) = 2E_H\{[(1 - F(X) - G(Y)]sign[1 - F(X) - G(Y)] + \\ -[F(X) - G(Y)]\text{sign}[F(X) - G(Y)]\}. \quad (2)$$

Combining the two signs in (2) the following formula may be derived

$$R_G(H) = 4E_H\{F(X)|F(X) < G(Y) < 1 - F(X)\} + \\ +4E_H\{G(Y)|G(Y) < F(X) < 1 - G(Y)\}. \quad (3)$$

If in (3) the marginal distribution functions are equal, that is if $F = G$, then the two terms in the right side are equal as well, and therefore the expression of Gini's cograduation index becomes

$$R_G(H) = 8E_H\{F(X)|F(X) < G(Y) < 1 - F(X)\}. \quad (4)$$

If the two marginal distribution functions are symmetric with respect to zero, (3) and (4) can be reduced as follows

$$R_G(H) = 4E_H\{F(X)|F(X) < G(Y) < F(-X)\} + \\ +4E_H\{G(Y)|G(Y) < F(X) < G(-Y)\}, \quad (5)$$

$$R_G(H) = 8E_H\{F(X)|F(X) < G(Y) < F(-X)\}. \quad (6)$$

Since the cumulative distribution functions are not decreasing, (5) and (6) can be written in terms of the arguments

$$R_G(H) = 4E_H\{F(X)|X < Y < -X\}+4E_H\{G(Y)|Y < X < -Y\},$$

$$R_G(H) = 8E_H\{F(X)|X < Y < -X\}.$$

3. The Gini's index of cograduation for the bivariate normal distribution

Let H be a bivariate normal distribution, denoted by Φ. Since correlation's measures, and therefore G also, are invariant with respect to standardizations, without loss of generality, we can assume ϕ to have means 0, standard deviations 1 and correlation coefficient ρ.

The functional representation of Gini's index for the bivariate normal standard population is given by

$$R_G(\phi_\rho) = 2E_\rho[|1 - \Phi(X) - \Phi(Y)| - |\Phi(X) - \Phi(Y)|], \quad (7)$$

where $\Phi(X)$ is the cumulative distribution function (CDF) of a standard normal. Since Φ depends only on ρ, henceforth, $R_G(\Phi_\rho)$ will be denoted by $R_G(\rho)$, which is a function only of the unknown parameter ρ.

The expectation of (7) is not easy to compute, first of all for the presence of absolute values, then because the CDF of a normal distribution is not known in its explicit form and finally for the complexity of the bivariate normal distribution. Nevertheless, through several analytical tricks applied in computing integrals, we obtained three equivalent expressions to evaluate it.

The first formula, which expresses $R_G(\rho)$ as a function of ρ, is given by

$$R_G(\rho) = \frac{4}{\sqrt{2\pi}}\int_0^\infty e^{-\frac{x^2}{2}}\mathrm{Erf}\left(\frac{x}{\sqrt{2}}\right)\left\{\mathrm{Erf}\left[x\sqrt{\frac{1+\rho}{2(1-\rho)}}\right] - \mathrm{Erf}\left[x\sqrt{\frac{1-\rho}{2(1+\rho)}}\right]\right\}dx, \quad (8)$$

the second formula can be derived by integrating by parts the expression (8)

$$R_G(\rho) = \sqrt{\frac{2}{\pi}}\int_0^\infty \left[\sqrt{\frac{1-\rho}{1+\rho}}e^{-\frac{x^2(1-\rho)}{1+\rho}} - \sqrt{\frac{1+\rho}{1-\rho}}e^{-\frac{x^2(1+\rho)}{1-\rho}}\right]\left(\mathrm{Erf}\left(\frac{x}{\sqrt{2}}\right)\right)^2 dx, \quad (9)$$

the third formula can be obtained by expanding the expression in braces of (8)

$$R_G(\rho) = \frac{4}{\pi}\int_0^\infty e^{-\frac{x^2}{2}}\mathrm{Erf}\left(\frac{x}{\sqrt{2}}\right)\sum_{k=0}^\infty \frac{(-1)^k\left(\sqrt{\frac{1+\rho}{1-\rho}}^{2k+1} - \sqrt{\frac{1-\rho}{1+\rho}}^{2k+1}\right)x^{2k+1}}{2^k(2k+1)k!}dx. \quad (10)$$

In formulas (8), (9) and (10), Erf(u) is the error function

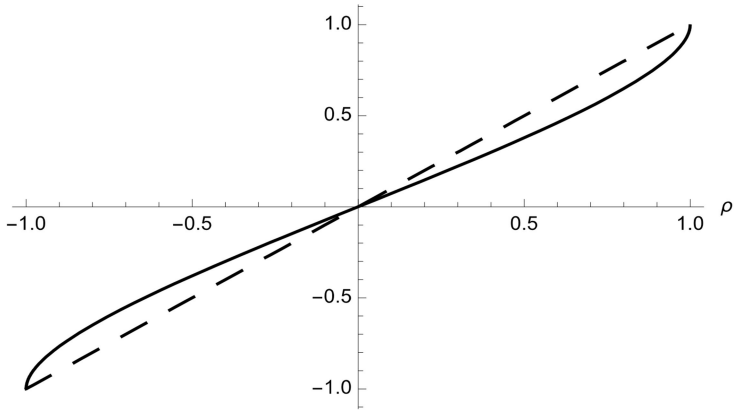
$$\mathrm{Erf}(u) = \frac{2}{\pi}\int_0^u e^{-t^2}dt.$$

The following Figure 1 represents $R_G(\rho)$ as a function of ρ from -1 to 1.

As it is easily seen in Fig. 1, the estimator $R_G(\rho)$ is equal to ρ only at $\rho = -1$, $\rho = 0$ and $\rho = 1$; $R_G(\rho)$ overestimates ρ at $-1 < \rho < 0$ and $R_G(\rho)$ underestimates ρ at $0 < \rho < 1$.

A Fisher consistent version of the Gini's index request to invert any one of the the $R_G(\rho)$ expressions, that is $\rho = R^{-1}(G)$.

Figure 1. *Gini's index of cograduation $R_G(\rho)$, given the bivariate standard normal distribution with parameter ρ as a model.*



4. An approximation of $R_G(\rho)$

Since it is not possible to invert one of the three expressions of the Gini's index of cograduation when the bivariate distribution is standard normal, we obtained an approximation of such an inversion. After many attempts we verified that $R_G(\rho)$ can be satisfactorily approximated by a linear combination of two arcsines

$$R_G(\rho) \simeq \left[1.4\ arcsin(\rho) + 1.8 \arcsin\left(\frac{\rho}{2}\right)\right]/\pi\ . \tag{11}$$

For a bivariate normal population, this result equals the weighted average of Kendall's and Spearman's transformed indexes. The following figures show the goodness of the approximation.

Figure 2. *Exact and approximated function of $R_G(\rho)$.*

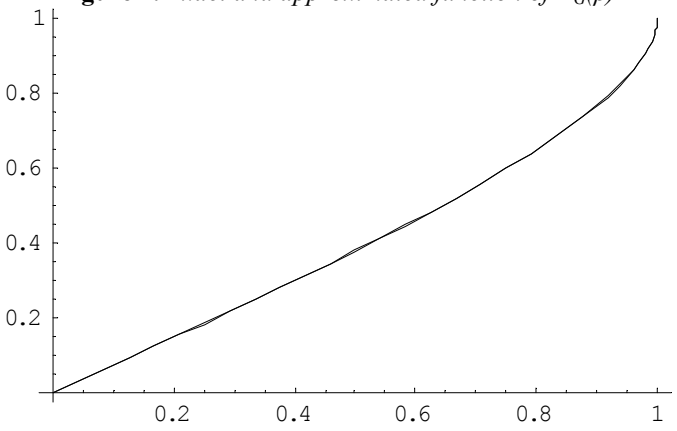


Figure 3. *Difference between the exact and the approximated function of $R_G(\rho)$.*

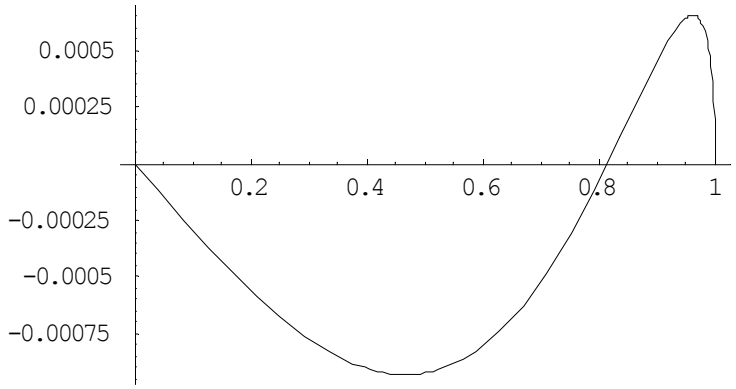


Figure 2 represents both the exact and the approximate function of $R_G(\rho)$, for $0 < \rho < 1$: the two curves almost coincide. Figure 3 shows the difference between the two functions: the maximum difference is less than a thousandth.

The approximation of $R_G(\rho)$ can not be inverted analytically, but inverting its expanded series we derived a good approximation

$$\begin{aligned}\rho \simeq\ & 1.36591g - 0.3000082g^3 - 0.027998g^5 - 0.0204595g^7 - 0.0096845g^9 + \\ & -0.00441897g^{11} - 0.0019345g^{13} - 0.00814741g^{15} - 0.000329198g^{17} + \\ & -0.00012671g^{19} - 0.0000453999g^{21} - 0.0000144617g^{23} + \\ & -0.00000359456g^{25} - 0.00000227507g^{27} - 0.000000542652g^{29}. \end{aligned} \quad (12)$$

In (12) for simplicity $g = R_G(\rho)$.

The expression (12) allows to obtain an approximate Fisher consistent estimator of the parameterρof a bivariate standard normal distribution from the Gini's index of cograduation. In fact as the inverse function is symmetric with respect to the principal diagonal of a function, it is also a good approximation of the inverse function of $R_G(H)$.

Alternatively, given the value g of $R_G(\rho)$ it is possible to obtain numerically the estimate by inverting $R_G(\rho)$.

5. The influence function of $R_G(\rho)$

Gini's index of cograduation, given a bivariate distribution H, is expressed by

$$R_G(H) = 2E_H[|1 - F(X) - G(Y)| - |F(X) - G(Y)|].$$

Let us suppose that the distribution $H(X,Y)$ of the population has an infinitesimal amount ε of contamination placed at the point (x,y); then the contaminated model distribution can be defined as

$$H_\varepsilon(X,Y) = (1-\varepsilon)H(X,Y) + \varepsilon\Delta_{(x,y)}$$

with the contaminated marginal cumulative distribution functions expressed by

$$F_\varepsilon(X) = (1-\varepsilon)F(X) + \varepsilon\Delta_x$$

and

$$G_\varepsilon(Y) = (1-\varepsilon)G(Y) + \varepsilon\Delta_y \ .$$

The Gini's index of cograduation at the contaminated population is given by

$$R_G(H_\varepsilon) = 2E_{H_\varepsilon}[|1 - F_\varepsilon(X) - G_\varepsilon(Y)| - |F_\varepsilon(X) - G_\varepsilon(Y)|]. \qquad (13)$$

By substituting in (13) the expressions of the contaminated model (H_ε) and of the marginal cumulative distribution functions ($F_\varepsilon, G_\varepsilon$), with simple algebraic steps, we obtain

$$\begin{aligned} R_G(H_\varepsilon) = 2E_H\{&|1-(1-\varepsilon)[F(X)+G(Y)] - \varepsilon[I(X>x)+I(Y>y)]| + \\ &-|(1-\varepsilon)[\mathrm{F(X)} - \mathrm{G(Y)}] + \varepsilon[\mathrm{I(X}>x) - \mathrm{I(Y}>y)]|\} + \\ -2\varepsilon E_H\{&|1-(1-\varepsilon)[F(X)+G(Y)] - \varepsilon[I(X>x)+I(Y>y)]| + \\ &-|(1-\varepsilon)[\mathrm{F(X)} - \mathrm{G(Y)}] + \varepsilon[\mathrm{I(X}>x) - \mathrm{I(Y}>y)]|\} + \\ +2\varepsilon\{&|1-(1-\varepsilon)[\mathrm{F(x)} + \mathrm{G(y)}] - \varepsilon[\mathrm{I(X}>x) + \mathrm{I(Y}>y)]| + \\ &-|(1-\varepsilon)[\mathrm{F(x)} - \mathrm{G(y)}] + \varepsilon[\mathrm{I(X}>x) - \mathrm{I(Y}>y)]|\} \end{aligned}$$

where $I(.)$ is the indicator function.

By differentiating the previous expression with respect to ε and putting $\varepsilon = 0$, after several simplifications, we derived the influence function of the Gini's index of cograduation

$$\begin{aligned} IF[(x,y), R_G(H)] = &-R_G(H) + 2E_H\{[F(X)+G(y)-I(X>x)-I(Y>y)]\cdot \\ &\cdot sign[1-F(X)-G(Y)] + [(F(X)-G(Y)-I(X>x)+I(Y>y)]\cdot \\ &\cdot sign[F(X)-G(Y)]\} + 2[|1-F(x)-G(y)| - |F(x)-G(y)|] \ , \end{aligned}$$

and since (see Sect. 2)

$$\begin{aligned} 2E_H\{&[F(X)+G(y)]\cdot sign[1-F(X)-G(Y)] + \\ &+[(F(X)-G(Y)]\cdot sign[F(X)-G(Y)]\} = -R_G(H), \end{aligned}$$

the influence function can be reduced as

$$IF[(x,y),R_G(H)] = 2E_H\{[-I(X>x)-I(Y>y)]\cdot sign[1-F(X)-G(Y)]+ \\ +[-\mathrm{I}(X>x)+\mathrm{I}(Y>y)]\cdot sign[F(X)-G(Y)]\} + \\ +2[|1-F(x)-G(y)|-|F(x)-G(y)|]-2R_G(H). \quad (14)$$

6. The influence function of the Gini's index of cograduation at the bivariate normal distribution

We assume that the model is given by a normal standard bivariate distribution with means 0 and variances 1

$$\phi(x,y) = \frac{e^{-\frac{(x^2-2\rho xy\quad ^2)}{2(1-\rho^2)}}}{2\pi\sqrt{1-\rho^2}}.$$

The marginal distribution functions are both standard normal and the cumulative distribution functions are $F = G = \Phi$. The symmetry of the normal model allows to simplify the calculation of the influence function.

The first two terms on the right side of the general expression of the influence function (14), derived in Section 5, because of the symmetry expressed by $1 - \Phi(X) = \Phi(-X)$, become

$$2E_\rho\{[-I(X>x)-I(Y>y)]\cdot sign[\Phi(-X)-\Phi(Y)]+ \\ +[-I(X>x)+I(Y>y)]\cdot sign[\Phi(X)-\Phi(Y)]\}, \quad (15)$$

and, since the cumulative distribution functions are not decreasing , is

$$2E_\rho\{[-I(X>x)-I(Y>y)]\cdot sign(-X-Y) \quad (16)$$

The first term in (16) becomes

$$2E_\rho\{I(X>x)\cdot[sign(X+Y)+sign(Y-X)]+ \\ +I(Y>y)\cdot[sign(X+Y)+sign(X-Y)]\}. \quad (17)$$

Since the two terms in (17) have an analogous structure, we calculate only one of them

$$A(u) = 2E_\rho\{I(X > u)[sign(X + Y) + sign(Y - X)]\}. \tag{18}$$

Combining the two signs in (18) we obtained the four quadrants

$$(X + Y > 0) \cap (Y - X > 0) \equiv (Y > 0) \cap (-Y < X < Y),$$
$$(X + Y > 0) \cap (Y - X < 0) \equiv (X > 0) \cap (-X < Y < X),$$
$$(X + Y < 0) \cap (Y - X > 0) \equiv (X < 0) \cap (X < Y < -X),$$
$$(X + Y < 0) \cap (Y - X < 0) \equiv (Y < 0) \cap (Y < X < -Y).$$

The sums of the two signs are, respectively, equal to 2, 0, 0 and −2. Consequently, the value of (18) for $u > 0$ is given by

$$A(u) = 4\int_u^\infty \int_x^\infty \phi(x,y)dydx - 4\int_u^\infty \int_{-\infty}^{-x} \phi(x,y)dydx \tag{19}$$

and for $u < 0$ is given by

$$\begin{aligned} A(u) = {} & 4\int_u^0 \int_{-x}^\infty \phi(x,y)dydx + 4\int_0^\infty \int_x^\infty \phi(x,y)dydx + \\ & -4\int_u^0 \int_{-\infty}^x \phi(x,y)dydx - 4\int_0^\infty \int_{-\infty}^{-x} \phi(x,y)dydx. \end{aligned} \tag{20}$$

Since expressions (19) and (20) are equal, as proved by some algebraic calculations, we only consider the first one.

In the same way the second term of (17) becomes

$$\begin{aligned} A(v) &= 2E_\rho\{I(Y > v) \cdot [sign(X + Y) + sign(X - Y)]\} = \\ &= 4\int_v^\infty \int_y^\infty \phi(x,y)dxdy - 4\int_v^\infty \int_{-\infty}^{-y} \phi(x,y)dxdy. \end{aligned} \tag{21}$$

Solving the second integrals we obtain

$$A(u) = 2\int_u^\infty \frac{e^{-\frac{x^2}{2}\left\{Erf\left[x\sqrt{\frac{1+\rho}{2(1-\rho)}}\right] - Erf\left[x\sqrt{\frac{1-\rho}{2(1+\rho)}}\right]\right\}}}{\sqrt{2\pi}} dx, \tag{22}$$

$$A(v) = 2\int_v^\infty \frac{e^{-\frac{y^2}{2}\left\{Erf\left[y\sqrt{\frac{1+\rho}{2(1-\rho)}}\right] - Erf\left[y\sqrt{\frac{1-\rho}{2(1+\rho)}}\right]\right\}}}{\sqrt{2\pi}} dy. \tag{23}$$

Henceforth, the variables u and v will be denoted by x and y.

The last two terms of the influence function (14) are the function of x and y

$$B(x,y) = 2[|1-\Phi(x)-\Phi(y)| - |\Phi(x)-\Phi(y)|] = \\ = \left|Erf\left(\frac{x}{\sqrt{2}}\right) + Erf\left(\frac{y}{\sqrt{2}}\right)\right| - \left|Erf\left(\frac{x}{\sqrt{2}}\right) - Erf\left(\frac{y}{\sqrt{2}}\right)\right| \quad (24)$$

and the constant $-2\,R_G(\rho)$.

The influence function of the Gini's index of cograduation for the bivariate normal distribution can be, consequently, expressed as

$$IF[(x,y), R_G(\rho)] = A(x) + A(y) + B(x,y) - 2R_G(\rho). \quad (25)$$

The expected value of each of the two functions A(x)+A(y) and B(x,y) is equal to $R_G(\rho)$ and so $E_\rho\{IF[(x,y), R_G(\rho)]\} = 0$.

7. The influence function of the Fisher consistent Gini's index of cograduation

The Gini's index of cograduation is not a Fisher consistent estimator of ρ (see Sect. 3) and it is not possible to obtain the analytic expression of the inverse function of $R_G(\rho)=g(\rho)$. Nevertheless it is possible to derive the influence function of the Fisher consistent Gini's index.

In general, if IF[(x,y), $R(\rho)$] is the influence function of a not Fisher consistent estimator and $\bar{R}_G(\rho) = R^{-1}[G]\big(R_G(\rho)\big)$ is the Fisher consistent estimator of ρ, the influence function of the Fisher consistent estimator of ρ can be obtained by the product between the influence function of the not Fisher consistent estimator of ρ and the derivative of the inverse estimator.

The derivative of the inverse Gini's index of cograduation, given a normal bivariate model, is

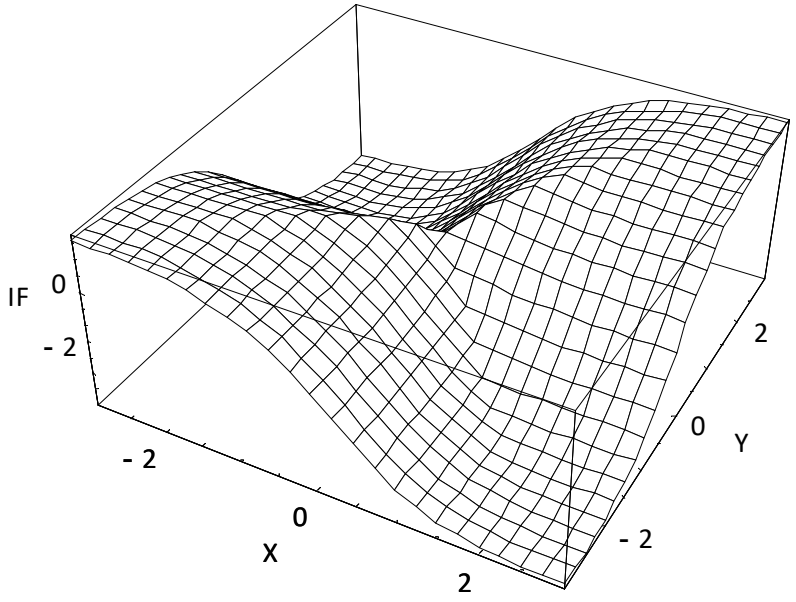
$$\frac{dR^{-1}(\rho)}{d\rho}\, g^{-1}[R_G(\rho)] = \frac{\pi\sqrt{(9-\rho^2)(1-\rho^2)}}{2(\sqrt{(3-\rho)(1+\rho)}+\sqrt{(3+\rho)(1-\rho)})},$$

hence

$$IF[(x,y), \bar{R}_G(\rho)] = \frac{\pi\sqrt{(9-\rho^2)(1-\rho^2)}}{2(\sqrt{(3-\rho)(1+\rho)}+\sqrt{(3+\rho)(1-\rho)})}\, IF[(x,y), R_G(\rho)]. \quad (26)$$

Figure 4 represents the influence function of the Fisher consistent Gini's index of cograduation $\bar{R}_G(\rho)$ for $\rho = 1/2$. This function is similar to those of Mosteller's, Kendall's and Spearman's indexes, but it is smoother.

Figure 4. *Influence function of the Fisher consistent Gini's index of cograduation for ρ=½.*



8. Gross-error sensitivity

The gross-error sensitivity (*GES*) is the maximal influence, in absolute value, an observation may have on the value of an index (Hampel et al. 1986). The *GES* for the Fisher consistent Gini's index of cograduation is given by

$$GES\left[\bar{R}_G(\rho), \phi_\rho\right] = \sup_\rho \left| IF\left[(x,y), \bar{R}_G(\rho), \phi_\rho\right]\right| =$$

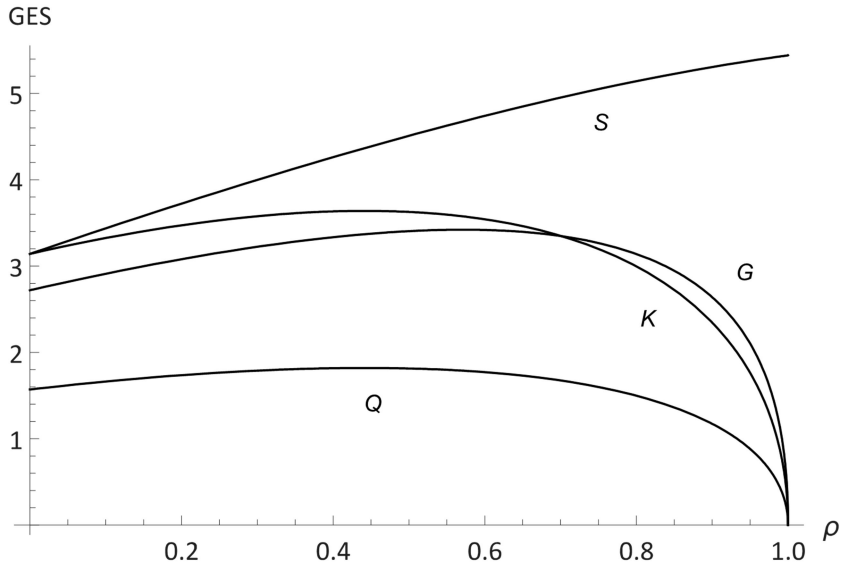
$$= \frac{\pi\sqrt{(9-\rho^2)(1-\rho^2)}}{2(\sqrt{(3-\rho)(1+\rho)}+\sqrt{(3+\rho)(1-\rho)})}\left[2 + |2R_G(\rho)|\right]. \qquad (27)$$

The previous result can be obtained for positive values of ρ, and consequently of $R_G(\rho)$, tending x to ∞ and y to $-\infty$ (or similarly tending x to $-\infty$ and y to ∞), and for negative values of ρ, and consequently of $R_G(\rho)$, tending both x and y to $-\infty$ or both x and y to ∞.

Figure 5 represents the *GES* of Mosteller's, Kendall's, Spearman's and Gini's indexes as a function of the correlation parameter ρ.

With regard to the aspect of the maximal influence of a contamination, Mosteller's index seems to be the best one, Spearman's the worst; Gini's and Kendall's indexes have an intermediate position, the first is better for $\rho < 0.701$ and the second for $\rho > 0.701$.

Figure 5. *Gross-error sensitivities of Mosteller's (Q), Kendall's (K), Spearman's (S) and Gini's (G) indexes as a function of the correlation ρ given a bivariate normal distribution as a model.*



9. Conclusion

In this note we derived the functional of the Gini's index of cograduation both for a generic bivariate distribution and for a normal bivariate distribution. We proved that Gini's index is not a Fisher consistent estimator of the parameter ρ of a normal bivariate distribution.

Since it is not possible to derive the analytic expression of the inverse Gini's index, a Fisher consistent estimate of the parameter ρ can be always obtained by using numerical techniques for inverting the index as function of ρ.

Then we obtained the influence function of the above index both for a generic bivariate distribution and for a normal bivariate distribution; the last result has been derived both for Fisher consistent and not Fisher consistent indexes.

Moreover we obtained the expression of the maximal effect, in absolute value, of a contamination on the value of the Gini's index. These results allows to evaluate the Gini's index in terms of consistency, of robustness with respect to contaminations of the model distribution.

With regard to the other indexes (Pearson, Mosteller, Kendall and Spearman) the Gini's index seems to have similar results overall and a better performance for some aspects.

References

Amato, V. (1954). Sulla distribuzione dell'indice di cograduazione del Gini. *Stati–stica*, 3: 505-519.

Cifarelli, D.M.; Conti, P.L.; Regazzini, E. (1996). On the asymptotic distribution of a general measure of monotone dependence. *The Annals of Statistics*, Vol. 24, No. 3: 1386-1399.

Croux, C.; Dehon, C. (2010). Influence functions of the Spearman and Kendall correlation measures. *Stat Methods Appl*, 19: 497-515.

Conti, P.L.; Nykitin, Y.Y. (2002). Rates of convergence for a class of rank tests for independence, *Journal of Mathematical Sciences*, vol 109, Issue 6, pag 2141-2147.

Cucconi, O. (1964). Sulla distribuzione dell'indice di cograduazione di Gini. *Statistica*, 24: 143-151.

Genest, C.; Nesleehova, J.; Ghorbal, N. (2010). Spearman's footrule and Gini's gamma: a review with complements, *Journal of Nonparametric Statistics*, 22, 8: 937-954.

Gini, C. (1916). Indici di concordanza. *Atti del Reale Istituto Veneto di Scienze, Lettere ed Arti*, 75-2, Venezia.

Gini, C. (1954). *Di una misura delle relazioni tra le graduatorie di due caratteri*. Saggi monografici del comune di Roma, Tip. Cecchini, Roma.

Girone, G.; Montrone, S.; Leogrande, D. (2010). La distribuzione campionaria dell'indice di cograduazione di Gini per dimensioni campionarie fino a 24. *Annali del Dipartimento di Scienze Statistiche*, Università degli Studi Bari, 9: 245-271.

Hampel, F.R.; Ronchetti, E.M.; Rousseeuw, P.J.; Stahel, W.A. (1986). *Robust statistics: the approach based on influence functions*. Wiley, New York.

Kendall, M.G. (1938). A new measure of rank correlation. *Biometrika*, 30: 81-93.

Mosteller, F. (1946). On some useful inefficient statistics. *Ann Math Stat*, 17: 377-408.

Spearman, C. (1904). General intelligence objectively determined and measured. *Am J Psycol*, 15: 201-293.

Salvemini, T. (1951). Sui vari indici di cograduazione. *Statistica*, 2: 133-154.

UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO
Metodi e analisi statistiche 2017
DIPARTIMENTO DI ECONOMIA E FINANZA
ISBN: 978-88-6629-013-1
pp. 21-33

The mean difference of selected continuous and discrete distributions

Giovanni Girone, Antonella Massari, Dante Mazzitelli, Francesco Campobasso, Angela Maria D'Uggento*, Fabio Manca, Claudia Marin

Università degli Studi di Bari Aldo Moro

Abstract:a paper on Gini mean difference (Yitzhaki, 2003) shows the superiority of mean difference as a measure of variability for non normal distributions and contains a survey of the main past and recent scholarly contributions on the topic. The two previous gaps of the mean difference (difficulty of calculation and lack of inferential results) appear overcome with the mean difference presently having a substantially basic knowledge comparable with that of other variability indexes, thus offering the prospect of more widespread use. This perspective requires a greater attention to the ability of the mean difference and other variability indexes to characterize distributions. Apart from its use in measuring the variability of a series of observations, mean difference can also be used to measure the distribution variability. Knowledge of the formulas of the mean difference of distributions, and particularly their relations with the parameters, is an unavoidable step towards an increase in characterization of distributions provided by other variability indexes. The aim of this paper is to obtain explicit formulas of the mean difference, if possible in compact form, for some continuous and discrete distributions.

Keywords: mean difference; variability indexes; continuous distributions; discrete distributions.

* Corresponding Author: angelamaria.duggento@uniba.it

Attributions: G. Girone wrote Sections 1 and 5; A. Massari Section 2 and part of Section 3 (3.7, 3.8, 3.9, 3.10); D. Mazzitelli part of Section 3 (3.1, 3.2, 3.3); F. Campobasso part of section 3 (3.4, 3.5, 3.6); A. M. D'Uggento subsection 3.11 and part of Section 4 (4.3, 4.4, 4.5, 4.6); C. Marin part of Section 4 (4.1, 4.2); F. Manca part of Section 4 (4.7, 4.8, 4.9).

1. Introduction

More than a century ago (Gini, 1912), the mean difference was suggested as a variability measure along with other indexes, in particular the standard deviation and the mean deviation, for the measure of variability, no longer understood as spread of observations around a location index, but rather as a measure of the inequality among the observations. Since then, there have been contributions from a large number of scholars, initially with the aim of proposing ingenious methods of calculation, then to study the relations with other indexes and suggest generalizations, and finally to study the sampling behavior of the mean difference under various conditions.

Apart from its use as a means of measuring the variability of a series of observations, like other variability indexes, the mean difference can also be used to measure the variability of distributions. For the most widespread (normal) or simple (exponential and uniform) continuous distributions, these studies produced some results which are to be found in works of statistical theory (Stuart, Ord, 1994). Knowledge of the mean difference of distributions contributes to the integration of their characterization provided by other variability indexes. Unfortunately, the most common handbooks of distributions (Johnson, Kotz, Balakrishnan, 1994 and 1995; Johnson, Kemp, Kotz, 2007; Patel, Kapadia, Owen, 1976) make no reference to the mean difference of the various distributions.

Two of the present authors (Girone, Mazzitelli, 2007) have carried out an initial systematic study of the mean difference, giving compact expressions of the mean difference for a number of continuous distributions.

2. Materials and methods

The aim of this paper is to obtain explicit formulas of the mean difference, if possible in compact form, for other continuous and discrete distributions. The formulas are obtained by particularizing general formulas of the mean difference of a distribution to single distributions. This work aims to solve difficult double integrals or double sums, a task initially accomplished with the help of the software Mathematica, after which hard simplifications are used to obtain expressions in a closed form. Main results are shown in the following chapters.

Knowledge of these formulas is necessary to enlarge the characterization of distributions through the relation of the mean difference and other variability indexes to their parameters. Although the formulas of this work have already been success-

fully used to analyze the relation between mean difference and standard deviation and between mean difference and mean deviation for many continuous distributions (Girone, Massari, Manca, 2017; D'Uggento, Girone, Marin, 2017), a number of other utilizations are possible, in order to enlarge the characterization of the distributions.

2.1 General formulas for the calculation of the mean difference

Continuous distribution is defined by the density function *f(x),* or the cumulative distribution function *F(x)*, or the first incomplete moment $F_1(x)$.

Discrete distribution is defined by the probabilities p_i or by the cumulative probabilities P_i.

The aforementioned functions can contain one or more parameters which may be location, scale and shape parameters. The mean difference is independent of the location parameter, homogeneous with respect to scale parameter and dependent on shape parameters. Consequently, for the sake of simplicity, the mean difference of a distribution can be calculated considering the distribution with unit scale parameter and multiplying the result by the value of this parameter.

To calculate the mean difference of a continuous distribution (Girone, Maz–zitelli, 2007) the direct formula may be used

$$\Delta = 2\int_{-\infty}^{+\infty}\int_{-\infty}^{x}(x-y)f(x)f(y)dydx, \qquad [1]$$

or the following formula based on the first incomplete moment

$$\Delta = 2\int_{-\infty}^{+\infty}[xF(x)-F_1(x)]f(x)dx, \qquad [2]$$

or the following formula based on the distribution function

$$\Delta = 2\int_{-\infty}^{+\infty}x[2F(x)-1]f(x)dx, \qquad [3]$$

or, finally, the following formula based on the distribution function and its complement at 1

$$\Delta = 2\int_{-\infty}^{+\infty}F(x)[1-F(x)]dx. \qquad [4]$$

The above formulas are exactly equivalent even though, according to the specific distribution, their calculation can be facilitated by some of them. By means

of their application, compact formulas of the mean differences were obtained for the following 14 distributions: normal, rectangular, exponential, Laplace, Weibull, Pareto, power, triangular, logistic, gamma, beta, Type 1 Gumbel, chi-square and Student (Girone, Mazzitelli 2007). For the Cauchy distribution the mean difference does not exist.

In paragraph 3 of this note, the formulas of the mean difference are obtained for the following 11 additional continuous distributions: parabolic, chi, semi-normal, Rayleigh, Maxwell-Boltzman, Dagum, Type-2 Gumbel, Type-3 Gumbel, Snedecor, correlation coefficient distribution in the bivariate normal with independent components and inverse normal.

In order to calculate the mean difference of a discrete distribution the direct formula may be used

$$\Delta = 2\sum_{i=2}^{s}\sum_{j=1}^{i-1}(x_i - x_j)p_i p_j \qquad [5]$$

or the formula based on the difference of each term from the previous one

$$\Delta = 2\sum_{i=2}^{s} P_i(1 - P_i)(x_i - x_{i-1})$$

or the formula

$$\Delta = 2\sum_{i=1}^{s} x_i p_i (2P_i - 1)$$

The three formulas are exactly the same. Since the cumulative probabilities almost always have complex expressions, the direct formula [5] has been used.

As shown in paragraph 4 of this note, the formulas of mean difference are obtained for the following 9 discrete distributions: Bernoulli, binomial, Poisson, negative binomial, uniform, geometric, hypergeometric, logarithmic and Zipf.

3. Formulas of the mean difference of 11 continuous distributions

3.1 Parabolic distribution

Parabolic distribution has density function

$$f(x) = \frac{3(2\mu + \omega - 2x)(2x + \omega - 2\mu)}{2\omega^3}, \quad \mu - \frac{\omega}{2} < x < \mu + \frac{\omega}{2}, \quad \omega > 0$$

in which μ is the location parameter equal to the mean value and ω is the scale parameter equal to the range of the variable. Using any one of the formulas [1]-[4], we simply arrive at

$$\Delta = \frac{9\omega}{35} = 0{,}257\omega,$$

thus demonstrating that the mean difference of the parabolic distribution is equal to slightly more than a quarter of the range.

3.2 Chi distribution

Chi distribution has density function

$$f(x) = \frac{x^{\nu-1}e^{-\frac{x^2}{2}}}{2^{\frac{\nu}{2}-1}\Gamma\left(\frac{\nu}{2}\right)}, \quad 0 < x < \infty, \qquad 0 < \nu < \infty, \qquad [6]$$

in which the sole parameterν, named degrees of freedom, is the shape parameter. By using formula [1], with significant simplification, we obtain

$$\Delta = \frac{2\sqrt{2}\Gamma\left(\nu+\frac{1}{2}\right)\left[\frac{2_2\mathrm{F}_1\left(\frac{\nu}{2},\ \nu+\frac{1}{2};\frac{\nu}{2}+1;-1\right)}{\nu} - \frac{2_2\mathrm{F}_1\left(\nu+\frac{1}{2},\ \frac{\nu}{2}+1;\frac{\nu+3}{2};-1\right)}{\nu+1}\right]}{\Gamma\left(\frac{\nu}{2}\right)^2} \qquad [7]$$

in which the values of two Gamma function and of two Gauss hypergeometric functions appear.

3.3 Seminormal distribution

Seminormal distribution is a particular case of chi distribution [6] with ν=1, possessing density function

$$f(x) = \sqrt{2/\pi}\, e^{-\frac{x^2}{2}}, \quad 0 < x < \infty,$$

and lacking parameters. By using one of the formulas [1]-[4] we obtain

$$\Delta = \frac{2\left(2-\sqrt{2}\right)}{\sqrt{\pi}} = 0{,}661.$$

After simplification, the same result is obtained with ν=1 in the formula of Δ of chi distribution [7].

3.4 Rayleigh distribution

Rayleigh distribution is a particular case of chi distribution [6] with ν=2, possessing density function

$$f(x) = xe^{-\frac{x^2}{2}}, \quad 0 < x < \infty,$$

and lacking parameters. By using one of the formulae [1]-[4] we obtain

$$\Delta = \left(\sqrt{2} - 1\right)\sqrt{\pi} = 0{,}734.$$

After simplification, the same result is obtained with ν=2 in the formula of Δ of the chi distribution [7].

3.5 Maxwell-Boltzman distribution

Maxwell-Boltzman distribution is a particular case of chi distribution [6] with ν=3, possessing density function

$$f(x) = \sqrt{\frac{2}{\pi}} x^2 e^{-\frac{x^2}{2}}, \quad 0 < x < \infty,$$

and lacking parameters. By using one of the formulas [1]-[4] we obtain

$$\Delta = \left(7 - 4\sqrt{2}\right)/\sqrt{\pi} = 0{,}758.$$

After simplification, the same result is obtained with ν=3 in the formula of Δ of chi distribution [7].

3.6 Dagum Distribution

Dagum distribution has density function

$$f(x) = \frac{\frac{\alpha\beta}{x}\left(\frac{x-\mu}{\theta}\right)^{\alpha\beta}}{\left[\left(\frac{x-\mu}{\theta}\right)^{\alpha} + 1\right]^{\beta+1}}, \quad \mu < x < \infty, \quad \theta > 0, \quad \alpha > 0, \quad \beta > 0$$

where μ is the minimum value of the variable, θ is the scale parameter, α and β are shape parameters.

After significant simplification, by using formula [1], we obtain

$$\Delta = 2\theta\Gamma\left(\frac{\alpha - 1}{\alpha}\right)\left[\frac{\Gamma\left(\frac{1}{\alpha} + 2\beta\right)}{\Gamma(2\beta)} - \frac{\Gamma\left(\frac{1}{\alpha} + \beta\right)}{\Gamma(\beta)}\right].$$

An analogous result was previously obtained (Girone, Viola, 2009).

3.7 Type-2 Gumbel distribution

Type-2 Gumbel distribution has density function

$$f(x) = \left[\alpha e^{-\left(\frac{x-\mu}{\theta}\right)^{-\alpha}} \left(\frac{x-\mu}{\theta}\right)^{-(\alpha+1)}\right]/\theta\,, \quad \mu < x < \infty, \quad \alpha > 0, \quad \theta > 0$$

in which μ is the minimum value of the variable, θ is the scale parameter and α is a shape parameter. By using formula [3] we obtain

$$\Delta = 2\theta\left(2^{1/\alpha} - 1\right)\Gamma\left(\frac{\alpha-1}{\alpha}\right), \quad \alpha > 1.$$

3.8 Type-3 Gumbel distribution

Type-3 Gumbel distribution has density function

$$f(x) = -\left[\alpha e^{-\left(\frac{x-\mu}{\theta}\right)^{\alpha}} \left(\frac{x-\mu}{\theta}\right)^{(\alpha+1)}\right]/\theta\,, \quad -\infty < x < \mu, \ \alpha > 0, \ \theta > 0$$

in which μ is the maximum value of the variable, θ is the scale parameter and α is the shape parameter. Utilizing formula [3], after some simplification, we obtain

$$\Delta = 2\theta(1 - 2^{-\frac{1}{\alpha}})\Gamma\left(\frac{\alpha+1}{\alpha}\right), \ \alpha > 1.$$

3.9 Snedecor distribution

Snedecor distribution has density function

$$f(x) = \frac{\left(\frac{\nu_1}{\nu_2}\right)^{\frac{\nu_1}{2}} x^{\frac{\nu_1}{2}-1}\left(1 + \frac{\nu_1}{\nu_2}x\right)^{-\frac{\nu_1+\nu_2}{2}}}{B\left(\frac{\nu_1}{2}, \frac{\nu_2}{2}\right)},$$

in which ν_1 and ν_2, named degrees of freedom, are positive integer shape parameters. Applying one of the formulas [1]-[4], exact values of Δ were calculated for the pairs of values of ν_1, from 1 to 10, and of ν_2, from 3 to 10. By examining these values, it was possible to determine their multiplicative structures by rows and columns. Relations were thus acquired which made it possible to obtain

$$\Delta = \frac{4\nu_2\Gamma\left(\frac{\nu_1+1}{2}\right)\Gamma\left(\frac{\nu_2-1}{2}\right)\Gamma\left(\frac{\nu_1+\nu_2}{2}\right)}{\nu_1(\nu_2-2)\sqrt{\pi}\Gamma\left(\frac{\nu_1}{2}\right)\Gamma\left(\frac{\nu_2}{2}\right)\Gamma\left(\frac{\nu_1+\nu_2-1}{2}\right)}.$$

An analogous result was previously obtained (Girone, Massari, 2015).

3.10 Distribution of the correlation coefficient

Distribution of the correlation coefficient in the bivariate normal distribution with independent components has density function

$$f(r) = \frac{\Gamma\left(\frac{n-1}{2}\right)(1-r^2)^{\frac{n-4}{2}}}{\sqrt{\pi}\,\Gamma\left(\frac{n-2}{2}\right)}, \quad -1 < r < +1, \quad n > 2,$$

in which parameter n is the sample size. By using one of the formulas [1]-[4] the exact numerical values of Δ were calculated for n=2, 3, …, 20. The values of Δ for the successive even values of n allow for determination of their multiplicative structure. Proceeding in a similar way for the odd values of n, the same result is obtained which leads to

$$\Delta = \frac{2^{n-2}\,\Gamma[(n-1)/2]^3}{\pi\Gamma[(2n-3)/2]\,\Gamma(n/2)}.$$

3.11 Inverse normal distribution

Inverse normal distribution has density function

$$f(x) = \sqrt{\frac{\lambda}{2\pi x^3}}\, e^{\frac{\lambda(x-\mu)^2}{2\mu^2 x}}, \quad 0 < x < \infty, \qquad \mu >, \qquad \lambda > 0.$$

Utilizing formula [3], after two transformations of variables and complex simplifications, we obtain

$$\Delta = \int_0^\infty \frac{8e^{-y^2}Erf(y)}{\sqrt{\pi}\sqrt{x^2+2\phi^2}},$$

Where $\phi = \sqrt{\lambda/\mu}$ and $Erf(x) = \frac{2}{\pi}\int_0^x e^{-t^2}dt$ is the error function, results which were obtained in a previous work (Girone, D'Uggento, 2016).

4 The mean difference of 9 discrete distributions

4.1 Bernoulli distribution

In Bernoulli distribution, the random variable X takes on two values only, 0 and 1, with probabilities 1-p and p. It is easy to show that the mean difference is

$$\Delta = 2p(1 - p).$$

The mean difference of Bernoulli distribution is equal to double the variance.

4.2 Binomial distribution

Binomial distribution, with p and n parameters is defined by the probability function

$$p_x = \binom{n}{x} p^x (1-p)^{n-x},$$

where x=0,1,…,n, $0 \leq p \leq 1$ and n is a positive integer.

If we apply formula [1], we obtain

$$\Delta = \sum_{i=0}^{n-1} \frac{2n\left[(1-n)^{(i)}\right]^2 p^{2(n-i)-1}(1-p)^{2i+1}}{(i!)^2}\left[1 + \frac{i\,p}{(1-p)(n-i)}\right], \qquad [8]$$

in which we find the Pochhammer function

$$(1-n)^{(i)} = \prod_{h=0}^{i-1}(h+1-n).$$

Subsequent to complex transformations, the formula [8], where the Pochhammer function is expressed as factorial, itself expressed in the Gauss hypergeometric functions, can also be put in a more closed form

$$\Delta = 2n(1-p)p^{(2n-1)} \cdot \cdot \left\{p_2\, F_1\left[1-n, 1-n, 1; \frac{(1-p)^2}{p^2}\right] + (n-1)(1-p)_2\, F_1\left[1-n, 2-n, 2; \frac{(1-n)^2}{p^2}\right]\right\}.$$

4.3 Poisson distribution

Poisson distribution, with λ parameter, is defined by the probability function

$$p_x = \frac{e^{-\lambda}\lambda^x}{x!}, \quad x = 0, 1, \ldots, \ \lambda > 0.$$

Applying formula [1], we obtain

$$\Delta = \sum_{x=1}^{\infty} \frac{2\left[\lambda^{1+2x} e^{-2\lambda} + \lambda^x e^{-\lambda}(x-\lambda)\Gamma(1+x,\lambda)\right]}{(x!)^2},$$

in which

$$\Gamma(\alpha, x) = \int_x^{\infty} e^{-t}\, t^{\alpha-1} dt$$

is the incomplete gamma function.

4.4 Negative binomial distribution

Negative binomial distribution, with p and s parameters, is defined by the probability function

$$p_x = \binom{s+x-1}{x} p^s (1-p)^x, \quad x = 0,1,2,\dots, \qquad 0 < p \leq 1, \qquad s > 0.$$

Through formula [1] we obtain

$$\Delta = 2sp^{s-1}(1-p) + \left[2\mathrm{p}^{\mathrm{s}-1}(1-\mathrm{p})^{2\mathrm{x}+1}\binom{\mathrm{s}+\mathrm{x}}{\mathrm{x}+1}^2 (\mathrm{x}+1)_2\mathrm{F}_1\,(1-\mathrm{s},\mathrm{x},2+\mathrm{x};1-\mathrm{p})\right]/(\mathrm{s}+\mathrm{x}),$$

which utilizes the value of a Gauss hypergeometric function.

4.5 Uniform distribution

Uniform distribution with N parameter is defined by the probability function

$$p_x = \frac{1}{N+1}, \qquad x = 0,1,2,\dots,N, \quad N \text{ positive integer.}$$

By applying formula [1] , the mean difference is

$$\Delta = \frac{2\mathrm{N}^2}{(1+\mathrm{N})^2}.$$

4.6 Geometric distribution

Geometric distribution with p parameter is defined by the probability function

$$p_x = p\,(1-p)^{x-1}, \quad x = 1,2,\dots, \quad 0 < p < 1.$$

By applying formula [1], we obtain

$$\Delta = \frac{2(1-p)}{p(2-p)}.$$

4.7 Hypergeometric distribution

Hypergeometric distribution, whose parameters are N, M and n, is defined by the probability function

$$p_x = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}},$$

Max(0,n-N+M)≤x≤Min(M,n), M≤ N,n,M and N positive integers.

Through formula [1], the mean difference can be obtained as follows

$$\Delta = \frac{Mnp_0}{N} + \sum_{x=1}^{n} p_x p_{x+1}\, {}_3F_2\left[\{2, 1-M+x, 1-n+x\}, \{2+x, 2-M-n+N+x\}, 1\right],$$

which utilizes the value of a generalized hypergeometric function.

4.8 Logarithmic distribution

Logarithmic distribution, with θ parameter, is defined by the probability function

$$p_x = \frac{\theta^x}{\log(1-\theta)\, x}, \quad x = 1, 2, \ldots, \quad 0 < \theta < 1.$$

Formula [1] leads to the following expression of the mean difference

$$\Delta = -\frac{2\left[\left(1-\theta+\theta^2\right) log\,(1-\theta) + log\,(1+\theta)\right]}{(1-\theta)[log\,(1-\theta)]^2} - 2\sum_{x=2}^{\infty} \frac{\theta^{2x}\Phi(\theta,1,x)}{[log\,(1-\theta)]^2},$$

in which we find the case s=1 of the Lerch transcendent function

$$\Phi(\theta, s, x) = \sum_{n=0}^{\infty} \frac{\theta^n}{(x+n)^s}.$$

4.9 Zipf distribution

Zipf distribution, with parameter α>0, is defined by the probability function

$$p_x = \frac{x^{-(a+1)}}{\zeta(a+1)}, \quad x = 1, 2, \ldots, \qquad \alpha > 0,$$

in which we find the Riemann zeta function

$$\zeta(\alpha+1) = \sum_{h=1}^{\infty} h^{-(\alpha+1)}.$$

Formula [1] leads to the following expression of the mean difference

$$\Delta = 2\sum_{x=2}^{\infty} \frac{x\, H_{x-1}^{(\alpha+1)} - H_{x-1}^{(\alpha)}}{x^{a+1}[\zeta(a+1)]^2},$$

where

$$H_n^{(r)} = \sum_{k=1}^{n} 1/k^r,$$

are the generalized harmonic numbers.

4. Conclusions

In this note, as a follow up to our previous study, expressions of the mean difference for other 11 continuous and 9 discrete distributions were obtained.

These expressions are compact in many cases and complicated in a few discrete instances, although also in the latter case, they are more easily calculated than the general formulas. They were obtained by solving the double integral or double sum, with the help of the software Mathematica, and by using a number of simplification expedients in order to obtain more compact formulas.

These new results contribute to integrating the characterization of distribution variability considered as a measure of the inequality among the values, rather than their spread. The resulting relations between the formulas of the mean difference and other variability measures with the parameters allows for fuller appreciation of the distributions.

Our hope is that, with this aim in mind, they may in future be included in the handbooks of statistical distributions.

References

D'Uggento, A.M.; Girone, G.; Marin, C. (2017). The relation between the mean difference and the mean deviation in 11 continuous distribution models. *Quality and Quantity*, Vol. 51, Issue 2: p. 595-615. doi: 10/1007/s11135-016-0427-x (online first, 2016, 06 october).

Gini, C. (1912). *Variabilità e mutabilità*, Studieconomico-giuridici, published by the Law Faculty of the Royal University of Cagliari, year III, part II, Cuppini, Bologna.

Girone, G.; D'Uggento, A. M. (2016). About the mean difference of the inverse normal distribution. *Applied Mathematics*, Vol.7 No.14. doi: 10.4236/am.2016.714130.

Girone, G.; Massari, A. (2015). La differenza media della variabile di Snedecor. *Studi in ricordo di Carlo Cecchi*, Università degli Studi di Bari Aldo Moro, Bari, p. 9-14.

Girone, G.; Massari, A; Manca, F. (2017). The relation between the mean difference and the standard deviation in continuous distribution models. *Quality and Quantity* Vol. 51, Issue 2: p. 481–507. doi: 10/1007/s11135-016-0398-y (online first, 2016, 14 september).

Girone, G.; Mazzitelli D. (2007). La differenza media nei principali modelli distributivi continui, *Annali del Dipartimento di Scienze statistiche "Carlo Cecchi"*, Università degli Studi di Bari, vol.VI, tomo I, p. 45-62.

Girone, G.; Viola, D. (2009). La differenza media della distribuzione di Dagum.*Annali del Dipartimento di Scienze statistiche "Carlo Cecchi"*, Università degli Studi di Bari, vol.8, Cleup, p. 101-106.

Johnson, N.; Kemp, A.W.; Kotz, S. (2005). *Univariate discrete distributions,* Wi–ley, New York, 2005.

Johnson, N.; Kotz, S.; Balakrishnan, N. (1994). *Continuous univariate distributions*, vol. 1, Wiley, New York.

Johnson, N.; Kotz, S.; Balakrishnan, N. (1995). *Continuous univariate distributions*, vol. 2, Wiley, New York.

Patel, K. J.; Kapadia, C. H.; Owen D. B. (1976). *Handbook of statistical distributions,* Marcel Dekker, New York & Basel.

Stuart, A.; Ord, J. K., (1994). *Kendall's Advanced Theory of Statistics*, vol. I, Distribution Theory, Oxford University Press, New York.

Yitzhaki, S. (2003). Gini's mean difference: a superior measure of variability for non normal distributions. *Metron,* Vol. LXI, n. 2: 285-316.

UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO *Metodi e analisi statistiche 2017* ISBN: 978-88-6629-013-1
DIPARTIMENTO DI ECONOMIA E FINANZA pp. 35-41

La differenza media della distribuzione normale generalizzata

Francesco Campobasso[1]*, Angela Maria D'Uggento[1], Claudia Marin[2]

[1] *Università degli studi di Bari-Dipartimento di Economia e Finanza,*
[2] *Università degli studi di Bari-Dip. di Scienze della Formazione, Psicologia, Comunicazione*

Riassunto: Lo scopo di questa nota è quello di fornire la formula della differnza media della distribuzione normale generalizzata, della quale sono noti invece i valori caratteristici. Colmare tale lacuna ha così consentito di ricavare le relazioni numeriche tra differenza media e scarto semplice medio, tra differenza media e scarto quadratico medio e tra scarto quadratico medio e scarto semplice medio di detto modello distributivo.

Keywords: differenza media; distribuzione normale; modelli distributivi.

1. Introduzione

La distribuzione normale generalizzata è un importante modello distributivo continuo e simmetrico (Nadarajah, 2005) con tre parametri: uno di posizione, uno di scala ed uno di forma. Al variare di quest'ultimo parametro si hanno distribuzioni diverse (laplaciana, normale, ..., uniforme).

Della distribuzione normale generalizzata sono noti i valori caratteristici (media, moda, mediana, scarto semplice medio, scarto quadratico medio, indice di disnormalità, entropia, ecc.). Non è nota la formula della differenza media.

Finalità di questa nota è quella di colmare tale lacuna. Si studieranno altresì le relazioni tra gli indici di variabilità.

* Autore corrispondente: francesco.campobasso@uniba.it
Il lavoro qui descritto è frutto di un progetto comune, ma F. Campobasso ha provveduto alla redazione dei paragrafi 1 e 3, mentre A. M. D'Uggento ha redatto il paragrafo 4 e 5 e C. Marin il paragrafo 2.

2. La distribuzione normale generalizzata

Al fine di ricavare la differenza media della distribuzione normale generalizzata è sufficiente considerare la sua forma ridotta in cui il parametro di posizione è nullo e quello di scala è unitario. Una volta ottenuta la espressione della differenza media per la distribuzione ridotta basterà moltiplicarla per il parametro di scala per ottenere la differenza media della distribuzione non ridotta.

La funzione di densità e la funzione di ripartizione della distribuzione normale generalizzata ridotta sono

$$f(x) = \frac{e^{-|x|^{\beta}}}{2\Gamma\left(\frac{1}{\beta}\right)}, \qquad -\infty < x < \infty, \qquad \beta > 0\,,$$

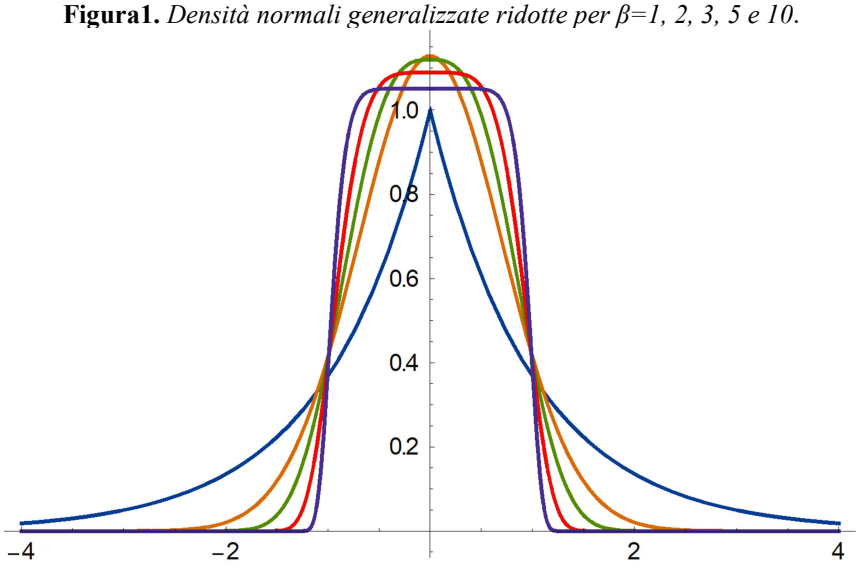
$$F(x) = \frac{1}{2} + segno(x)\frac{\Gamma\left(\frac{1}{\beta}\right) - \Gamma\left(|x|, \frac{1}{\beta}\right)}{2\Gamma\left(\frac{1}{\beta}\right)}, \qquad \beta > 0\,,$$

nelle quali compaiono la funzione segno, la funzione gamma e la funzione gamma incompleta.

Nella seguente Fig. 1 sono rappresentate le densità della distribuzione normale generalizzata ridotta per i valori del parametro β = 1, 2, 3, 5 e 10:

È agevole accertare la varietà dei modelli distributivi al variare del parametro β.

Figura1. *Densità normali generalizzate ridotte per β=1, 2, 3, 5 e 10.*



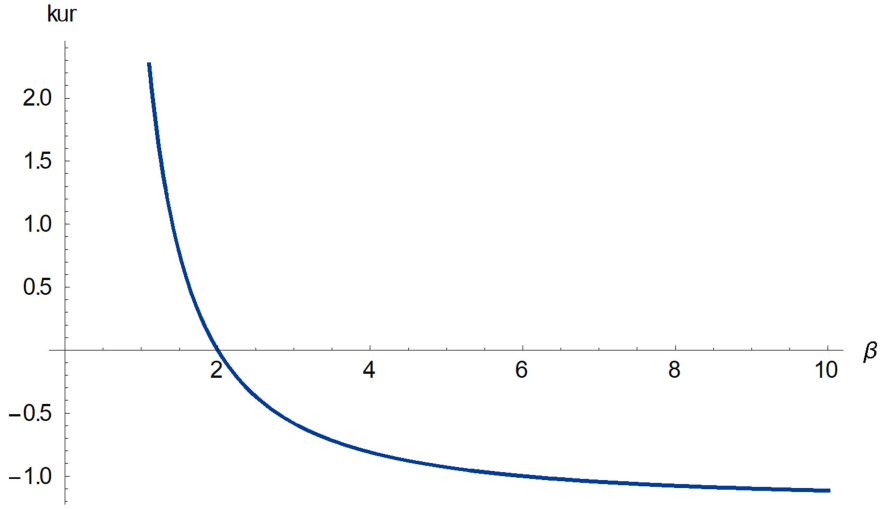
L'indice di disnormalità della distribuzione normale generalizzata è pari a

$$kur(X) = \frac{\Gamma\left(\frac{1}{\beta}\right)\Gamma\left(\frac{5}{\beta}\right)}{\left[\Gamma\left(\frac{3}{\beta}\right)\right]^2} - 3.$$

Nella seguente Fig. 2 è riportato l'andamento dell'indice di disnormalità al variare del parametro β.

Come può vedersi il parametro di forma β è anche un indice di disnormalità. Per i valori di $\beta < 2$ la distribuzione normale generalizzata è ipernormale di tipologia laplaciana, in particolare per $\beta = 1$ è esattamente laplaciana. Per $\beta = 2$ la distribuzione è esattamente normale. Per $\beta > 2$ la distribuzione normale generalizzata è iponormale e, al crescere di detto parametro, tende rapidamente alla uniformità.

Figura 2. *Andamento dell'indice di disnormalità al variare del parametro β della distribuzione normale generalizzata.*



3. Gli indici di variabilità della distribuzione normale generalizzata ridotta

Lo scarto semplice medio e lo scarto quadratico medio della distribuzione normale generalizzata sono

$$\delta = \frac{\Gamma\left(\frac{2}{\beta}\right)}{\Gamma\left(\frac{1}{\beta}\right)},\ \beta > 0, \qquad \sigma = \sqrt{\frac{\Gamma\left(\frac{3}{\beta}\right)}{\Gamma\left(\frac{1}{\beta}\right)}},\quad \beta > 0.$$

La differenza media invece non è nota.

Una delle formule più semplici di calcolo della differenza media di una distribuzione continua utilizza la funzione di ripartizione *F(x)*:

$$\Delta = \int_{-\infty}^{\infty} 2F(x)[1 - F(x)]dx.$$

Applicando detta formula, dopo pesanti semplificazioni, si perviene alla

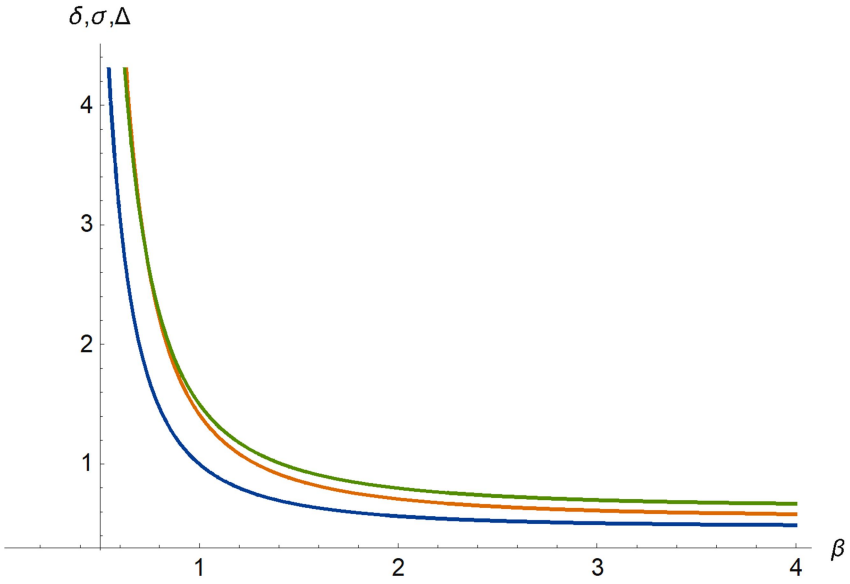
$$\Delta = \frac{\Gamma\left(\frac{2}{\beta}\right)}{\Gamma\left(\frac{1}{\beta}\right)} + \frac{\beta\Gamma\left(\frac{3}{\beta}\right)\left[2\,{}_2F_1\left(\frac{1}{\beta},\frac{3}{\beta},\frac{\beta+1}{\beta},-1\right) - {}_2F_1\left(\frac{2}{\beta},\frac{3}{\beta},\frac{\beta+2}{\beta},-1\right)\right]}{2\left[\Gamma\left(\frac{1}{\beta}\right)\right]^2}, \beta > 0.$$

Nella precedente espressione, oltre alla funzione gamma, compare la funzione ipergeometrica di Gauss. Si noti poi che il primo addendo della espressione è lo scarto semplice medio.

La formula fornisce espressioni reali esatte della differenza media della distribuzione normale generalizzata nel caso in cui il parametro β assume valori razionali positivi e conduce a valori numerici esatti se detto parametro assume altri valori reali positivi.

Nella seguente Fig. 3 sono riportati gli andamenti dei tre suddetti indici di variabilità in relazione al parametro β.

Figura 3. *Differenza media, scarto quadratico medio e scarto semplice medio della distribuzione normale generalizzata, in relazione al parametro β.*



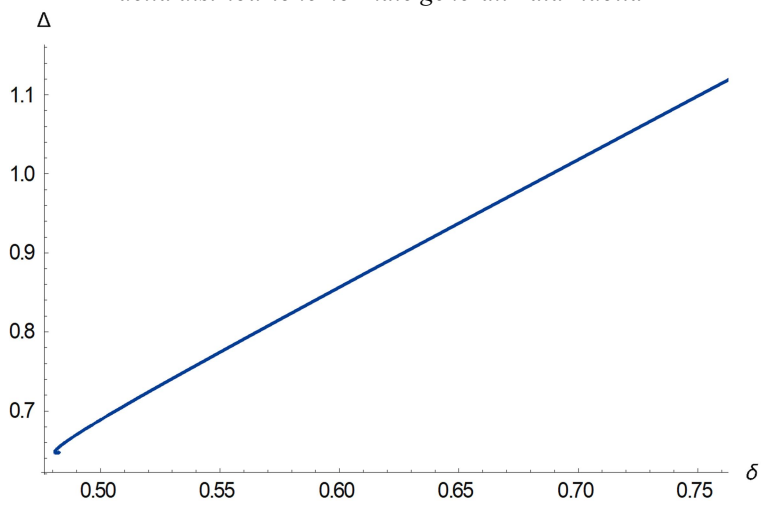
Come può vedersi, tutti i tre indici diminuiscono al crescere del parametro β, inoltre lo scarto semplice medio per ogni valore di β assume il valore più piccolo; lo scarto quadratico medio è maggiore della differenza media per $\beta < 0{,}742$, per i valori maggiori la differenza media, invece, supera lo scarto quadratico medio. Al limite per $\beta \to \infty$ la distribuzione normale generalizzata tende alla distribuzione uniforme tra -2 e $+2$; conseguentemente i tre indici di variabilità tendono a quelli di tale distribuzione limite, rispettivamente $2/3$, $1/\sqrt{3}$ e $1/2$.

4. Relazioni tra gli indici di variabilità della distribuzione normale generalizzata ridotta

Non è possibile invertire le espressioni degli indici di variabilità della distribuzione normale generalizzata in funzione del parametro β, per cui non è possibile ricavare le relazioni analitiche tra lo scarto semplice medio e lo scarto quadratico medio, tra la differenza media e lo scarto semplice medio e tra la differenza media e lo scarto quadratico medio della distribuzione normale generalizzata. È invece possibile ricavare le relazioni numeriche calcolando i valori dei tre indici per una congrua serie di valori del parametro β e rappresentando le coppie di indici sul piano cartesiano.

Nella seguente Fig. 4 è riportata la relazione numerica tra differenza media e scarto semplice medio della distribuzione normale generalizzata. Come può vedersi la relazione è pressoché lineare.

Figura 4. *Relazione tra differenza media e scarto semplice medio della distribuzione normale generalizzata ridotta.*



Nelle successive Figure 5 e 6 sono riportate le relazioni numeriche, rispettivamente, tra la differenza media e lo scarto quadratico medio della distribuzione normale generalizzata e tra la lo scarto semplice medio e lo scarto quadratico medio della medesima distribuzione. Come è evidente, anche queste relazioni sono pressoché lineari.

Figura 5. *Relazione tra differenza media e scarto quadratico medio della distribuzione normale generalizzata ridotta.*

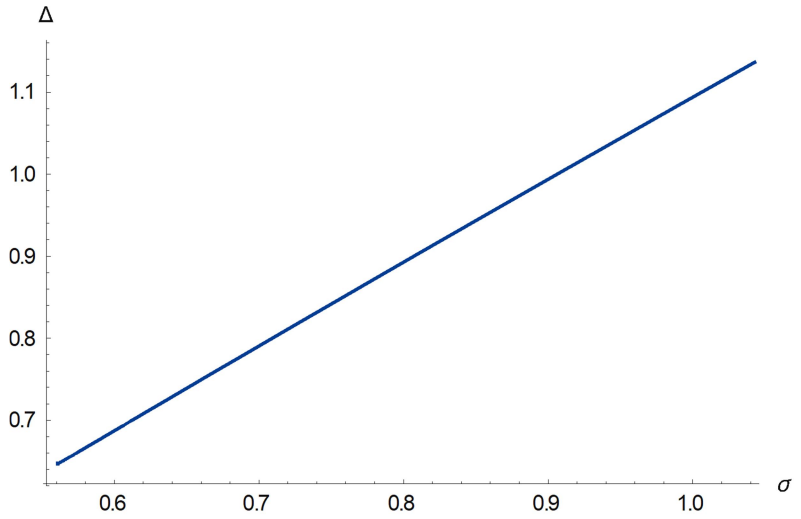
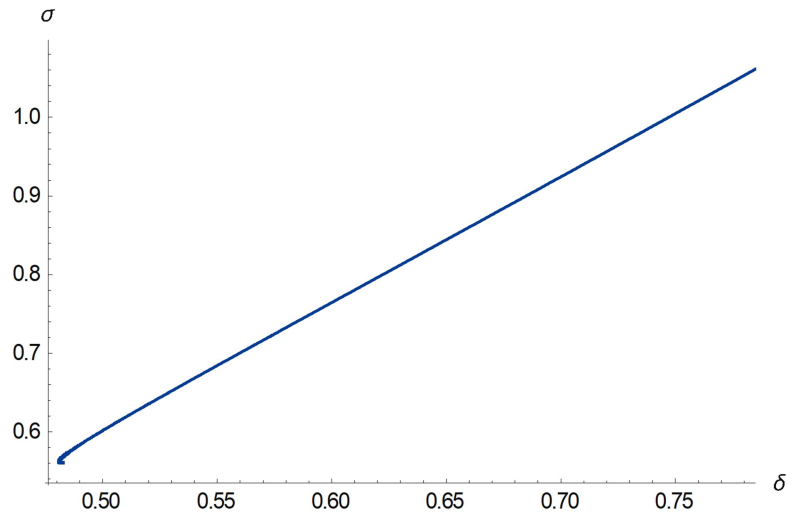


Figura 6. *Relazione tra scarto semplice medio e scarto quadratico medio della distribuzione normale generalizzata ridotta.*



5. Conclusioni

La distribuzione normale generalizzata è un importante modello distributivo che comprende, in relazione ai valori del suo parametro di forma, numerose altre distribuzioni (Laplace, normale, uniforme, ecc.). A cagione della sua duttilità è stata largamente applicata e ne sono state ricavate le principali caratteristiche.

Una delle carenze è costituita dal fatto che non se ne conosceva la espressione generale della differenza media.

In questa nota, tale lacuna è stata colmata ottenendo una espressione dell'indice di variabilità giniano in una forma generale e di facile calcolo. Grazie a tale risultato si è reso possibile ricavare le relazioni numeriche tra differenza media e scarto semplice medio, tra differenza media e scarto quadratico medio e tra scarto quadratico medio e scarto semplice medio di detto modello, relazioni che sono risultate tutte pressoché lineari.

Riferimenti bibliografici e sitografici

Nadarajah, S. (1982). *A generalized normal distribution.* Journal of Applied Statistics. **32** (7), September 2005, pag. 685–694.

Wikipedia (2017). https://en.wikipedia.org/wiki/Generalized_normal_distribution (aggiornamento 16 Ottobre)

UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO
Metodi e analisi statistiche 2017 ISBN: 978-88-6629-013-1
DIPARTIMENTO DI ECONOMIA E FINANZA pp. 43–50

Una caratterizzazione per la risolubilità dell'equazione di Black-Scholes-Merton

Sabrina Diomede [1*], **Giovanni Taglialatela** [2]

[1] *Dipartimento di Economia, Management e Diritto dell'Impresa*
[2] *Dipartimento di Economia e Finanza,*
Università degli Studi di Bari Aldo Moro

Riassunto. In questo lavoro si propone una caratterizzazione della risolubilità di un problema differenziale associato all'equazione di Black-Scholes-Merton che utilizza alcune tecniche della teoria dei semigruppi. A tal riguardo, una piccola sezione preliminare fornisce dei richiami utili ad illustrare un'interconnessione fra la teoria degli operatori e quella delle equazioni alle derivate parziali.

Keywords: equazione di Black-Scholes-Merton; formula asintotica; processo di approssimazione mediante operatori positivi; saturazione locale; semigruppi di operatori sullo spazio delle funzioni continue.

1. Introduzione

Uno dei problemi fondamentali dell'economia è quello di assegnare un prezzo a beni o titoli finanziari da acquistare o vendere sul mercato.

In particolare fra questi beni vi sono le opzioni, contratti finanziari con cui il compratore acquista il diritto (ma non l'obbligo) di scambiare in una certa data prefissata, o entro una certa data, un'attività finanziaria (titoli, valute, etc...) (detta *sottostante*) ad un prezzo concordato precedentemente (detto *prezzo di esercizio*). Si parla di *opzione di tipo Call* in caso di diritto di acquisto e *opzione di tipo Put* in caso di diritto di vendita.

* Autore corrispondente: sabrina.diomede@uniba.it

Si chiamano invece *opzioni di tipo Europeo* quelle per cui è possibile esercitare il proprio diritto solo alla scadenza del contratto e *opzioni di tipo Americano* quelle per cui è possibile esercitare il proprio diritto in qualsiasi momento fino alla scadenza del contratto.

Per quanto concerne l'altra parte del contratto, ossia colui il quale sottoscrive il contratto insieme al compratore, questi ha, a differenza del compratore, un potenziale obbligo: egli è tenuto a vendere/acquistare il sottostante nel momento in cui il compratore decide di esercitare il suo diritto allo scambio.

Poiché l'opzione conferisce a chi la acquista un diritto senza obbligo, essa ha un suo valore. Il problema di determinare tale valore, che dipende certamente dal valore intrinseco del sottostante e dal fattore tempo, ma anche da numerosi altri parametri caratteristici del mercato finanziario sul quale si verificano gli scambi, è stato affrontato in modo rigoroso da Black e Scholes i quali, nel 1973, proposero un modello successivamente esteso dal punto di vista matematico da Merton, noto come modello di Black-Scholes-Merton (BSM).

In esso, mediante un'equazione differenziale alle derivate parziali di tipo parabolico (denominata appunto equazione di Black-Scholes-Merton), si stima il prezzo delle opzioni europee e inoltre si dimostra che l'opzione ha un prezzo unico indipendentemente dal rischio e dal rendimento atteso del sottostante.

La possibilità di attribuire un prezzo alle opzioni, o più in generale agli strumenti derivati, ha enormemente ampliato ed agevolato l'attività di trading e lo sviluppo di mercati dei derivati (negli Stati Uniti il mercato di riferimento è il Chicago Board of Trade, in Italia l'Italian Derivative Market).

Il modello BSM ha consentito e consente peraltro di effettuare investimenti nei quali, mediante opportuni acquisti o vendite del sottostante (delta hedging) si riesce a mitigare il rischio implicitamente riposto nei mercati finanziari.

Nonostante siano passati molti anni da allora e l'ingegneria finanziaria abbia fatto grandi progressi, la formula di BSM è tutt'ora la piú diffusa tra gli operatori del settore i quali continuano ad averla quale riferimento nelle loro attività di trading, sia pure con opportuni aggiustamenti.

Le assunzioni originarie alla base del modello si distinguono in assunzioni sui titoli e assunzioni sul mercato.

Le prime sono:

- esiste almeno un titolo privo di rischio e almeno un titolo rischioso;
- il tasso d'interesse del titolo privo di rischio r è costante;
- i rendimenti dei titoli sono continui e i titoli sono scambiati sul mercato in tempo continuo;

- il rendimento logaritmico del sottostante S è un moto browniano geometrico con media e varianza costanti;
- il titolo rischioso non paga dividendi.

Le assunzioni che riguardano il mercato sono:

- non vi sono opportunità d'arbitraggio;
- sono consentite vendite allo scoperto;
- non vi sono costi di transazione, tassazione, né altri tipi di attriti nel mercato;
- è possibile comprare o vendere qualunque ammontare di S;
- è possibile prestare o prendere in prestito qualunque ammontare al tasso privo di rischio.

Versioni aggiornate del modello danno conto di tassi non costanti, dividendi e costi di transazione.

Nel modello BSM il prezzo dell'opzione u dipende solo dal prezzo del sottostante S e dal tempo T, avendo supposto tutte le altre quantità costanti, ed è dato dalla soluzione dell'equazione

$$\frac{\partial u}{\partial T}(S,T) = ru(S,T) - rS\frac{\partial u}{\partial S}(S,T) - \frac{\sigma^2 S^2}{2}\frac{\partial^2 u}{\partial S^2}(S,T)\,, \qquad (S>0, T<T_0) \tag{1}$$

soggetta alla condizione iniziale

$$u(S,T_0) = \max(S-K,0) \qquad (S>0)\,, \tag{2}$$

dove K è il prezzo di esercizio dell'opzione, e T_0 è il tempo di scadenza dell'opzione.

Sostituendo per semplicità T con $T_0 - t$ e S con x, la (1) diventa

$$\frac{\partial u}{\partial t}(x,t) = \frac{\sigma^2 x^2}{2}\frac{\partial^2 u}{\partial x^2}(x,s) + rx\frac{\partial u}{\partial x}(x,t) - ru(x,t)\,, \qquad (x>0, t>0)\,.$$

Tale equazione è un'equazione di tipo parabolico degenere, riconducibile all'equazione del calore mediante un opportuno cambio di variabili. È dunque possibile ottenere la soluzione esplicita

$$u(t,x) = x\,N\left(\frac{\log(x/K)}{\sigma\sqrt{t}} + \frac{\sigma}{2}\sqrt{t}\right) - K\,N\left(\frac{\log(x/K)}{\sigma\sqrt{t}} - \frac{\sigma}{2}\sqrt{t}\right) \tag{3}$$

dove $N(x)$ è la funzione di ripartizione della distribuzione normale standardizzata

$$N(x) := \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-t^2/2}\,dt\,.$$

Sebbene sia possibile esprimere la soluzione dell'equazione (1) in forma esplicita, essa non risulta sempre di facile applicazione.

Ad esempio uno dei problemi delle finanza è quello di dedurre la *volatilità implicita*, ovvero, dati t, x, K, dedurre σ dalla (3). Sfortunatamente, la presenza della funzione di ripartizione $N(x)$ nella (3) non permette di scrivere in forma esplicita σ in funzione di t, x, K e numerose formule approssimate sono state proposte (si veda ad esempio Orlando e Taglialatela (2017) per un'esposizione sull'argomento).

Un approccio a tale problema è studiare come approssimare l'operatore differenziale in (1) in modo da ottenere soluzioni facilmente invertibili e quindi ottenere approssimazioni della soluzione (3).

In questa nota ci proponiamo di caratterizzare i dati iniziali per cui il problema di Cauchy associato sia risolubile. Un riferimento per una trattazione non troppo complessa tuttavia esaustiva per quanto concerne gli aspetti matematici e finanziari del modello BSM può essere rappresentato da Wilmott e al. (1995) (cfr. in particolare Part. 1, capp. 1-6 e i riferimenti bibliografici ivi contenuti).

2. Richiami

In quanto segue denoteremo con $\mathcal{C}\big(]0,+\infty[\big)$ lo spazio delle funzioni reali continue definite sull'intervallo $]0,+\infty[$ e con $\mathcal{C}^2\big(]0,+\infty[\big)$ quello delle funzioni reali su $]0,+\infty[$ differenziabili due volte con derivata seconda continua.

Siano $\alpha, \beta, \gamma \in \mathcal{C}\big(]0,+\infty[\big)$ e si supponga che γ sia limitata. Si consideri il seguente problema ai valori iniziali

$$\text{(4)} \qquad \begin{cases} \dfrac{\partial u}{\partial t}(x,t) = \alpha(x)\,\dfrac{\partial^2 u}{\partial x^2}(x,t) + \beta(x)\,\dfrac{\partial u}{\partial x}(x,t) + \gamma(x)\,u(x,t) \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (x>0, \quad t\geq 0)) \\ u(x,0) = u_0(x) \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (x>0) \end{cases}$$

soggetta alle *condizioni di compatibilità*

$$\text{(5)} \qquad \begin{cases} \lim\limits_{x\to 0} \alpha(x)\dfrac{\partial^2 u}{\partial x^2}(x,t) + \beta(x)\dfrac{\partial u}{\partial x}(x,t) = 0 \qquad (t\geq 0) \\ \lim\limits_{x\to +\infty} \alpha(x)\dfrac{\partial^2 u}{\partial x^2}(x,t) + \beta(x)\dfrac{\partial u}{\partial x}(x,t) = 0 \quad (t\geq 0) \end{cases}$$

Diremo che una funzione continua

$$v\colon\ ]0,+\infty[\,\times\,[0,+\infty[\to \mathbb{R}$$

è una *soluzione classica* di (4)-(5) se:

(1) esiste $\dfrac{\partial v}{\partial t}$ ed è continua su $]0,+\infty[\,\times\,[0,+\infty[$;

(2) esistono $\dfrac{\partial^2 v}{\partial x^2}, \dfrac{\partial v}{\partial x}$ e sono continue su $]0,+\infty[\times [0,+\infty[$;

(3) v soddisfa (4) e (5).

L'esistenza di una soluzione del problema (4)-(5) si può ricondurre ad un opportuno problema di Cauchy astratto considerando l'insieme $D(A)$ delle funzioni

$$u \in \mathcal{C}\big([0,+\infty[\big) \cap \mathcal{C}^2\big(]0,+\infty[\big)$$

per cui

$$\lim_{x\to 0} \alpha(x)\, u''(x) + \beta(x)\, u'(x) = 0$$
$$\lim_{x\to +\infty} \alpha(x)\, u''(x) + \beta(x)\, u'(x) = 0$$

e, per ogni $u \in D(A)$, l'operatore

$$A(u)(x) := \begin{cases} \alpha(x)u''(x) + \beta(x)u'(x) + \gamma(x)u(x) & (x > 0) \\ 0 & (x = 0) \end{cases}$$

Allora, considerato il Problema di Cauchy astratto

$$\text{(6)} \qquad \begin{cases} \dfrac{du}{dt}(t) = Au(t) & t \geq 0 \\ u(0) = u_0 & u_0 \in D(A) \end{cases}$$

si riconosce che $u\colon t \in [0,+\infty[\mapsto u(t) \in \mathcal{C}\big([0,+\infty[\big)$ è una soluzione di (6) se, e solo se, la funzione

$$v(x,t) := u(t)(x)$$

è una soluzione di (4)-(5)

Ebbene, se $\big(A, D(A)\big)$ è un operatore chiuso su di uno spazio di Banach $\big(E, \|\cdot\|\big)$ che genera un semigruppo fortemente continuo $\big(T(t)\big)_{t\geq 0}$ di operatori su E, allora per ogni $u_0 \in D(A)$ esiste una ed una sola soluzione di (6), ed in tal caso la soluzione si esprime mediante il semigruppo secondo la relazione

$$u(t) = T(t)(u_0) \qquad (t \geq 0)$$

(si veda, a.e., Engel e Nagel (2000), Capitolo II, Teorema 6.7).

Si dimostra inoltre che la soluzione dipende in modo continuo dal dato iniziale. Infatti, poiché per ogni semigruppo fortemente continuo $\big(T(t)\big)_{t\geq 0}$ esistono $\omega \in \mathbb{R}$ ed $M \geq 1$ tali che, per ogni $t \geq 0$

$$\big\|T(t)\big\| \leq Me^{\omega t}\,,$$

se u_0 e v_0 sono due dati iniziali appartenenti al dominio $D(A)$ di A, risulta

$$\big\|u(t)-v(t)\big\| \le \big\|T(t)\big\|\,\|u_0-v_0\| \le Me^{\omega t}\,\|u_0-v_0\|\,.$$

3. La caratterizzazione

Si consideri lo spazio delle funzioni continue con peso definito ponendo

$$\mathcal{C}_0^{w_2}\big(]0,+\infty[\big) := \left\{ f\in\mathcal{C}\big(]0,+\infty[\big) \;\middle|\; \lim_{x\to 0} f(x) = \lim_{x\to+\infty}\frac{f(x)}{1+x^2}=0 \right\}$$

che, munito della norma con peso

$$\|f\|_2 := \sup_{x\ge 0}\frac{\big|f(x)\big|}{1+x^2}$$

risulta essere uno spazio di Banach.

Teorema 1. *Si consideri il problema differenziale associato all'equazione di Black-Scholes-Merton*

$$\text{(7)}\qquad \begin{cases} \dfrac{\partial u}{\partial t}(x,t) = \dfrac{\sigma^2x^2}{2}\dfrac{\partial^2 u}{\partial x^2}(x,t) + rx\dfrac{\partial u}{\partial x}(x,t) - ru(x,t) & (x>0, t>0)\,, \\ u(x,0)=u_0(x) & (x>0)\,, \end{cases}$$

con le condizioni al bordo

$$\text{(8)}\qquad \begin{cases} \lim\limits_{x\to 0^+} \dfrac{\sigma^2}{2}x^2\dfrac{\partial^2 u}{\partial x^2}(x,t) + rx\dfrac{\partial u}{\partial x}(x,t) = 0 & (t\ge 0)\,, \\ \lim\limits_{x\to+\infty}\dfrac{1}{1+x^2}\Big(\dfrac{\sigma^2x^2}{2}\dfrac{\partial^2 u}{\partial x^2}(x,t) + rx\dfrac{\partial u}{\partial x}(x,t)\Big) = 0 & (t\ge 0)\,, \\ \lim\limits_{x\to 0^+} u(x,t) = 0 & (t\ge 0)\,, \\ \lim\limits_{x\to+\infty}\dfrac{u(x,t)}{1+x^2}=0 & (t\ge 0)\,. \end{cases}$$

Allora, le funzioni $u_0\in\mathcal{C}_0^{w_2}\big(]0,+\infty[\big)$ *tali che*

(i) $u_0\in\mathcal{C}^2\big(]0,+\infty[\big)$,

(ii) $\lim\limits_{x\to 0}\Big(\dfrac{x^2}{2}\,u''(x)+\dfrac{r}{\sigma^2}\,xu'(x)\Big)=0$,

(iii) $\lim\limits_{x\to+\infty}\dfrac{1}{1+x^2}\Big(\dfrac{x^2}{2}u''(x)+\dfrac{r}{\sigma^2}xu'(x)\Big)=0$

sono tutte e sole le funzioni dato iniziale per le quali il problema differenziale (7)–(8) *ammette una* (*ed una sola*) *soluzione* $u\colon\ ]0,+\infty[\times[0,+\infty[\to\mathbb{R}$ *la quale dipende in modo continuo da* u_0.

Dimostrazione. Sia $u_0 \in \mathcal{C}_0^{w_2}(]0,+\infty[$ una funzione verificante le (i), (ii) ed (iii) dell'enunciato. Si consideri inoltre il sottospazio $D(A)$ di $\mathcal{C}_0^{w_2}\big(]0,+\infty[\big)$ delle funzioni verificanti

$$\lim_{\substack{x\to 0^+\\ x\to+\infty}} \frac{1}{1+x^2}\left(\frac{\sigma^2x^2}{2}u''(x)+rxu'(x)\right)=0\,,$$

il quale coincide con il dominio dell'operatore differenziale

$$Au(x)=\frac{\sigma^2x^2}{2}u''(x)+rxu'(x)-ru(x). \qquad (x>0)$$

L'operatore $\big(A, D(A)\big)$ è il generatore di un semigruppo fortemente continuo di operatori sullo spazio $\mathcal{C}_0^{w_2}\big(]0,+\infty[\big)$ (cf. sez. 4 di Altomare e Attalienti (2002)). Dalla teoria dei semigruppi, secondo quanto richiamato nella sezione 2, consegue l'esistenza e l'unicità della soluzione prevista dall'asserto.

4. Approssimazione della soluzione

Alla luce del legame fra un'ampia classe di problemi differenziali, teoria dei semigruppi ed approssimazione delle soluzioni dei problemi differenziali documentato in Altomare e al. (2014), capp. 1, 2 e 3 e i riferimenti ivi indicati, è utile sottolineare che le funzioni $u_0 \in \mathcal{C}_0^{w_2}\big(]0,+\infty[\big)$ verificanti (i), (ii) e (iii) del Teorema precedente sono tutte e sole quelle per cui la successione

$$\Big(n\big(\mathscr{Q}_n(u_0)-u_0\big)\Big)_{n\geq 1}$$

converge puntualmente su $]0,+\infty[$, ove per ogni $n \geq 1$, l'operatore $\mathscr{Q}_n$ è definito ponendo

(9)

$$\mathscr{Q}_n(f)(x) := \begin{cases} \Big(1-\dfrac{r}{\sigma^2 n}\Big)\dfrac{n^{\frac{n}{2}}}{(2x)^{\frac{n}{2}}\,\Gamma\Big(\dfrac{n}{2}\Big)}\displaystyle\int_0^{+\infty} v^{\frac{n}{2}-1}\,e^{-\frac{nv}{2x}}\,f\left(\Big(1+\frac{r}{\sigma^2}\Big)v\right)dv & x>0 \\ \Big(1-\dfrac{r}{\sigma^2 n}\Big)\,f(0) & x=0\,, \end{cases}$$

ove $r, \sigma > 0$ e $n \geq \dfrac{r}{\sigma^2}$ per ogni f in

$$\left\{\, g\in\mathcal{C}\big([0,+\infty[\big) \;\;\Big|\;\; \sup_{x\geq 0}\frac{|g(x)|}{1+x^4}<+\infty \,\right\},$$

e $\Gamma(x)$ è la funzione Gamma di Eulero.

Inoltre, come giustificato nella sez. 5.3 di Altomare e Diomede (2010) il

$$\lim_{n\to+\infty} n\big(\mathscr{L}_n(u_0)-u_0\big)(x)=\frac{x^2}{2}\,u_0''(x)+\frac{r}{\sigma^2}x\,u_0'(x)-\frac{r}{\sigma^2}\,u_0(x)$$

e la soluzione al problema differenziale (7) può esprimersi mediante opportune iterate degli operatori $\mathscr{L}_n$ secondo la relazione

$$u(x,t)=\lim_{n\to+\infty}\mathscr{L}_n^{K(n)}(u_0)(x)$$

ove $\big(K(n)\big)_{n\geq 1}$ è un'arbitraria successione di interi positivi tale che

$$\lim_{n\to+\infty}\frac{K(n)}{n}=t\,.$$

(ad esempio $K(n)=[nt]$).

Riferimenti bibliografici

Altomare, F., Attalienti A. (2002). Degenerate evolution equations in weighted continuous function spaces, Markov processes and the Black-Scholes equation – Part II, *Result. Math.* Vol. 42: 212–228.

Altomare F., Cappelletti M., Leonessa V., Raşa, I. (2014). *Markov operators, positive semigroups and approximation processes*, De Gruyter Studies in Mathematics Vol. 61, Walter de Gruyter, Berlin-New York.

Altomare, F., Diomede S. (2010). Asymptotic formulae for positive linear operators: direct and converse results, *Jaen Journal on Approximation*, Vol. 2(2), 255–287.

Black F., Scholes, M. (1973). The pricing of options and corporate liabilities, *The journal of political economy*, Vol. 81, no. 3: 637–654.

Engel K.-J., Nagel, R. (2000). One-parameter semigroups for linear evolution equations, *Graduate Texts in Mathematics*, Vol. 194, Springer-Verlag, New York.

Hull, J. C. (2006). *Options, futures, and other derivatives*. Pearson Education India.

Margrabe W. (1978). The value of an option to exchange one asset for another, *Journal of Finance* Vol. 33: 177–186.

Merton R. (1973), Theory of rational option pricing, *Bell Journal of Economics and Management Science* Vol. 4: 141–183.

Orlando, G. and Taglialatela, G. (2017). A review on implied volatility calculation, *Journal of Computational and Applied Mathematics* Vol. 320: 202–220.

Wilmott, P., Howison S. and Dewynne J. (1995). The mathwematics of financial derivatives. A student Introduction. *Cambridge University Press*.

UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO *Metodi e analisi statistiche 2017*
DIPARTIMENTO DI ECONOMIA E FINANZA
ISBN: 978-88-6629-013-1
pp. 51-64

Il repricing gap nella valutazione del margine d'interesse

Mauro Bisceglia*

Università degli Studi di Bari Aldo Moro

Riassunto: Il *repricing gap* è in stretta dipendenza con il rischio dei tassi di interesse, a loro volta dipendenti dalla variabile *tassi di mercato* Δi_M, quindi una funzione composta definita sul mercato e con conseguenti ripercussioni sul margine di interesse (Resti e Sironi, 2008). Questo lavoro propone un'analisi della gestione del repricing gap, per poter meglio prevedere le performance del margine di interesse MI, di una banca, tenendo debitamente conto delle sue eventuali variazioni ΔMI. Il rischio dei tassi si ripercuote principalmente su tre fattori, quali le diverse *maturity*, lo squilibrio delle strutture tra attivo e passivo e la diversa incidenza sulle variazione dei tassi attivi e passivi, Δi_a e Δi_p. Il rischio dei tassi di interesse può essere osservato in base al rischio di prezzo o in base al rischio di reinvestimento: il repricing gap serve quindi alla valutazione di quest'ultimo. In tale lavoro si sono pertanto studiate le diverse metodologie di valutazione del gap, in presenza di maturity o di intervalli distinti, relativi alle attività e passività, sensibili alle variazioni del mercato, con l'opportunità di modificare le maturity, al fine di migliorare la valutazione del margine di interesse. Infine si suggerisce una ponderazione dei *gaps* marginali, che tiene conto delle vite a scadenza residue e del valore, in base al costo del mercato, degli stessi.

Keywords: repricing gap; margine d'interesse; rischio di rendimento; rischio di mercato; banche non quotate.

1. Introduzione

Il repricing gap contribuisce, da un punto di vista reddituale, ad una puntuale valutazione prospettica del margine di interesse in dipendenza di Δi_M, servendosi appunto del gap che si crea tra le attività e le passività finanziarie, sensibili al mercato

* Autore corrispondente: maurogianfranco.bisceglia@uniba.it.

$$G_t = AS_t - PS_t \Leftrightarrow G_t = \sum_{j=1}^{n} as_{t,j} - \sum_{i=1}^{m} ps_{t,i} \qquad [1]$$

dove G_t è il *Gap* relativo al periodo $[0, t]$ preso in esame, mentre AS_t e PS_t sono rispettivamente le attività e passività sensibili ovvero quelle che giungono a scadenza o che sono soggette a revisione di tasso di interesse nel corso del periodo $[0, t]$, $as_{t,i}$ rappresenta la *j-esima* attività sensibile e $ps_{t,i}$ la *i-esima* passività sensibile (Dermine e Bissada 2002).

Il margine di interesse, globale, relativo all'attività di intermediazione creditizia, risulta quale differenza tra gli interessi delle attività finanziarie, *IA* e quelli delle passività finanziarie, *IP*:

$$MI = IA - IP \Leftrightarrow MI = i_a A - i_p P \Leftrightarrow MI = i_a \left(AS + ANS\right) - i_p \left(PS + PNS\right)$$

da cui

$$\Delta MI = \Delta i_a AS - \Delta i_p PS\,.$$

La variazione ΔMI si basa sulla semplice considerazione che le Δi_M producono effetti sui tassi di rendimenti della banca, attivi e passivi, i_a e i_p, e conseguentemente sulle AS_t e PS_t.

In questa prima fase di studio, per comodità consideriamo

$$\Delta i_a = \Delta i_p = \Delta i$$

quindi

$$\Delta MI = \Delta i\left(AS - PS\right) \Leftrightarrow \Delta MI = \Delta i\left(\sum_{j=1}^{n} as_{t,j} - \sum_{i=1}^{m} ps_{t,i}\right) \Leftrightarrow \Delta MI = \Delta i \cdot G_t$$

Da quest'ultima si osserva come la variazione del margine di interesse può essere considerata (De Giuli *et al*. 2008) come *funzione* dei fattori gap e Δi, quindi $\Delta MI = \varphi(\Delta i, G_t)$.

Possiamo quindi sintetizzare le prime osservazioni sulle variazioni del margine di interesse in dipendenza delle sue variabili:

$$\Delta MI = \Delta i \cdot G_t \Leftrightarrow \Delta MI : \begin{cases} > 0 \Leftrightarrow \begin{cases} \Delta i > 0 \quad e \quad G_t > 0 \\ \qquad\quad o \\ \Delta i < 0 \quad e \quad G_t < 0 \end{cases} \\ = 0 \Leftrightarrow \Delta i = 0 \quad o \quad G_t = 0 \\ < 0 \Leftrightarrow \begin{cases} \Delta i > 0 \quad e \quad G_t < 0 \\ \qquad\quad o \\ \Delta i < 0 \quad e \quad G_t > 0 \end{cases} \end{cases}$$

2. Le vite residue a scadenza

La valutazione delle vite residue muove dal principio che la variazione del tasso di interesse produca i propri effetti solo nel periodo di vita residua, tempo compreso fra la data sensibile e la fine del *gapping period* (Fraser, Philips e Rose 1974). In particolare, nel caso di una qualunque attività sensibile *j-esima*, che frutta un tasso di interesse $i_{a,i}$ alla scadenza s_i produrrà un interesse attivo, relativo al futuro anno:

$$Ia_j = as_j \cdot i_{a,j} \cdot s_j + as_j \left(i_{a,j} + \Delta i_{a,j}\right)\left(1 - s_j\right).$$

Si nota come il rendimento di un'attività sensibile sia scindibile in due componenti, una certa, ovvero $as_j \cdot i_{a,j} \cdot s_j$, e una incerta, ovvero $as_j \cdot (i_{a,j} + \Delta i_{a,j})\ (1 - s_j)$, la quale rappresenta il valore atteso in funzione del mercato (Simonsons, Stowe e Watson 1983). Pertanto è naturale dedurre che la variazione degli interessi attivi sia determinata esclusivamente dalla componente incerta

$$\Delta Ia_j = as_j \Delta i_{a,j} \left(1 - s_j\right)$$

e considerando tutte le *n* attività di una banca, avremo che

$$\Delta IA = \sum_{j=1}^{n} as_j \Delta i_{a,j} \left(1 - s_j\right).$$

ed in modo analogo per le passività, avremo

$$\Delta Ip_i = ps_i \Delta i_{p,i} \left(1 - s_i\right)$$

e quindi per tutte le passività di una banca

$$\Delta IP = \sum_{i=1}^{m} ps_i \Delta i_{p,i} \left(1 - s_i\right).$$

Ipotizzando che le variazioni dei tassi di interesse attivi e passivi siano uniformi, ovvero

$$\Delta i_{a,j} = \Delta i_{p,i} = \Delta i\ ,\ \forall \left(j,i\right) \in \left\{1,2,...,n\left(m\right)\right\},$$

potremo stimare quindi la variazione del margine di interesse di una banca, come

$$\Delta MI = \Delta IA - \Delta IP \Leftrightarrow$$

$$\Leftrightarrow \Delta MI = \Delta i \left(\sum_{j=1}^{n} as_j \left(1 - s_j\right) - \sum_{i=1}^{m} ps_i \left(1 - s_i\right) \right) \Leftrightarrow \Delta MI \equiv \Delta i \cdot G^{MA} \quad ,$$

in cui G^{MA} rappresenta la vita residua ponderata, detta anche *maturity-adjusted gap*, dato dalla differenza fra attività e passività sensibili, ponderate in un gapping period, fissato a 1 anno (Saita, 2000); ponderazione che risulterebbe poco efficace se non fuorviante, riducendo il gap e rendendolo meno incisivo nella sua naturale espressione.

3. Gap marginali e cumulati

È utile pensare che lavorare sulle maturity delle attività sensibili è alquanto laborioso e complesso vista l'elevata numerosità delle stesse, pertanto ha senso raggruppare le maturity sensibili in specifici periodi futuri ed osservare quindi tali gap marginali. Pertanto, indichiamo con G'_{t_1} il primo gap marginale nel periodo $[0,t]$, si perviene al gap cumulato $G_{t_2} = G'_{t_1} + G'_{t_2}$ relativo al secondo sottoperiodo nell'intervallo di osservazione $[0,t]$. Trattandosi di gap periodale è necessario fare riferimento, con leggera imprecisione, ad una maturity del periodo, quale

$$t_j^* = \frac{t_j + t_{j-1}}{2}$$

maturity che consente la ponderazione del gap marginale secondo il principio del *maturity-adjusted gap*

$$\Delta MI \cong \Delta i \sum_{j \mid t_j \le 1} G'_{t_j}\left(1 - t_j^*\right) \Leftrightarrow \Delta MI \cong \Delta i G_1^W \qquad [2]$$

in cui G_1^W rappresenta il gap cumulato ponderato ad un anno. L'indicatore ottenuto come somma dei gap di periodo ponderati è un indicatore della sensibilità del margine di interesse a variazioni dei tassi di mercato e viene anche definito come *duration del margine di interesse*.

I gap marginali si prestano meglio alla valutazione del margine di interesse in presenza di un mercato bizzarro (Forestieri e Mottura 2009).

4. Ipotesi correttiva al repricing gap

Una parte del repricing gap che potrebbe essere rivista è il tipo di ponderazione utilizzata, la quale considera le poste sensibili proporzionate solo alla propria vita residua. Si comprende come in tal modo, il peso delle poste sensibili, resta limitato

nella sua espressione, rendendo poco efficace il gap stesso, e conseguentemente con ripercussioni sulla corretta valutazione del ΔMI, sulle eventuali politiche di immunizzazione da intraprendere.

Le singole poste sensibili hanno non solo effetto sulla ΔMI per la loro specifica grandezza, ma il tutto si ripercuote conseguentemente sul MI della banca in questione; pertanto sarebbe opportuno che ogni posta non subisse la sola correzione dovuta alla sua vita residua, ma che si tenesse conto anche della redditività futura di questa e non con la semplice variazione dei tassi, bensì con la struttura a scadenza di competenza, quindi in alternativa risulterebbe:

$$G'^{*}_{t_k} = \left(1 - t^{*}_{k}\right) G'_{t_k}\, m\left(t_0, t_k, t_m\right)$$

ed in presenza di un solo gap si avrebbe:

$$\Delta MI^{*} = \left(1 - t^{*}_{1}\right) G'_{t_1}\, m\left(t_0, t_1, t_m\right)$$

mentre nell'ipotesi di più gap marginali:

$$\Delta MI^{*} = \sum_{k=1}^{m} \left(1 - t^{*}_{k}\right) G'_{t_k}\, m\left(t_0, t_k, t_m\right).$$

dove $m\left(t_0, t_k, t_m\right)$ rappresenta appunto il valore del reinvestimento unitario a scadenza in base al rendimento risk free corrente (Castellani, De Felice e Moriconi 2005). Per cui da un paniere di titoli di Stato a breve, sapendo che

$$v\left(t_0, t_k\right) = \left[1 + j\left(t_0, t_k\right)\right]^{-1} \text{ o } v\left(t_0, t_k\right) = \left[1 + i\left(t_0, t_k\right)\right]^{-t_k}$$

Si ha (Fisher, 1965):

$$\left[1 + j\left(t_0, t_k\right)\right]^{-1} = \left[1 + i\left(t_0, t_k\right)\right]^{-t_k} \Leftrightarrow i\left(t_0, t_k\right) = \sqrt[t_k]{1 + j\left(t_0, t_k\right)} - 1$$

Pertanto, noto il rendimento unitario $i\left(t_0, t_k\right)$ costante per l'intero periodo $\left[t_0, t_k\right]$, e considerando che $m\left(t_0, t_k\right) = e^{\delta\left(t_k - t_0\right)}$, si ricava la relativa forza di interesse unitaria mensile δ_m

$$\left[1 + i\left(t_0, t_k\right)\right]^{\left(t_k - t_0\right)} = e^{\delta_m\left(t_k - t_0\right)} \Leftrightarrow \delta_m = \log\left(1 + i\left(t_0, t_k\right)\right).$$

Di conseguenza, dalla struttura dei *RendiBot 2016* (http://www.bancaditalia.it), relativa ad un intervallo identico a quello in esame, noto il rendimento $j(t_0,t_0,t_{12})$, si ricava $i(t_0,t_0,t_{12})$ e conseguentemente la forza di interesse mensile $\delta_m(t_0,t_0,t_{12})$ necessaria alla ponderazione dei gap marginali presenti nel primo intervallo $[t_0,t_1]$; mentre per i gap marginali relativi all'intervallo successivo $[t_1,t_2]$, l'utilizzo del rendimento $j(t_0,t_0,t_{12})$ non risponderebbe pienamente alla realtà del mercato cui il gap si riferisce, in quanto subentrerebbe anche il rendimento $j(t_0,t_1,t_{13})$, che interseca nell'intervallo in esame $[t_1,t_2]$ il rendimento $j(t_0,t_0,t_{12})$ e che ci sembra opportuno e corretto tenerne conto.

Ricavata la seconda forza di interesse $\delta_m(t_0,t_1,t_{13})$ che s'interseca con la precedente nell'intervallo $[t_1,t_{12}]$, ed essendo appunto unitaria, possiamo pensare a un'incidenza media tra le due, e così di seguito per i periodi di osservazione successivi. In generale, nota la struttura dei rendimenti $\{\delta_m(t_0,t_k,t_{m+k}),\forall k\in\{0,1,2,...,m-1\}\}$, possiamo determinare

$$\delta_m(t_0,t_k,t_m)=\frac{1}{k+1}\sum_{\gamma=0}^{k}\delta_m(t_0,t_\gamma,t_{m+\gamma}),\forall k\in\{0,1,2,...,m-1\}$$

necessaria alla rivalutazione dei gap marginali periodale relativi all'intervallo $[t_k,t_{k+1}]$. Il ragionamento fin qui condotto è replicabile per ogni ulteriore intervallo successivo.

In tal modo si viene a costruire la seguente struttura per scadenze di rendimenti, ovvero

$$\{\delta_m(t_0,t_0,t_{12}),\delta_m(t_0,t_1,t_{12}),\delta_m(t_0,t_2,t_{12}),...,\delta_m(t_0,t_{11},t_{12})\} \qquad [3]$$

necessaria alla ponderazione di ogni gap marginale del periodo in esame.

5. Indici in funzione del gap

Oltre servire nella valutazione futura del margine di interesse, il gap viene anche utilizzato nella determinazione di alcuni indici, quali:

1. *Indice di redditività del capitale*, che fornisce una misura della variazione della redditività del margine di interesse sui mezzi propri, in dipendenza delle variazioni del mercato.

$$\Delta\left(\frac{MI}{MP}\right)=\frac{G}{MP}\Delta i$$

2. *Indice di redditività della gestione* che fornisce una misura della variazione del margine di interesse sulle attività fruttifere, in dipendenza delle variazioni del mercato

$$\Delta\left(\frac{MI}{AF}\right) = \frac{G}{AF}\Delta i$$

3. *Gap ratio* che mette in relazione attività e passività sensibili, fornendo possibili confronti nel tempo e nello spazio fra banche, anche di dimensioni diverse.

$$gap\ ratio = \frac{AS}{PS}.$$

6. Problematiche legate al repricing gap

Il metodo del repricing gap, nonostante le sue versioni più accurate, presenta alcune criticità, di seguito descritte (Matz e Neu 2006).

6.1 L'ipotesi di variazioni uniformi dei tassi attivi e passivi e dei tassi di diversa scadenza

L'ipotesi fondamentale del modello è appunto l'uniformità delle variazioni dei tassi di interesse del mercato, sulle attività e sulle passività. Una possibile soluzione sarebbe quella di considerare la sensibilità di adeguamento in modo esplicito nel computo del gap, tramite l'identificazione del tasso di riferimento, la stima della sensibilità dei diversi tassi bancari attivi e passivi, quindi il computo del *gap corretto*. Per quanto detto

$$\Delta MI = \sum_{j=1}^{n} as_j \Delta i_{a,j} - \sum_{i=1}^{m} ps_i \Delta i_{p,i} \Leftrightarrow \Delta MI \cong \sum_{j=1}^{n} as_j \beta_j \Delta i_a - \sum_{i=1}^{m} ps_i \gamma_i \Delta i_p \Leftrightarrow$$

$$\Leftrightarrow \Delta MI \cong \left(\sum_{j=1}^{n} as_j \beta_j - \sum_{i=1}^{m} ps_i \gamma_i\right)\Delta i \Leftrightarrow \Delta MI \equiv G^s \Delta i$$

in cui β_j e γ_k rappresentano i coefficienti di sensibilità delle poste attive e passive, mentre G^s il *gap standardizzato*.

6.2 Il trattamento delle poste a vista

Si parla di poste a vista intendendo quelle poste attive o passive di cui la scadenza non è determinata. Secondo lo schema di suddivisione del *gapping period*, tali poste

andrebbero annoverate tra quelle sensibili il cui periodo di riferimento può anche essere quello giornaliero (Sakovich 2012). Da analisi empiriche è risultato che tali poste non si adeguino prontamente alle variazioni dei tassi di mercato; si è notato che l'adeguamento dei rendimenti delle poste a vista è asimmetrico. In questo caso una soluzione è possibile attraverso la stima dei ritardi medi per le diversi poste a vista all'adeguamento dei tassi rispetto all'istante in cui la variazione dei tassi si verifica, e generalmente si adopera l'analisi statistica dei dati passati (Resti e Sironi, 2008).

6.3 Omessa considerazione degli effetti delle variazioni dei tassi di interesse sulla quantità di fondi intermediati

Nel modello del repricing gap non si tiene conto dei valori stock, ma solo di quelli flusso, e cioè non si tiene conto di eventuali importi di *assets* o *liabilities* negoziate dalla banca (Wilson 1988). Muovendo dalla soluzione adoperata per l'ipotesi di variazioni uniformi dei tassi attivi e passivi, si possono modificare i coefficienti β e γ per tener conto dell'elasticità delle quantità rispetto ai prezzi. Nella pratica basta costruire β' come

$$\beta' = \beta(1 + x\%)$$

indicando con $x\%$ la percentuale relativa alla variazione dei volumi; lo stesso discorso può essere fatto per il coefficiente della sensibilità delle passività γ. In realtà però, anche la scelta di β' come funzione lineare di $x\%$ non sembra essere molto corretta, ma questo richiederebbe l'utilizzo di un modello econometrico sofisticato.

6.4 Omessa considerazione degli effetti di variazioni dei tassi sui valori di mercato

Un rialzo dei tassi di interesse non produce i suoi effetti unicamente sui flussi reddituali connessi alle attività e passività sensibili, ma anche sul valore stesso di tali poste. Per questo motivo il modello di repricing gap non è adatto a catturare gli impatti che variazioni di tasso possono avere sul valore degli assets, al contrario di un modello di tipo patrimoniale: il modello del duration gap.

7. Caso applicativo

Ai fini di una prima analisi di quanto trattato sono stati utilizzati dei dati di bilancio chiuso al 31/12/2015 di una BCC di Puglia e Basilicata, che per questioni di riser-

vatezza denoteremo come Banca Alpha. Il campo di previsione è riferito quindi all'anno 2016, sul seguente quadro di partenza dell'anno precedente

Tabella 1. *Situazione patrimoniale della Banca Alpha al 31/12/2015*

Attività sensibili 74.741.925	Passività sensibili 54.868.264
	Passività non sensibili 183.243.524
Attività non sensibili 186.507.716	Patrimonio netto 23.084921
Totale 261.249.641	Totale 261.249.641

Nello specifico, le attività e passività che matureranno la loro sensibilità nell'area del periodo di osservazione risultano così suddivise:

Tabella 2. *Suddivisione di attività e passività sensibili per data di scadenza*

	A vista	1-7gg	7-15gg	15-30gg	1-2 mesi	2-3 mesi	3-6 mesi	6-12 mesi
Attività	134.235.771	13.203.432	9.105.674	2.738.607	1.497.100	2.133.699	10.790.254	35.157.225
Passività	18.269.771	1.197.398	224.014	2.210.092	3.924.641	8.629.517	4.118.259	34.501.341

Tabella 3. *Suddivisione di attività e passività sensibili per data di riprezzamento*

	A vista	1-3 mesi	3-6 mesi	6-12 mesi
Attività	67.353	17.534	93.748	4.652
Passività	131.114	16.981	10.802	35.219

Quindi il gap calcolato secondo la $[1]$ relativo ai dati delle **Tab. 2** e **3**, risulta

$$G_t = AS_t - PS_t = \sum_{j=1}^{n} as_{t,j} - \sum_{i=1}^{m} ps_{t,j} = 74.741.925 - 54.868.264 = 19.873.661,$$

escludendo le poste a vista, le quali ammontano rispettivamente a 134.303.124 e 18.400.885.

7.1 Calcolo del gap dopo la ridistribuzione delle poste vista

Al fine di superare il problema del trattamento delle poste a vista, viene stimata per ognuna di esse, la struttura dei ritardi medi di adeguamento dei tassi rispetto al momento in cui si verifica una variazione dei tassi di mercato. Si assume quindi che il coefficiente complessivo di sensitività[1] al tasso Euribor a 3 mesi risulta pari all'80%; e questi 8000 p.b. sono così distribuiti:

[1]Fonte: http://www.bancaditalia.it

Tabella 4. *Ridistribuzione progressiva dei depositi a vista*

Orizzonte temporale	Variazione percepita	Riallocazione attività	Riallocazione passività
A vista	0%	/	/
A 1 mese	10%	13.430.312	1.826.977
A 3 mesi	50%	67.151.562	9.134.886
A 6 mesi	12%	16.116.375	2.192.373
A 1 anno	8%	10.744.250	1.461.582
Totale	80%	107.442.499	14.615.817

da cui la nuova composizione delle poste sensibili

Tabella 5. *Suddivisione attività sensibili a seguito della rivalutazione delle poste a vista attive*

	1-7gg	7-15gg	15-30gg	1-2 mesi	2-3 mesi	3-6 mesi	6-12 mesi
per data di scadenza	13.203.432	9.105.674	2.738.607	1.497.100	2.133.699	10.790.254	35.157.225
data di riprezzamento	/			17.534		93.784	4.652
Poste a vista	13.430.312			67.151.562		16.116.375	10.744.250

Tabella 6. *Suddivisione passività sensibili a seguito della rivalutazione delle poste a vista passive*

	1-7gg	7-15gg	15-30gg	1-2 mesi	2-3 mesi	3-6 mesi	6-12 mesi
per data di scadenza	1.197.398	224.014	2.210.092	3.924.641	8.629.517	4.118.259	34.501.341
data di riprezzamento	/			16.981		10.802	35.219
Poste a vista	1.826.977			9.134.886		2.192.373	1.461.582

e pertanto $AS_t = \sum_{j=1}^{n} as_{t,j} = 182.184.460$ e $PS_t = \sum_{i=1}^{m} ps_{t,i} = 69.484.082$.

La ridistribuzione delle poste a vista modifica lo scenario patrimoniale della banca

Tabella 7. *Situazione patrimoniale post ridistribuzione*

Attività sensibili 182.184.424	Passività sensibili 69.484.082
	Passività non sensibili 23.084.921
Attività non sensibili 79.065.217	Patrimonio netto 23.084921
Totale 261.249.641	Totale 261.249.641

Possiamo quindi osservare il nuovo Gap a seguito della forte distribuzione delle masse a vista:

$$G_t = AS_t - PS_t = \sum_{j=1}^{n} as_{t,j} - \sum_{i=1}^{m} ps_{t,i} \Leftrightarrow$$

$$\Leftrightarrow G_t = 182.184.424 - 69.484.082 = 112.700.342$$

Se ipotizziamo una $\Delta i_a = \Delta i_p = \Delta i = 1\%$, la variazione positiva del margine di interesse risulterà

$$\Delta MI = \Delta i(AS - PS) \Leftrightarrow \Delta MI = \Delta i G_t \Leftrightarrow \Delta MI = 0.01 \cdot 112.700.342 = 1.127.003,42$$

superiore di oltre cinque volte rispetto a quella che si avrebbe senza le poste a vista.

7.2 Ponderazione dei gap marginali

Per una migliore lettura dell'esposizione al rischio dei tassi di mercato della banca è opportuno analizzare i gap relativi alle diverse scadenze.

Quindi si calcolano le scadenze medie $\left(t_j^*\right)$ che saranno considerate quali maturity delle poste sensibili nell'intervallo e quindi si pondera il gap marginale $G_t^{'}$, secondo la $[2]$, così come riportato in **Tab. 8**.

Tabella 8. *Calcolo del gap cumulato ponderato*

(t_{j-1}, t_j)	G_t'	G_t	$t_j - t_{j-1}$	t_j^*	$1 - t_j^*$	$G_t'(1 - t_j^*)$
1-7 giorni	12.006.034	12.006.034	(7-1)/360	4/360	356/360	11.872.633,62
7-15 giorni	8.881.660	20.887.694	(15-7)/360	11/360	349/360	8.610.275,94
15-30 giorni	528.515	21.416.209	(30-15)/360	22,5/360	337,5/360	495.482,81
0-1 mese	11.603.335	33.019.544	(1-0)/12	1/24	23/24	11.119.862,71
1-2 mesi	-2.427.541	30.592.003	(2-1)/12	1,5/360	10,5/12	-2.124.185,88
2-3 mesi	-6.495.818	24.096.185	(3-2)/12	2,5/12	11/12	-5.954.499,83
1-3 mesi	58.017.229	82.113.414	(3-1)/12	2/12	10/12	48.347.690,83
3-6 mesi	20.678.943	102.792.357	(6-3)/12	4,5/12	7,5/12	12.924.339,38
6-12 mesi	9.907.985	112.700.342	(12-6)/12	9/12	3/12	2.476.996,25

$$\Delta MI \cong \Delta i \left(\sum_{j \mid t_j \le 1} G_{t_j}^{'} \left(1 - t_j^*\right) \right) \Leftrightarrow \Delta MI \cong \Delta i G_t^W \Leftrightarrow$$

$$\Leftrightarrow \Delta MI \cong 0.01 \cdot 87.768.595,83 = 877.685,96 .$$

7.3 Ipotesi alternativa di ponderazione

Per poter mettere in atto tale ulteriore correzione è necessario disporre dei tassi di rendimento privi di rischio relativi al periodo in esame; pertanto ci serviamo della

seguente struttura dei tassi a pronti su base annua, relativa ai rendimenti medi dei BOT a breve termine, come da pubblicazione del *Rendistato anno 2016 della Banca d'Italia*[2], dato che i dati della Banca Alpa sono fotografati al 31/12/2015

$$\left\{j(t_0,t_0,t_{12}), j(t_0,t_1,t_{13}), j(t_0,t_2,t_{14}),...,j(t_0,t_{11},t_{23})\right\},$$

ovvero

$$\{-0.141;-0.110;-0.141;-0.191;-0.255;-0.187;-0.264;-0.276;-0.308;-0.339;-0.314;-0.370\}$$

Quindi, considerando quanto trattato nel precedente paragrafo 4, si ottiene la relativa struttura per scadenza delle intensità istantanee d'interesse

$$\left\{\delta_m(t_0,t_0,t_{12}), \delta_m(t_0,t_1,t_{12}), \delta_m(t_0,t_2,t_{12}),...,\delta_m(t_0,t_{11},t_{12})\right\},$$

ovvero

$$\begin{aligned}&\{-0{,}000051;-0{,}0000454;-0{,}0000473;-0{,}0000527;-0{,}0000607;-0{,}0000618;\\&-0{,}0000667;-0{,}0000708;-0{,}0000754;-0{,}0000802;-0{,}0000832;-0{,}0000874\}\end{aligned} \quad [3^*]$$

e tenendo conto delle le scadenze medie $t_k^* = \frac{t_k + t_{k-1}}{2}$ quali maturity, la ponderazione alternativa da noi proposta risulta

$$G_{t_k}^{'*} = (t_m - t_{k-1}^*) G_{t_k}^{'} m(t_0,t_{k-1},t_m) \Leftrightarrow G_{t_k}^{'*} = (t_m - t_{k-1}^*) G_{t_k}^{'} e^{\delta_m(t_0,t_{k-1},t_m)(t_m - t_k^*)}$$

Tabella 9. *Ponderazione dei gap marginali secondo la logica dai noi proposta*

Periodo	$G_t^{'}$	t_k^*	$1-t_k^*$	$\delta_m(t_0,t_k,t_m)$	$G_{t_k}^{'*}$
1-7 gg	12.006.034,00	7+1/2	(360-4)/30	-0,000051	11.998.768,11
7-15 gg	224.014,00	15+7/2	(360-11)/30	-0,000051	8.876.393,58
15-30 gg	528.515,00	30+15/2	(360-22,5)/30	-0,000051	528.221,85
0-1 m	11.590.223,90	1+0/2	12-0,5	-0,000051	11.596.531,64
1-2 m	-2.427.541,00	2+3/2	12-1,5	-0,0000454	-2.426.484,01
2-3 m	-6.495.818,00	3+2/2	12-2,5	-0,0000473	-6.492.899,76
1-3 m	57.951.672,50	3+1/2	12-2	-0,0000473	57.989.793,34
3-6 m	20.663.209,68	6+3/2	12-4,5	-0,0000618	20.669.396,51
6-12 m	106.595.745,40	12+6/2	12-9	-0,0000874	9.905.387,47

[2] Fonte: http:/www.bancaditalia.it

$$\Delta MI^{*} \cong \Delta i \sum_{k=1}^{m} \left(t_m - t_{k-1}^{*}\right) G_{t_k}^{'} m\left(t_0, t_k, t_m\right) \Leftrightarrow$$

$$\Leftrightarrow \Delta MI^{*} \cong 0.01 \cdot 112.645.098{,}70 = 1.126.450{,}98$$

Tabella 10. *Quadro riepilogativo*

	senza poste a vista	con poste a vista	con ponderazione	ponderazione alternativa
G_t	19.873.661	112.700.342	87.768.595,83	112.645.099
ΔMI	198.736,61	1.127.003,42	877.685,96	1.126.450,99
$\Delta(MI/MP)$	0,00861	0,0488	0,0380	0,0479
$\Delta(MI/AF)$	0,000793	0,004497	0,0035	0,004495
AS/PS	1,36	2,62	2,62	2,62

8. Conclusioni

L'analisi fin qui condotta fornisce, per le opportune considerazioni, il seguente quadro riepilogativo di repricing gap, il quale potrebbe essere facilmente ampliato su un campione di Banche omogenee.

Un'ulteriore considerazione va fatta, in merito alla struttura di tassi attualmente presente sul mercato, che come si è potuto osservare presenta valori negativi, e questo restituisce un'apparente impoverimento delle considerazioni fatte a livello teorico. In ogni caso non perde di significatività il confronto fatto tra le due diverse ponderazioni, vista la divergenza di risultati, e conseguentemente con previsioni sulla valutazione del margine d'interesse e sulle opportune decisioni inerenti il risk management. Problema che si accentua in maniera ancora più significativa nei periodi caratterizzati da un'alta volatilità dei tassi di interesse ed in particolar modo, in presenza di variazione a ribasso.

Riferimenti bibliografici

Castellani, G.; De Felice, M.; Moriconi, F. (2005). *Manuale di finanza vol.1*. Il Mulino, Bologna.

De Giuli, M. E.; Giorgi, G.; Maggi, M.; Magnani, U. (2008). *Matematica per l'economia e la finanza*. Zanichelli, Bologna.

Dermine, J.; Bissada, Y. F. (2002). *Asset and liability management a guide to value creation and risk control*. Pearson education limited, London.

Fisher, I. (1930). *The theory of interest*. Macmillan, New York; tard. Ital. (1974). *Teoria dell'interesse*, in *Opere di Irving Fisher*, a cura di Pelladana A. Utet, Torino.

Forestieri, G.; Mottura, P. (2009). *Il sistema finanziario,* quinta edizione. Egea, Milano.

Fraser, D. R.; Philips, W.; Rose, P. S. (1974). A Canonical Analysis of Bank Performance. *The Journal of Finance and Quantitative Analysis*, **9**: 287-295.

Matz, L.; Neu, P. (2006). *Liquidity Risk, Measurement and Management*. Wiley, Stati Uniti.

Resti, A.; Sironi, A. (2008). R*ischio e valore nelle banche, Misura, regolamentazione, gestione*. Egea, Milano.

Saita, F. (2000). *Il risk mamagement in banca. Performance corretta per il rischio e allocazione del capital*. Egea, Milano.

Sakovich, M. (2012). Asset-Liability Management in banking as an instrument for minimization of expenses in the implementation of Basel III requirements. Available at SSRN: http://dx.doi.org/10.2139/ssrn.2189606.

Simonson, D.; Stowe, J.; Watson, C. (1983). A Canonical Correlation Analysis of Commercial Bank Asset/Liability Structures. *The Journal of Financial and Quantitative Analysis*, **18**: 125-140 (https://doi.org/10.2307/2330808).

Wilson, J. S. G. (1988). *Managing bank assets and liabilities*. Euromoney publications, London.

Sitografia

http:/www.bancaditalia.it

http:/www.federpb.bcc.it

UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO *Metodi e analisi statistiche 2017* ISBN: 978-88-6629-013-1
DIPARTIMENTO DI ECONOMIA E FINANZA pp. 65-101

Previsione del rischio a partire da report finanziari mediante l'utilizzo di modelli per topic latenti

Massimo Bilancia[1*], Gianluca Novembre[2]

[1]*Dipartimento Jonico in "Sistemi Giuridici ed Economici del Mediterraneo: società, ambiente, culture", Università degli Studi di Bari Aldo Moro. Taranto (TA – IT),*
[2]*Apulia Distribuzione s.r.l., Ufficio Marketing, Rutigliano (BA – IT).*

Riassunto: La volatilità dei rendimenti di un asset finanziario è generalmente considerata come una misura del rischio associato a tale attività. In questo lavoro ci concentreremo su un approccio atipico per la previsione della volatilità, basato su un insieme di algoritmi che permettono di risolvere problemi di classificazione testuale. A tale fine, abbiamo utilizzato un corpus messo a disposizione dal Center for Research in Security Prices (CRSP), contenente un insieme di report il cui modello di riferimento è noto come *Form 10-K*, la cui compilazione è richiesta annualmente dalla SEC (U.S. Securities and Exchange Commission) alle aziende quotate le cui attività circolanti siano superiori a 10 milioni di dollari. Abbiamo messo a confronto le performance previsive di due classificatori testuali, utilizzando una versione discretizzata della volatilità come target previsivo. Il primo classificatorte è il *modello di regressione logistica multinomiale penalizzato*, che utilizza una particolare norma sullo spazio dei parametri per risolvere il problema dell'overfitting, spingendo verso lo zero la maggior parte dei coefficienti di regressione. Il secondo è il modello *Supervised Latent Dirichlet Allocation* (SLDA), che permette di classificare un corpus testuale sulla base di un insieme di topic latenti (ossia di contenuti tematici) che vengono identificati all'interno di ciascuno documento. I risultati ottenuti indicano una leggera prevalenza del modello di regressione logistica multinomiale, prevalenza che si rafforza quando si considera che i tempi di calcolo richiesti sono nettamente inferiori a quelli del modello SLDA.

Keywords: Rischio finanziario; Modelli per topic latenti; Latent Dirichlet Allocation (LDA); Supervised Latent Dirichlet Allocation (SLDA).

---

[*] Autore corrispondente: massimo.bilancia@uniba.it.
Gli autori hanno collaborato in parti uguali alla stesura del presente saggio, che è stato oggetto di revisione anonima a doppio cieco.

1. Introduzione

L'obiettivo generale di questo lavoro consiste nella previsione della volatilità dei rendimenti di azioni quotate, utilizzando un approccio atipico basato sulla classificazione testuale di un corpus di documenti, contenenti un riassunto delle performance economiche, in un determinato anno d'esercizio, di aziende che hanno emesso azioni o altri strumenti rappresentativi del capitale.

La volatilità dei rendimenti di un'azione quotata, è generalmente considerata come una misura del rischio associato a tale attività finanziaria. Se definiamo il rendimento r_t come la variazione relativa tra il prezzo di chiusura P_t al giorno di trading $t-1$ e il prezzo di chiusura al giorno di trading t:

$$r_t = \frac{P_t}{P_{t-1}}, \tag{1}$$

la *volatilità campionaria* misurata su una finestra temporale avente lunghezza pari a τ giorni di trading è uguale alla deviazione standard dei rendimento nello stesso periodo:

$$\upsilon_{[t-\tau,t]} = \sqrt{\frac{1}{\tau}\sum_{i=0}^{\tau}(r_{t-i} - \overline{r})}, \tag{2}$$

dove $\overline{r}$ indica la media campionaria dei rendimenti nel periodo considerato. La volatilità è direttamente collegata all'ampiezza delle fluttuazioni dei redimenti (Bouchaud and Potters, 2003). Se le quotazioni di una determinata azione sono molto volatili, sarà possibile registrare ampie escursioni nei livelli dei rendimenti, che potranno garantire guadagni in conto capitale elevati, così come perdite di livello elevato (e quindi, maggiore è la volatilità, maggiore è il *rischio*).

Sulla base dell'*ipotesi dei mercati efficienti* (EMH, *Efficient Market Hypothesis*), non è possibile realizzare rendimenti superiori a quelli del marcato utilizzando l'informazione pubblicamente disponibile (Fama, 1970). Sotto certe condizioni, l'ipotesi dei mercati efficienti afferma che la previsione dei rendimenti, se possibile, sarebbe immediatamente messa in atto dai mercati, e quindi il prezzo (quotazione) finirebbe per scontare immediatamente tutte le nuove informazioni, rendendo impossibile qualsiasi possibilità di arbitraggio. Nella sua forma debole, l'EMH postula che sia impossibile formulare strategie di trading con un rendimento superiore a quello del mercato, utilizzando come set informativo i soli prezzi di chiusura passati. Se l'EMH in senso debole è vera, è impossibile prevedere correttamente i ren-

dimenti futuri utilizzando sistemi comportamentali come l'analisi tecnica, oppure tutti i modelli che trattano l'evoluzione dei rendimenti come un processo stocastico a parametro discreto o continuo per il quale sia possibile esplicitare un previsore lineare di minima varianza. Senza entrare nei dettagli sul dibattitto sottostante a queste problematiche, e sulla possibilità che esistano o meno anomalie che permettano che inficino le ipotesi che sono alla base dell'EMH (almeno per periodi limitati; Foye et al., 2013), è invece interessante sottolineare che la previsione della volatilità sulla base di dati storici (e quindi del rischio) è universalmente considerata possibile. Ciò in quanto la volatilità ha delle proprietà statistiche universali, che possono essere opportunamente modellate ed utilizzate per migliorare le previsioni dei rendimenti sui mercati che, altrimenti, nella maggior parte dei casi continuano a dimostrarsi altamente impredicibili. L'approccio standard a questo modo di procedere è basato sulla costruzione di processi stocastici ad hoc che incorporino queste caratteristiche, e che al tempo stesso siano trattabili dal punto di vista matematico: da questo punto la classe dei modelli ARCH e GARCH, e tutti i modelli da questi derivati, riveste un ruolo estremamente importante nella finanza contemporanea (Cumby et al., 1993; Poon e Granger, 2003).

In questo lavoro perseguiremo, invece, un approccio del tutto differente, basato sulla *classificazione testuale*. In sintesi, un problema di classificazione testuale è basato su un *corpu*s di documenti testuali D, ed un insieme di categorie predefinite (etichette) $\Xi = \{1,\ldots,C\}$, con $C \geq 2$, tali che a documento possa essere attribuita una ed una sola etichetta. In questo contesto, un classificatore è una funzione $\gamma : D \rightarrow \Xi$, che viene appresa sulla base del corpus D mediante un algoritmo appropriato, ed è utilizzata per classificare documenti futuri non disponibili nella fase di apprendimento, ossia per attribuire a ciascuno di questi documenti futuri un'unica etichetta dell'insieme Ξ (Sebastiani, 2002; Manning et al., 2008).

Per trasformare il problema della previsione della volatilità in un problema di classificazione testuale, abbiamo utilizzato un corpus di test, messo a disposizione dal Center for Research in Security Prices (CRSP), contenente un insieme di report no- ti come Form 10-K[1], la cui compilazione è richiesta annualmente dalla SEC (U.S. Securities and Exchange Commission) alle aziende quotate le cui attività circolanti siano superiori a 10 milioni di dollari (Kogan et al., 2009). Ogni report è un documento testuale che contiene una serie di sezioni specializzate dedicate, per esempio, alla storia dell'azienda, alla sua struttura organizzativa e alla struttura finanziaria. Tra le altre sezioni, quella che utilizzeremo nello specifico contiene una

[1] https://www.sec.gov/fast-answers/answers-form10khtm.html

serie di '*forward-looking statements*', che sulla base delle informazioni correnti stimano le aspettative, e proiettano in avanti le performance dell'azienda all'interno dei mercati sui quali essa opera.

Per quanto riguarda la volatilità, a ciascun report disponibile è associata la quantità $v^{(+12)}$, detta *volatilità forward*. Tale misura è stata calcolata e resa disponibile da Kogan et al. (2009), e coincide con la volatilità campionaria (2) misurata su una finestra temporale pari ai 12 mesi successivi alla pubblicazione del report. La previsione della volatilità non è effettuata direttamente sulla variabile continua sottostante, ma questa viene opportunamente discretizzata in due o più livelli: per esempio, con una discretizzazione a due classi la previsione riguarda la distinzione naturale tra un livello elevato di volatilità, caratterizzato da ampie fluttuazioni, e un livello basso durante il quale il mercato presenta, tipicamente, un movimento laterale dei prezzi e scarse escursioni.

Un modello di base per prevedere l'etichetta (nel nostro caso, il livello di volatilità) associata a ciascun documento di test, è la *regressione logistica multinomiale* (Hastie et al., 2009; James et al., 2013). In questo setting, ciascun documento del corpus di apprendimento viene innanzitutto tokenizzato, ossia ridotto ad un insieme di token che prenderà il nome di vocabolario dei termini V. Le fasi di preprocessing che portano alla costruzione di V saranno illustrate con maggior dettaglio dettaglio nel Paragrafo 4. La regressione logistica multinomiale è un classificatore discriminativo (Ng e Jordan, 2001), nel quale le probabilità a posteriori delle etichette sono apprese, documento per documento, direttamente dal modello sulla base di una opportuna funzione di un insieme di variabili previsive (insieme che prende anche il nome di *feature vector*). Nello specifico, per ciascun documento, il feature vector contiene le relative frequenze di occorrenza di ciascun token del vocabolario dei termini. Poiché in applicazioni realistiche l'ordine di grandezza di V è pari almeno a 10^2 token, il modello di regressione appreso dal corpus di training sarà generalmente molto sparso, con molti coefficienti specifici associati ai token prossimi a zero tanto in valore assoluto, quanto sulla scala definita dal relativo errore standard. Per evitare ovvi problemi di overfitting, la soluzione generalmente accettata (e che seguiremo in questo lavoro) è quella di penalizzare le stime rispetto ad un certa norma definita sullo spazio dei parametri, in modo che il vincolo complessivo su tale norma spinga verso lo zero la maggior parte delle stime dei parametri (e quindi la maggior parte dei token cesseranno di essere influenti nel determinare la previsione delle etichette; Friedman et al, 2010).

L'idea che le sole frequenze di occorrenza dei token siano rilevanti per prevedere le etichette è stata utilizzata in modelli più elaborati, che come ipotesi di base si

fondano sull'idea che la probabilità di occorrenza di un token all'interno di un documento non dipende dalla posizione nella quale il token compare. In questo caso si parla di *modello unigram*, e il modello multinomiale prodotto è la formalizzazione naturale del modello unigram (Manning et al., 2008). In esso, la verosimiglianza dipende dai token solo attraverso le relative frequenze di occorrenza nei documenti. Se abbiamo un insieme di etichette, ciascuna delle quali rappresenta uno specifico contenuto tematico di un documento, possiamo estendere il modello unigram in modo non-supervisionato definendo un *miscuglio di unigram*, nel quale la probabilità condizionale di occorrenza di ciascun token data l'etichetta è ancora descritta tramite una distribuzione multinomiale definita su V, e l'insieme delle etichette è descritto da una distribuzione di probabilità discreta (Nigam et al., 2000). In questo caso, il significato delle etichette non è disponibile a priori, né è noto il numero di etichette. Una volta appreso il modello, tutto quello che siamo in grado di fare è di partizionare i documenti nelle classi a disposizione, ossia attribuire a ciascun documento una ed una sola etichetta (*model-based hard clustering*) sulla base delle stime delle probabilità a posteriori delle etichette stesse.

Lo sviluppo più importante in questa classe di modelli è la specificazione nota come *Latent Dirichlet Allocation* (LDA), introdotta in Blei at al. (2003), assieme ad un metodo di stima dei parametri ad-hoc particolarmente efficiente dal punto di vista computazionale. Se nell'approccio non-supervisionato possiamo identificare ciascuna etichetta con il contenuto tematico univoco del documento (ad esempio sport, politica, scienza, e così via), nel modello LDA è permessa la coesistenza di più contenuti tematici in uno stesso documento, contenuti che prendono il nome di *topic*. La probabilità di occorrenza di ciascun token, pur continuando a non dipendere dalla posizione occupata nel documento, è un miscuglio di multinomiali le cui probabilità dipenderanno dal topic che genera il particolare token che stiamo considerando. In questo modo, token distinti possono essere generati da topic diversi (non esiste un contenuto tematico globale), e l'identificazione dei topic latenti diventa un potente strumento di riduzione della dimensionalità del corpus di documenti che stiamo studiando.

Infine, quando abbiamo un'etichetta che, come nel nostro caso, non rappresenta un contenuto tematico latente, ma una informazione ausiliaria disponibile accanto a ciascun documento (ossia la volatilità forward discretizzata, ed associata a ciascun Form), possiamo definire un modello che cerca di prevedere le etichette non solo attraverso le frequenze di occorrenza dei token, bensì anche attraverso il contenuto semantico latente espresso attraverso i topic. Il modello risultante è la versione supervisionata del modello LDA, conosciuto appunto come *Supervised Latent Diri-*

chlet Allocation (SLDA; Blei e McAuliffe, 2007; Zhang e Kjellström, 2015). Considerando che l'algoritmo variazionale utilizzato per la stima a posteriori dei parametri del modello SLDA ha un importante costo computazionale (Blei et al., 2017), in questo lavoro vogliamo confrontare, in termini di accuratezza previsiva su un insieme di test, quest'ultimo modello con la regressione logistica multinomiale, il cui algoritmo di stima è invece enormemente meno esigente in termini di tempo di calcolo e di allocazione di memoria. La questione di ricerca è se un eventuale guadagno in termini di accuratezza da parte del modello SLDA (rispetto alla regressione logistica multinomiale), possa essere giustificato sulla base del costo computazionale aggiuntivo richiesto.

Per quanto detto, il presente lavoro è strutturato come segue. Data la sua importanza, nel Paragrafo 2 passeremo in rassegna in profondità la struttura gerarchica del modello LDA, discutendo ad un certo livello di dettaglio anche il relativo algoritmo di stima. Nello stesso paragrafo presenteremo anche un esempio di utilizzo del modello LDA in ottica non-supervisionata, ove l'interesse risiede univocamente nell'identificazione dei topic latenti. Il Paragrafo 3 è invece dedicato a richiamare brevemente la specificazione e le caratteristiche tanto del modello di regressione logistica multinomiale, quanto del modello SLDA. Infine, il Paragrafo 4 descrive le fasi di pre-processing del corpus dei documenti, ed illustra i risultati ottenuti. Le conclusioni del lavoro, e le possibili direzioni per le ricerche future, sono brevemente descritte nel Paragrafo 5.

2. Latent Dirichlet Allocation (LDA)

Come abbiamo detto, il modello descritto in questa sezione ha natura non supervisionata, ed è stato introdotto con lo scopo di permettere l'estrazione dei topic latenti presenti in collezioni di documenti testuali. Data la sua importanza, nonché la particolarità dei metodi di inferenza che sono stati introdotti per la stima dei parametri (metodi che sono generalizzabili all'approccio supervisionato che discuteremo nel Paragrafo 3), ne daremo una trattazione approfondita degli elementi essenziali, rinviando il lettore alla letteratura per i dettagli più sottili.

Introduciamo innanzitutto le definizioni rilevanti, ricalcando la notazione originale introdotta in Blei et al. (2003). Come già detto in precedenza, un token è un *item* da un vocabolario dei termini V, i cui elementi sono indicizzati dall'insieme $\{1,\ldots,V\}$. Ogni token è rappresentato da un vettore unitario $|V|$-dimensionale; ad esempio, il token il cui indice nel vocabolario è v, sarà rappresentato mediante il

vettore w le cui componenti, indicate tramite apici sono, rispettivamente, $w^v = 1$ e $w^u = 0$ per $u \neq v$. L'indice associato a ciascun token è arbitrario, e qualsiasi permutazione di $\{1,\ldots,V\}$ risulterà essere altrettanto valida. Con queste ipotesi, l'occorrenza di ciascun singolo token è una variabile aleatoria che si presta ad essere descritta in modo naturale da una distribuzione multinomiale su una singola prova: le relative probabilità multinomiali determinano la frequenza di occorrenza di ciascun token all'interno di un dato documento, $w = (w_1,\ldots,w_N)$, visto come uno stream ordinato di N token dal vocabolario V.

A priori, non ci sono motivi per considerare uguali due documenti che differiscono solo per l'ordine dei token. In molti casi, tuttavia, l'ordine non è particolarmente rilevante per decodificare la semantica del testo. Come vedremo subito, questa caratteristica è direttamente incorporata all'interno del modello LDA. Infine, supporremo di avere a disposizione un corpus di D documenti, ciascuno dei quali è, sulla base della notazione introdotta, è uno stream ordinato di N_d token. Altre ipotesi preliminari che sono necessarie a definire il modello sono le seguenti:

- in ciascun documento sono presenti K temi semantici latenti (*topic*), indicati con $\beta_{1:K} = (\beta_1,\ldots,\beta_K)$. Ciascun elemento di $\beta_{1:K}$ è una distribuzione di probabilità discreta $|V|$-dimensionale a supporto sugli elementi di V. Di conseguenza, il generico elemento di $\beta_{1:K}$ è il vettore $\beta_i = (\beta_{i1},\ldots,\beta_{i|V|})$. Si noti che $\beta_{1:K}$ non varia al variare di d nel corpus D, ed è isomorfo ad una matrice di dimensione $K \times |V|$;
- per ciascun documento $d \in D$, $z_{d,1:N_d} = (z_{d,1},\ldots,z_{d,N_d})$ è un vettore di vettori unitari K-dimensionali, ognuno dei quali indica il topic che ha generato il token $w_{d,n}$ (per $d = 1,\ldots,D$, $n = 1,\ldots,N_d$). In altre parole, la generica variabile indicatrice z, che descrive l'occorrenza dell'i-esimo topic, è rappresentata da un vettore unitario tale che $z^i = 1$ e $z^s = 0$ per $s \neq i$. Facendo variare l'indice d sui documenti, ciascun vettore $z_{d,1:N_d}$ può essere raccolto nel vettore $z_{1:D}$, che contiene le variabili indicatrici dei topic per qualsiasi token di ogni documento $d \in D$;
- con la stessa notazione, un generico documento verrà rappresentato tramite il vettore $w_{d,1:N_d} = (w_{d,1},\ldots,w_{d,N_d})$, mentre $w_{1:D}$ indica tutti i token del corpus raccolti in un unico vettore.

La variabile latente z indicatrice del topic seleziona in modo univoco una ed una sola distribuzione di probabilità $\beta_{1:K}$ nel modo seguente:

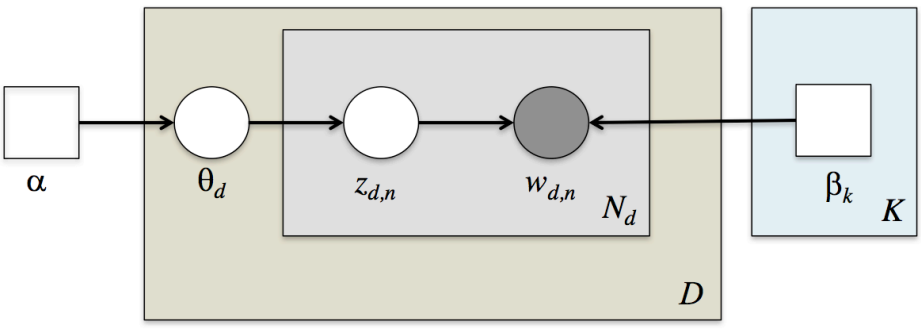
$$\beta_{z_{d,n}} \equiv \beta_i \quad \text{quando} \quad z^i_{d,n} = 1,\ n \in \{1,\ldots,N_d\}, \tag{3}$$

e pertanto $\beta_{ij} = \Pr\{w^i_{d,n} = 1 \mid z^i_{d,n} = 1\}$. Con questa notazione, il modello LDA è basato sulla seguente modello gerarchico generativo, valido per ciascun documento indipendentemente l'uno dall'altro:

- generiamo le proporzioni dei topic da una distribuzione di Dirichlet simmetrica K-dimensionale: $\theta_d \mid \alpha \sim \text{Dirichlet}_K(\alpha)$;
- per ciascun token $w_{d,n}$, per $n = 1,\ldots,N_d$, ed indipendentemente l'uno dall'altro:
  - generiamo il topic $z_{d,n}$ da una distribuzione Multinomiale K-dimensionale, con probabilità multinomiali dipendenti da θ_d, ossia: $z_{d,n} \mid \theta_d \sim \text{Multinomial}_{|K|}(1;\theta_d)$;
  - generiamo il token $w_{d,n}$ da una distribuzione Multinomiale $|V|$-dimensionale, con probabilità multinomiali dipendenti da $z_{d,n}$ e $\beta_{1:K}$, ossia: $w_{d,n} \mid z_{d,n}, \beta_{1:K} \sim \text{Multinomial}_{|V|}(1;\beta_{z_{d,n}})$.

Graficamente, il modello può essere rappresentato come nella Figura 1. In questa rappresentazione basata su un *grafo orientato*, dobbiamo distinguere tra *nodi stocastici* (indicati con cerchi), ossia tutte quelle variabili ai quali è associata una distribuzione di probabilità (la verosimiglianza sui dati e le distribuzioni di probabilità a priori sulle variabili latenti), e *nodi deterministici* (indicati con quadrati), ossia quegli *iper-parametri* che, sebbene incogniti, non sono stati dotati di una distribuzione di probabilità a priori.

Figura 1. *Rappresentazione del modello LDA sotto forma di grafo orientato. I parametri α e $\beta_{1:K}$ sono trattati come iper-parametri incogniti da stimare sulla base dei dati, piuttosto che come nodi stocastici dotati di una distribuzione di probabilità a priori.*



La specificazione gerarchica che abbiamo introdotto corrisponde, evidentemente, alla seguente distribuzione congiunta sulle quantità osservabili (occorrenze dei token) e sulle variabili latenti:

$$p(\theta_{1:D}, z_{1:D}, w_{1:D} \mid \alpha, \beta_{1:K}) = \prod_{d=1}^{D} p(\theta_d \mid \alpha) \left[\prod_{n=1}^{N_d} p(z_{d,n} \mid \theta_d) p(w_{d,n} \mid z_{d,n}, \beta_{1:K}) \right]. \quad (4)$$

Non insisteremo ulteriormente sulla struttura del modello, poiché essa è ampiamente approfondita in letteratura. Vogliamo però solo mettere in evidenza che marginalizzando la (4) documento per documento rispetto alle proporzioni dei topic θ_d, arriviamo alla seguente espressione:

$$p(z_{d,1:N_d}, w_{d,1:N_d} \mid \alpha, \beta_{1:K}) = \int \left[\prod_{n=1}^{N_d} p(z_{d,n} \mid \theta_d) p(w_{d,n} \mid z_{d,n}, \beta_{1:K}) \right] p(\theta_d \mid \alpha) \, d\theta_d \,. \quad (5)$$

Poiché il processo generativo può, almeno potenzialmente, generare uno stream infinito di token, la (5) afferma che la successione bivariata $(w_{d,1:N_d}, z_{d,1:N_d})$ è infinitamente scambiabile nel senso di De Finetti (qualunque sia il valore effettivo di N_d). Ciò significa che ogni sotto-successione finita di token (ossia un documento nel senso in cui lo abbiamo definito) ha distribuzione di probabilità congiunta invariante per permutazioni. In altre parole, l'ordine con il quale appaiono i token non è rilevante, ma solo la loro frequenza di occorrenza ha significato. Questa sottile conseguenza della specificazione gerarchica che abbiamo introdotto si è dimostrata utile in molti contesti e, come vedremo, funziona bene per il problema che dobbiamo affrontare. Tuttavia, è bene notare che questa '*bag-of-word assumption*' potrebbe risultare inadeguata per altri obiettivi, come ad esempio la traduzione automatica, per quale la posizione occupata dalle parole può essere rilevante.

2.1 Il ruolo della distribuzione di Dirichlet

Nella definizione del modello che abbiamo descritto nel Paragrafo precedente, un ruolo centrale è giocato dalla distribuzione di Dirichlet, che all'interno della specificazione gerarchica del modello definisce la distribuzione a priori per le frequenze di occorrenza dei topic in ciascun documento. Da un punto di vista formale, la distribuzione di Dirichlet è una distribuzione di probabilità in $\mathbb{R}^K$ a supporto sul simplesso unitario, ed avente la seguente funzione di densità di probabilità (per semplicità, elimineremo il riferimento al documento d):

$$p(\theta \mid \alpha) = \frac{\Gamma(\sum_i \alpha_i)}{\prod_i \Gamma(\alpha_i)} \prod_{i=1}^{K} \theta_i^{\alpha_i - 1}, \tag{6}$$

con $\alpha = (\alpha_1, \ldots, \alpha_K)$, $\alpha_i \geq 0$ per $i = 1, \ldots, K$. Tale distribuzione può essere evidentemente interpretata come un modello probabilistico per le distribuzioni di probabilità discrete a supporto finito su K elementi, ossia $\theta = (\theta_1, \ldots, \theta_K)$, $\theta_i \geq 0$ per ogni $i = 1, \ldots, K$, con l'ovvio vincolo $\sum_i \theta_i = 1$.

La geometria della distribuzione di Dirichlet può essere meglio apprezzata ricorrendo al cosiddetto *parametro di concentrazione*, $\alpha_0 = \alpha_1 + \cdots + \alpha_K$, e *alla misura di base*, definita da $\alpha_i' = \alpha_i / \alpha_0$ per $i = 1, \ldots, K$. La misura di base determina il valore atteso di ciascuna componente marginale, poiché si dimostra che:

$$E(\theta_i \mid \alpha) = \frac{\alpha_i}{\alpha_0} = \alpha_i', \tag{7}$$

mentre il parametro di concentrazione influisce direttamente sulle varianze delle distribuzioni marginali:

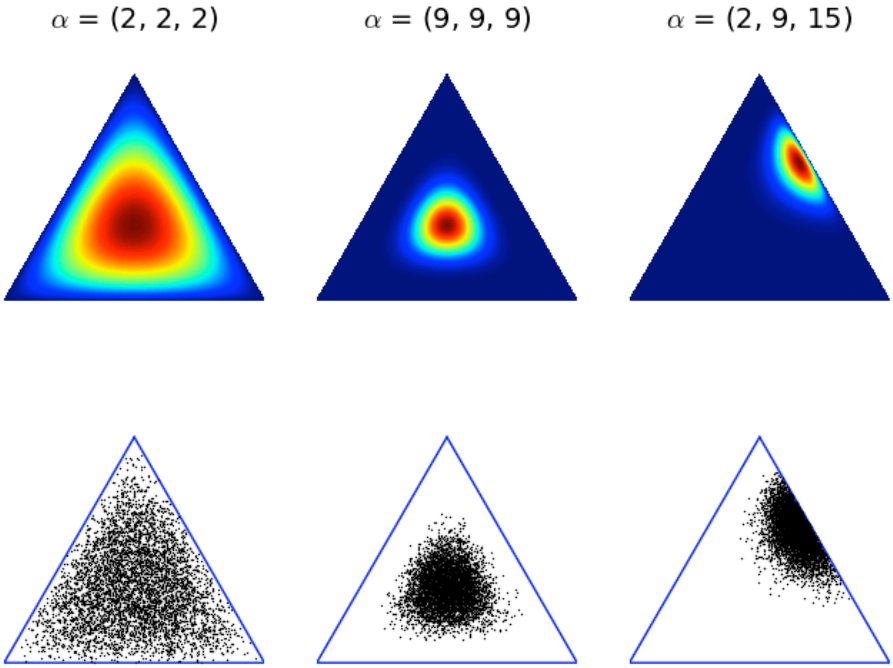
$$Var(\theta_i \mid \alpha) = \frac{\alpha_i'(1 - \alpha_i')}{\alpha_0 + 1}. \tag{8}$$

Quando $\alpha_1 = \cdots = \alpha_K = \alpha$ si parla di distribuzione di Dirichlet *simmetrica*, ed α può essere trattato come un singolo parametro scalare. In questo caso la misura di base è uguale ad $\alpha_i' = 1 / K$ per $i = 1, \ldots, K$, le componenti marginali hanno tutte la stessa distribuzione, e la distribuzione multivariata è simmetrica rispetto al baricentro del simplesso unitario. In questo caso saranno favorite distribuzioni approssimativamente uniformi sulle proporzioni dei topic (la variabilità dipenderà dal parametro di concentrazione). Alcuni esempi di queste proprietà sono stati visualizzati nella Figura 2 per $K = 3$, proiettando opportunamente sul simplesso unitario i contorni della distribuzione e 5000 generazioni aleatorie. Si tenga inoltre presente che, nel caso simmetrico:

- se $\alpha = 1$ la distribuzione è simmetrica sul simplesso unitario;
- se $\alpha > 1$ la distribuzione ha concavità rivolta verso il simplesso, e quindi è concentrata attorno al baricentro del simplesso, con dispersione che dipenderà dal parametro di concentrazione;
- se $\alpha < 1$ la distribuzione ha concavità in direzione opposta a quella del caso precedente, e quindi concentra la massa probabilistica attorno ai vertici

del simplesso unitario, caratterizzati da una o più probabilità marginali prossime a zero.

Figura 2. *Distribuzione di Dirichlet in* $\mathbb{R}^3$. *In alto: contorni della distribuzione corrispondenti a due casi simmetrici ed uno non simmetrico. In basso: 5000 generazioni aleatorie dalle stesse distribuzioni, proiettate sul simplesso unitario.*



Pertanto, nell'ultimo dei tre casi appena descritti, saranno favorite distribuzioni sparse sulle proporzioni dei topic. In questo caso i documenti tenderanno ad essere generati come un miscuglio di un piccolo numero di topic. Tuttavia, va sottolineato che una distribuzione a priori di questo tipo è poco informativa a causa della varianza piuttosto elevata (il parametro di concentrazione è molto piccolo), e può essere facilmente dominata dalla verosimiglianza. In generale, la scelta del parametro α può risultare particolarmente influente sulla distribuzione a posteriori delle proporzioni dei topic. Per quanto riguarda il modello LDA, l'algoritmo variazionale di stima che descriveremo nelle prossime pagine, permette direttamente la stima di α a partire dai dati. L'uso di una distribuzione di Dirichlet non simmetrica può portare ad alcuni vantaggi (Wallach et al., 2009), ma questa complicazione non sarà ulteriormente analizzata nel seguito del presente lavoro.

2.2 L'inferenza nel modello LDA

Sebbene la struttura gerarchica del modello LDA sembri adatta ad un approccio alla stima dei parametri di natura pienamente bayesiana, sono stati sviluppati metodi alternativi computazionalmente efficienti che tengono conto della peculiare struttura del modello che stiamo considerando. In particolare, possiamo considerare tanto i nodi stocastici che governano le allocazioni dei token ai topic, nonché quelli che definiscono la distribuzione dei topic nei singoli documenti, come variabili latenti che contribuiscono a governare la distribuzione dei dati (ossia le frequenze di occorrenza dei token). In questo modo, l'algoritmo EM per la ricerca di stime di massima verosimiglianza in presenza di dati mancanti diventa la via naturale per l'inferenza, poiché è ben noto che i dati mancanti sono trattati alla stessa stregua di variabili latenti non osservate.

Esiste tuttavia un'importante limitazione all'utilizzo dell'algoritmo EM nella sua forma originale. Per comprenderne la natura, consideriamo un insieme di variabili latenti $\ell_{1:D}$ di osservazioni $w_{1:D}$ (per semplicità, sopprimiamo momentaneamente la dipendenza da eventuali parametri incogniti). In generale, supporremo che la distribuzione congiunta dei dati e delle variabili latenti dati non osservati possa essere fattorizzata nel modo seguente (questa ipotesi è, nello specifico, vera per il modello LDA):

$$p(\ell_{1:D}, w_{1:D}) = p(\ell_{1:D}) p(w_{1:D} \mid \ell_{1:D}) . \tag{9}$$

Affinché l'algoritmo EM possa essere applicato, è necessario esplicitare la distribuzione a posteriori delle variabili latenti:

$$p(\ell_{1:D} \mid w_{1:D}) = \frac{p(\ell_{1:D}, w_{1:D})}{p(w_{1:D})}, \tag{10}$$

dove il denominatore contiene la *verosimiglianza marginale* dei dati, ottenuta marginalizzando la distribuzione congiunta rispetto alle variabili latenti del modello:

$$p(w_{1:D}) = \int p(\ell_{1:D}, w_{1:D}) d\ell_{1:D} . \tag{11}$$

Nel modello LDA la (11) non è esplicitabile in forma chiusa, e quindi la distribuzione a posteriori delle variabili latenti non è direttamente disponibile. Infatti, si può dimostrare che l'espressione della (11) per un singolo documento è la seguente (Blei at al., 2003):

$$p(w_{1:N_d} \mid \alpha, \beta_{1:K}) = \frac{\Gamma\left(\sum_i \alpha_i\right)}{\prod_i \Gamma(\alpha_i)} \int \left(\prod_{i=1}^{K} \theta_{d,i}^{\alpha_i - 1} \right) \left(\sum_{n=1}^{N_d} \sum_{i=1}^{K} \prod_{j=1}^{|V|} (\theta_{d,i} \beta_{ij})^{w_{d,n}^j} \right) d\theta_d \,. \quad (12)$$

Questa espressione non è esplicitabile a causa dell'accoppiamento tra θ_d e $\beta_{1:K}$ quando andiamo ad eseguire la somma sui topic latenti. Non è possibile neppure un approccio diretto all'inferenza di tipo brute-force, poiché Sontag e Roy (2011) dimostrano che relativa la complessità computazione è NP-hard nella maggior parte dei casi pratici.

Nell'approccio variazionale (Wainwright e Jordan, 2007; Tzikas et al., 2008; Blei et al., 2017) la soluzione consiste nello scegliere una famiglia di distribuzioni di probabilità Q, e nell'approssimare direttamente la distribuzione a posteriori risolvendo il seguente problema di approssimazione funzionale:

$$\arg\min_{q(\ell_{1:D}) \in Q} \mathrm{KL}\left(q(\ell_{1:D}) \,\|\, p(\ell_{1:D} \mid w_{1:D}) \right). \quad (13)$$

In altre, l'obiettivo dell'ottimizzazione è la funzione $q^*(\ell_{1:D})$ che minimizza la divergenza di Kullback-Leibler con la distribuzione a posteriori esatta. Questo approccio variazionale rappresenta, pertanto, un approccio all'inferenza bayesiana alternativo a quello basato sui metodi MCMC. Mentre in questi ultimi la parte essenziale è giocata dal campionamento, nell'approccio variazionale è invece centrale l'ottimizzazione. Si può facilmente dimostrare che la divergenza di Kullback-Leibler ha la seguente espressione alternativa:

$$\begin{aligned} &\mathrm{KL}\left(q(\ell_{1:D}) \,\|\, p(\ell_{1:D} \mid w_{1:D}) \right) = \\ &\quad = E_q\left[\log q(\ell_{1:D}) \right] - E\left[\log p(\ell_{1:D}, w_{1:D}) \right] + \log p(w_{1:D}) \end{aligned}, \quad (14)$$

e dipende essa stessa dalla verosimiglianza marginale (intrattabile). Tuttavia, possiamo considerare la seguente funzione obiettivo, che è equivalente alla divergenza di Kullback-Leibler:

$$\mathrm{ELBO}(q) = E\left[\log p(\ell_{1:D}, w_{1:D}) \right] - E_q\left[\log q(\ell_{1:D}) \right], \quad (15)$$

ottenuta cambiando di segno la (14), ed aggiungendo $\log p(w_{1:D})$, poiché la log-verosimiglianza marginale non dipende da q. La quantità (15) prende il nome di *energia libera* o ELBO (*Evidence Lower Bound*), in quanto con semplici manipolazioni algebriche si può dimostrare che:

$$\log p(w_{1:D}) = \mathrm{KL}\left(q(\ell_{1:D}) \,\|\, p(\ell_{1:D} \mid w_{1:D})\right) + \mathrm{ELBO}(q), \tag{16}$$

e poiché $KL(\bullet \,\|\, \bullet) \geq 0$, ne consegue che $\mathrm{ELBO}(q)$ è un limite inferiore della log-verosimiglianza marginale, che sarà tanto più stretto quanto migliore è l'approssimazione di $p(\ell_{1:D} \mid w_{1:D})$ tramite $q(\ell_{1:D})$. Nel caso più estremo possibile, $\mathrm{ELBO}(q) \equiv \log p(w_{1:D})$, il che accade quando $q(\ell_{1:D}) \equiv p(\ell_{1:D} \mid w_{1:D})$.

Per il modello LDA, una tipica scelta della famiglia variazionale Q la si ottiene prendendo:

$$q(\ell_{1:D} \mid \gamma_{1:D}, \phi_{1:D}) = \prod_{d=1}^{D} q(\theta_d \mid \gamma_d) \prod_{n=1}^{N_d} q(z_{d,n} \mid \phi_{d,n}), \tag{17}$$

ossia l'inferenza avviene separatamente per ciascun documento e per le proporzioni dei topic e le relative variabili indicatrici, dove $q(\theta_d \mid \gamma_d)$ e $\prod_{n=1}^{N_d} q(z_{d,n} \mid \phi_{d,n})$ sono rispettivamente una distribuzione di Dirichlet e una distribuzione Multinomiale-prodotto che dipendono, rispettivamente, dai parametri variazionali γ_d e $\phi_{d,1:N_d}$. In questo modo, l'obiettivo variazionale (13) si trasforma nel modo seguente:

$$\underset{q(\gamma_d, \phi_{d,1:N_d}) \in Q}{\arg\min}\ \mathrm{KL}\left(q(\theta_d, z_{d,1:N_d} \mid \gamma_d, \phi_{d,1:N_d}) \,\|\, p(\theta_d, z_{d,1:N_d} \mid w_{d,1:N_d}, \alpha, \beta_{1:K})\right). \tag{18}$$

Se disponiamo di un set di parametri variazionali ottimali γ_d^* e $\phi_{d,1:N_d}^*$, che soddisfano il problema di ottimizzazione funzionale (18), avremo un limite inferiore trattabile della log-verosimiglianza marginale, che può essere come surrogato di quest'ultima nella versione dell'algoritmo EM che prende appunto il nome di variational EM, e si compone dei seguenti step:

1. (Variational E-step) Per ciascun documento del corpus determiniamo i valori ottimali dei parametri variazionali $\{\gamma_d^*, \phi_{d,1:N_d}^*; d \in D\}$;
2. (Variational M-Step) In corrispondenza delle stime dei parametri variazionali appena ottenute, massimizziamo il limite inferiore $\mathrm{ELBO}(q^*)$ rispetto ai parametri α e $\beta_{1:K}$.

Ovviamente, la distribuzione variazionale è completamente determinata una volta che i suoi parametri siano stimati. Poiché la distribuzione variazionale è trattabile analiticamente, possiamo immediatamente scrivere in forma chiusa il valore atteso delle sue componenti marginali, e determinare un insieme di stime delle variabili latenti $z_{1:D}$ e $\theta_{1:D}$ in funzione delle osservazioni e dei parametri variazionali appena determinati dall'algoritmo.

Nella pratica, per ottenere un set finale di stime, i due step che abbiamo descritto sono ripetuti in modo alternante, fino a quando il limite inferiore della log-verosimiglianza converge. La versione generale dell'algoritmo è delineata in Blei et al. (2017): poiché all'interno di ciascuno step i parametri sono ottimizzati uno alla volta mantenendo fissi tutti gli altri gli altri al loro valore corrente, l'algoritmo è noto anche come CAVI (Coordinate Ascent Variational Inference). Dettagli più specifici sull'implementazione nel caso del modello LDA sono invece contenuti in Blei et al. (2003) e Blei (2012). Globalmente, l'algoritmo implementa un approccio di tipo Empirical Bayes, nel quale gli iper-parametri di interesse α e $\beta_{1:K}$ non sono considerati come nodi stocastici dotati a loro volta di una distribuzione di probabilità a priori, ma piuttosto come nodi deterministici incogniti che vanno stimati a partire dai dati.

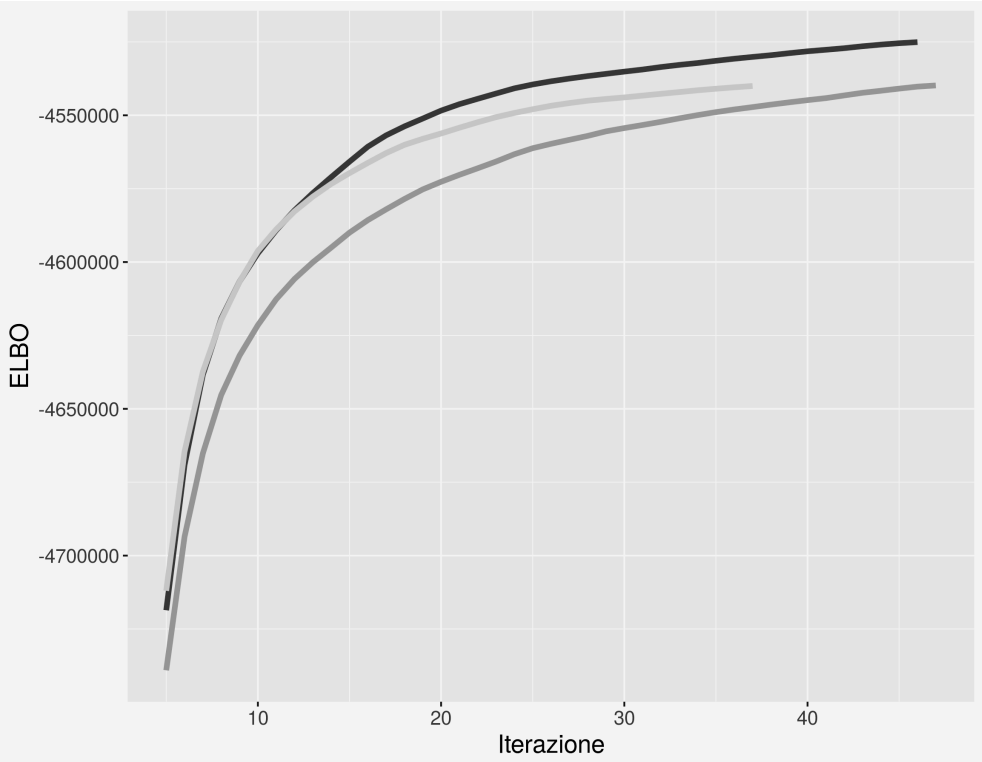
Come nell'algoritmo EM classico, il limite inferiore della log-verosimiglianza è presumibilmente sempre crescente nelle successive iterazioni dell'algoritmo. Abbiamo usato una espressione dubitativa in quanto, al momento, non sono note condizioni sufficienti di natura generale sotto le quali la convergenza dell'algoritmo è garantita, e molte ricerche sono in atto per arrivare a dei risultati in tal senso (Awasthi, 2015). Inoltre, poiché l'obiettivo variazionale è in generale una funzione non convessa, l'algoritmo convergerà ad un punto di minimo locale che potrebbe non coincidere con il massimo globale desiderato.

La soluzione a questo problema è quella tipicamente utilizzata con la maggior parte degli algoritmi iterativi: l'intero processo è ripetuto per un certo numero di run indipendenti, che differiscono esclusivamente per i valori iniziali di innesco scelti in modo casuale, e in ciascun run viene monitorato il limite inferiore della log-verosimiglianza. Al termine del processo verrà scelto quel run in corrispondenza del quale l'ELBO ha raggiunto il valore più elevato in fase di convergenza. Nella Figura 3 presentiamo un esempio relativo a tre run indipendenti, basati sullo stesso dataset che sarà utilizzato nel successivo Paragrafo.

Concludiamo infine osservando che l'algoritmo che abbiamo descritto ha molte similarità con quell'algoritmo di tipo MCMC noto come *Gibbs Sampling* (si veda Blei at al., 2017, per i dettagli). Tuttavia, esistono degli algoritmi MCMC specializzati per la stima dei parametri di tipo collapsed Gibbs Sampling (Griffiths e Steyvers, 2004), nel quale il campionamento avviene sulle sole variabili indicatrici dopo aver marginalizzato rispetto agli altri parametri del modello. Questo particolare schema di campionamento ha dato origine a molte varianti migliorative dal punto di vista dell'efficienza, incluse implementazioni parallele basate su architettura MapReduce, per migliorare la scalabilità su grandi collezioni di documenti. I

dettagli di queste implementazioni e il confronto con l'algoritmo Variational EM esulano dall'obiettivo di questo lavoro, e per essi rimandiamo alla letteratura sull'argomento (Steyvers e Griffith, 2007; Porteous et al.; 2008; Liu et al., 2011, Speh et al., 2013; Chen et al., 2015).

Figura 3. *Esempio di tre traiettorie seguite dal limite inferiore della log-verosimiglianza (ELBO), durante tre run distinti sullo stesso dataset dall'algoritmo Variational EM, che differiscono solo per i valori di innesco scelti in modo casuale.*



2.3 Esempio

Per l'esempio che presentiamo in questa sezione, utilizzeremo i dati dei Form 10-K messi per l'anno 2006, messi a direttamente a disposizione da Kogan et al. (2009) su sito apposito[2]. Il testo di ciascun Form non è riportato nella sua interezza, ma contiene la sola Sezione 7, intitolata '*Management's Discussion and Analysis*' (MD&A), all'interno della quale particolare interesse riveste la Sezione 7A '*Quan-*

[2] www.cs.cmu.edu/~ark/10K/

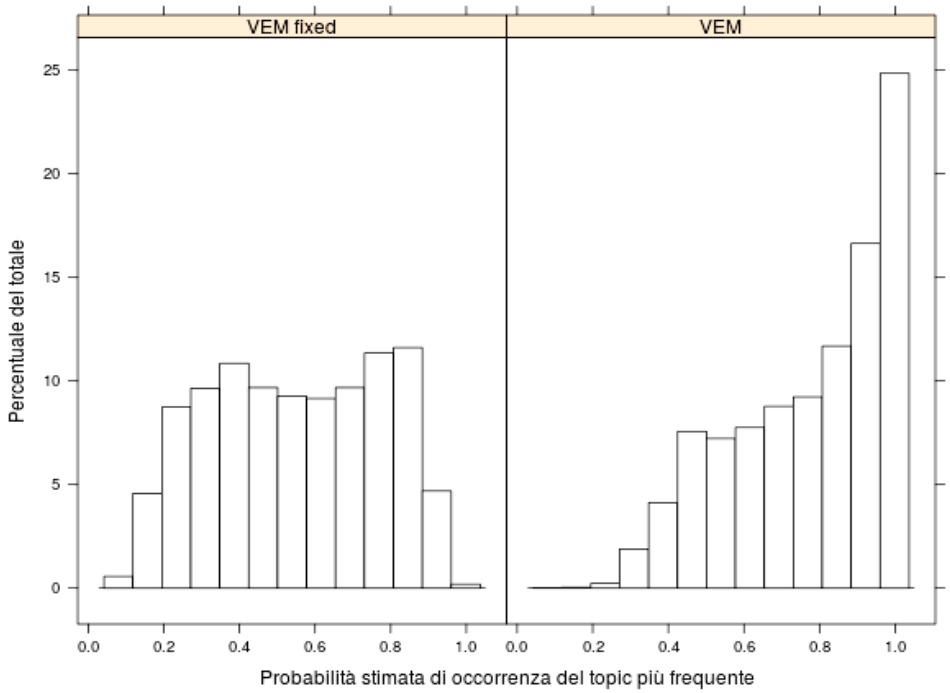
titative and Qualitative Disclosures about Market Risks', contenente un insieme di *forward-looking statements* che, sulla base delle informazioni correnti, esprimono il punto di vista del management sulle aspettative del mercato e riportano le proiezioni sulle performance attese.

La verosimiglianza del modello LDA è di tipo multinomiale-prodotto, e quindi il modo naturale di strutturare i dati in input è attraverso la *matrice documenti-termini*, che per ciascuna riga (documento) riporta le frequenze di occorrenza del vocabolario dei termini V. Questa strutturazione dei dati riflette in modo naturale il fatto che il modello tratta i documenti testuali come delle *bag-of-words*, all'interno della quale non conta la posizione di occorrenza dei token. La creazione del vocabolario dei termini V a partire dal testo grezzo è un processo complesso, che necessità di una serie di fasi di *pre-processing*, che descriveremo in maggior dettaglio nel Paragrafo 4 dedicato ai risultati. Per il momento, possiamo dire che il corpus originale per l'anno 2006 contiene 3306 Form, per un totale di circa 3.92×10^7 termini (contando ovviamente le ripetizioni), con una media di 11854 termini per ciascun documento. In questo contesto, *termine* sta indicare qualsiasi unità semanticamente determinata, includendo anche i numeri e tutti quei termini noti come *stopword*, che occorrono molto frequentemente all'interno di un testo (come ad esempio articoli, avverbi e congiunzioni), ma che non trasportano una quantità di informazione semantica particolarmente rilevante. Al termine del pre-processing, che prevede la rimozione di tutti i numeri e delle stopword, più altre operazioni specifiche di normalizzazione, il vocabolario dei termini è risultato costituito da 4376 token distinti, con un coefficiente di sparsità del 99% (nel senso che il 99% delle celle della matrice documenti-termini conteneva una frequenza di occorrenza del relativo token pari allo zero).

L'iper-parametro α, che governa la distribuzione delle proporzioni dei topic nei documenti, può essere stimato direttamente attraverso l'algoritmo variazionale, ovvero è possibile fissarne il valore ad un livello plausibile (e lasciare all'algoritmo il solo compito di stimare le distribuzioni contenute in $\beta_{1:K}$). Per esempio, Griffiths e Steyvers (2004) suggeriscono $\alpha = 50/K$, dove K è il numero di topic (ossia il numero di componenti del miscuglio di distribuzioni multinomiali definite attraverso la struttura gerarchica del modello). Per valori non troppo elevati di K, questa scelta corrisponde ad un setting non informativo (si riveda la Figura 2), e che in ogni caso manterrà costante il valore del parametro di concentrazione al variare di K. Quando invece a è stimato direttamente dai dati sulla base dell'algoritmo variazionale, tale parametro tenderà spesso ad assumere valori inferiori ad uno (Grün, e Hornik, 2011), con la conseguenza, già analizzata nel Paragrafo precedente, che

molti documenti tenderanno a contenere solo pochi topic con probabilità elevata, mentre la maggior parte dei restanti topic avrà una probabilità di occorrenza tra i token praticamente prossima a zero. Questo effetto è ben documentato nella Figura 4, dove abbiamo presentato i risultati relativi alla stima di un modello con $K = 15$ topic: quando α è stimato tramite l'algoritmo variazionale (*VEM*), all'incirca nel 25% dei documenti viene identificato un solo topic con probabilità pari ad uno. La sparsità nella distribuzione dei topic è molto evidente, e la conseguente riduzione di dimensionalità è molto marcata. Invece, quando $\alpha = 50 / K$ (*VEM fixed*) la sparsità si riduce notevolmente, e solo in una percentuale molto esigua di documenti viene identificato un unico topic latente.

Figura 4. *Distribuzione empirica della probabilità stimata di occorrenza del topic più frequente, calcolata mediante un modello LDA sul corpus dei Form* 10-*K per l'anno* 2006*, settando il numero di topic pari a* $K = 15$. *Nel grafico di sinistra (VEM fixed) il parametro* α*, che regola la distribuzione delle proporzioni nei topic, è stato settato come* $\alpha = 50 / K$*, mentre nel grafico di sinistra è stato stimato direttamente dall'algoritmo variazionale.*



Ovviamente, non è possibile decidere a priori se la sparsità sia una virtù o un difetto, poiché la risposta è funzione dello specifico problema che stiamo trattando e degli obiettivi che ci siamo posti nell'analisi (anche se, in genere, la riduzione di

dimensionalità si riflette in un miglioramento dell'interpretabilità e dell'espressività dei risultati).

Come per ogni altro modello basato su un miscuglio di distribuzioni, un problema centrale è quello dell'individuazione del numero di componenti del miscuglio, problema che nel nostro caso coincide con la determinazione del numero ottimale di topic. Assegnati un insieme di parametri $\hat{\alpha}$ e $\hat{\beta}_{1:K}$ stimati su insieme di apprendimento (*training set*), un modo standard per valutare un insieme di modelli alternativi è quello di calcolare la verosimiglianza marginale su un insieme di D' documenti di test, ossia:

$$\log p(w_{1:D'} \mid \hat{\alpha}, \hat{\beta}_{1:K}) = \sum_{d=1}^{D'} \sum_{n=1}^{N_d} \log p(w_{d,n} \mid \hat{\alpha}, \hat{\beta}_{1:K}), \tag{19}$$

Sebbene nella (19) la variabilità campionaria non è presa in considerazione, poiché stiamo effettuando i calcoli condizionalmente alle stime ottenute sul particolare insieme di training considerato, l'utilizzo di un dataset di test distinto dall'insieme di training permette di ridurre l'ottimismo sulle performance del modello stimato in termini di generalizzazione (ossia di descrivere con elevata probabilità l'occorrenza di documenti futuri che non sono stati osservati al momento della stima). Dunque, in un insieme di modelli candidati (che ad esempio differiscono solo per il numero di componenti K), sceglieremo quel modello nel quale la log-verosimiglianza marginale K assume il valore più elevato. Tuttavia, è pratica comune considerare una funzione alternativa monotona della log-verosimiglianza marginale, chiamata *perplexity*, e definita nel modo seguente:

$$perplexity = \exp\left\{ -\frac{\sum_{d=1}^{D'} \sum_{n=1}^{N_d} \log p(w_{d,n} \mid \hat{\alpha}, \hat{\beta}_{1:K})}{\sum_{d=1}^{D'} N_d} \right\}, \tag{20}$$

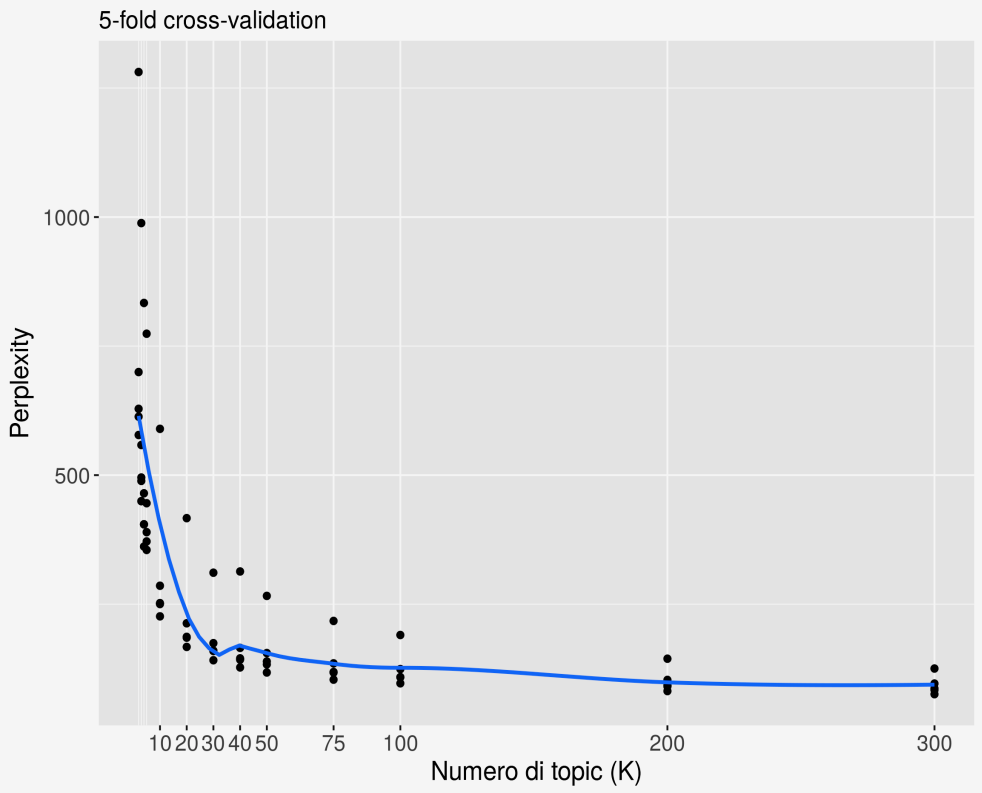
ossia la media geometrica per-token della log-verosimiglianza marginale. Naturalmente, la perplexity è una funzione monotona decrescente della log-verosimiglianza marginale. Dunque, il 'miglior' modello un set di modelli candidati è quello la cui perplexity assume il valore più basso.

Sebbene la (20) sia una misura *data-based* il cui significato intuitivo è semplice e evidente, motivo per il quale è particolarmente utilizzata negli studi applicativi, essa è stata critica da alcuni autori, che hanno scoperto empiricamente che i risultati ottenuti tramite la perplexity e il giudizio umano sono spesso *anti-correlati* (in

altre parole, l'utilizzo della perplexity porterebbe a risultati nei quali i topic latent stimati sono spesso non facilmente interpretabili; Chang et al., 2009). Inoltre, la verosimiglianza marginale che appare nella (19), sebbene dipenda dalle stime ottenute nella fase di training, è intrattabile matematicamente in quanto la sua espressione esatta può essere ottenuta solo marginalizzando rispetto alle proporzioni dei topic $\theta_{1:D'}$. Abbiamo pertanto bisogno di riutilizzare in modo opportuno l'algoritmo variazionale per ottenere le variabili di assegnazione $z_{1:D'}$ per i documenti di test, utilizzando la relativa distribuzione a posteriore condizionale (condizionale ad $\hat{\alpha}$ e $\hat{\beta}_{1:K}$), nonché di costruire uno stimatore Monte Carlo efficiente che permetta di marginalizzare numericamente rispetto alle proporzioni dei topic. Non insisteremo ulteriormente su questi aspetti, poiché essi sono passati esaurientemente in rassegna in Wallach et al. (2009b). Per quanto riguarda i nostri dati, la curva di determinazione del valore ottimale di K presentata nella Figura 5 non è stata ottenuta attraverso una unica suddivisione dell'intero corpus in un insieme di training e in un insieme di test. Piuttosto, abbiamo utilizzato una 5-fold cross-validation, calcolando quindi cinque volte la perplexity in corrispondenza di ciascun fold e di ciascun valore prefissato di K (Hastie e Tibshirani, 2009; James et al., 2013). Operando in questo modo, viene ridotta l'influenza del particolare test set sul calcolo dell'errore di generalizzazione. Inoltre, ai valori calcolati è stato sovrapposto uno smoother (linea continua) per delineare la tendenza di fondo della perplexity al crescere di K.

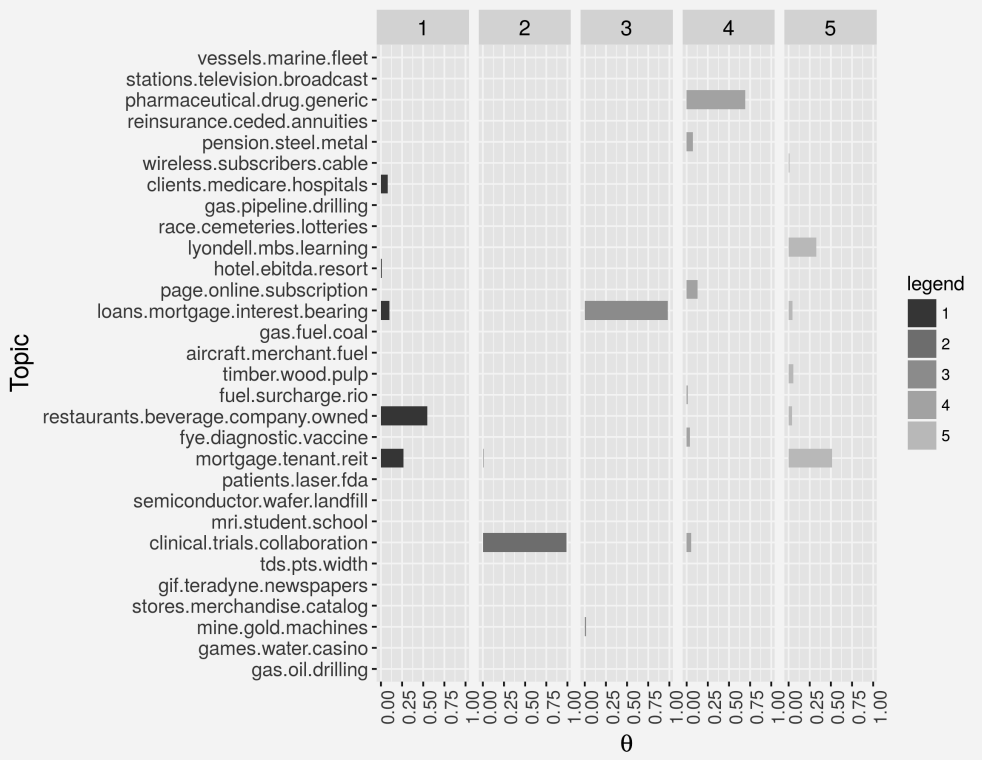
Basandosi su un esame visivo, sembra evidente che il valore della perplexity non raggiunge un minimo ben definito, ma continua costantemente a decrescere, anche in corrispondenza di modelli iper-parametrizzati che contengono $K = 200$ e più topic. Tuttavia, a partire da $K = 30$ la perplexity inizia a decrescere molto lentamente, e le variazioni successive della relativa media sui cinque fold tendono a diventare praticamente insignificanti per $K \geq 40$. Dunque, anche in assenza di un minimo locale ben definito la scelta $K = 30$ appare come un buon compromesso tra parsimonia, espressività del modello e tempo di calcolo. Del resto, il numero di topic del modello non è fonte di preoccupazione eccessiva per quanto attiene al costo computazionale dell'algoritmo variazionale descritto nel Paragrafo 2.2. Infatti, Blei et al. (2003) dimostrano che la complessità temporale per un singolo documento è dell'ordine $O(N_d^2 K)$, e quindi il ruolo più importante è giocato dalla lunghezza dei documenti del corpus. Ad ogni modo, come abbiamo già notato in precedenza, la ricerca di tecniche computazionalmente più efficienti è un argomento di studio estremamente attuale in questo settore della linguistica computazionale.

Figura 5. *Curva per la determinazione del valore ottimale di K, ottenuta inserendo il calcolo della perplexity in una 5-fold cross-validation. Ai valori calcolati è stato sovrapposto uno smoother (linea continua) per delineare la tendenza di fondo della perplexity al crescere del numero di componenti (topic) K.*



Infine, la Figura 6 presenta un esempio di classificazione, su cinque documenti di test, ottenuta mediante il classificatore empirico standard MAP (Maximum a Posteriori). In altre parole, ciascun documento di test viene classificato in modo non supervisionato in base al topic che occorre con la frequenza relativa più elevata nel documento considerato (ciascun topic è stato riassunto mostrando i tre token che appaiono con la frequenza più elevata). Come si può notare (il parametro α è stato stimato all'interno dell'algoritmo variazionale) la distribuzione delle proporzioni dei topic è notevolmente sparsa, e il topic più frequente è sempre dominante rispetto a tutti gli altri (anche in termini di interpretabilità) che, invece, sono presenti con una probabilità prossima allo zero o comunque molto piccola.

Figura 6. *Classificazione non supervisionata MAP (Maximum a Posteriori) di un insieme di cinque documenti scelti dal corpus dei Form-10K. I trenta topic latenti identificati dal modello LDA sono descritti attraverso i tre token che occorrono con le frequenza più elevate. Ciascun documento è assegnato a quel topic per il quale la relativa proporzione di occorrenza nel documento (ottenuta attraverso la stima a posteriori di* θ_d*) è massima.*



3. Approcci supervisionati

Come abbiamo già messo in evidenza nell'introduzione, l'obiettivo finale di questo lavoro non è quello di individuare il contenuto tematico latente del corpus che stiamo analizzando, bensì quello di utilizzare (almeno indirettamente) tali topic latenti per la previsione di un'etichetta, che rappresenta un'informazione ausiliaria associata a ciascun documento (nel nostro caso la volatilità forward a 12 mesi, opportunamente discretizzata). A questo approccio 'strutturato', possiamo contrapporre un approccio 'diretto', che utilizza direttamente le frequenze di occorrenza dei token nei documenti, modellizzate attraverso un modello di regressione logistica multinomiale, opportunamente penalizzato per ridurre la sparsità (e rendere pos-

sibile la stima dei parametri anche quando la dimensione dello spazio dei parametri eccede quella del numero di documenti disponibile).

3.1 La regressione logistica multinomiale penalizzata

Per la famiglia Multinomiale, ricordando che l'etichetta del generico documento sarà indicata con $c \in \Xi = \{1,\ldots,C\}$, la verosimiglianza assume la seguente forma:

$$p(c_d \mid x_d, \beta_{1:C}) = \text{softmax}(x_d, \beta_{1:C}) = \frac{\exp(\beta_c^T x_d)}{\sum_{\ell=1}^{C} \exp(\beta_\ell^T x_d)}, \tag{21}$$

dove x_d è un vettore di dimensione $p = |V| + 1$ (includendo anche un termine unitario corrispondente all'intercetta) che contiene le frequenze di occorrenza di ciascun token nel documento $d \in D$. Con la stessa notazione utilizzata in precedenza, $\beta_{1:C}$ raccoglie i coefficienti di regressione relativi a ciascuna etichetta, e quindi è isomorfo ad una matrice di dimensione $p \times C$, ottenuta sistemando i vettori β_c lungo le colonne. Si noti che il modello (21) è stato espresso tramite una parametrizzazione log-lineare compatta (che nella letteratura che si occupa di machine learning è nota anche *funzione di attivazione softmax*), che è assolutamente equivalente alla forma standard nella quale viene di solito espresso il modello di regressione logistica Multinomiale (Hastie e Tibshirani, 2009), dati i vincoli sulle probabilità di appartenenza alle classi (che devono, ovviamente, sommare ad uno).

Se introduciamo una matrice di variabili indicatrici $Y = \{y_{d,c}\}$ di dimensione $D \times C$, tale che $y_{d,\ell} \equiv I(c_d = \ell)$ per $\ell = 1,\ldots,C$, allora la verosimiglianza penalizzata, utilizzando la specifica penalizzazione che prende il nome di *elastic-net*, ha la seguente espressione (Zou e Hastie, 2005; Friedman et al., 2010):

$$\begin{aligned} l(\beta_{1:C}) = &-\left[\frac{1}{D}\sum_{d=1}^{D}\left(\sum_{c=1}^{C} y_{d,c}(\beta_c^T x_d) - \log\left(\sum_{c=1}^{c}\exp(\beta_c^T x_d)\right)\right)\right] + \\ &+\lambda\left[\frac{(1-\alpha)\left\|\beta_{1:C}\right\|_F^2}{2} + \alpha\sum_{j=1}^{p}\left\|\beta^j\right\|_1\right] \end{aligned}, \tag{22}$$

In questa espressione, $\|\bullet\|_F$ indica la norma matriciale di Frobenius, mentre $\|\bullet\|_1$ indica la norma L_1 standard, e β^j indica la *j*-esima riga della matrice $\beta_{1:C}$. Nell'espressione (22), la penalità di tipo *elastic-net* è controllata dal parametro $0 \le \alpha \le 1$, in base al quale si ottiene una penalità standard di tipo *lasso* quando

$\alpha = 1$ (Tibshirani, 1996), mentre per $\alpha = 0$ otteniamo una penalità di tipo *ridge* (Hoerl e Kennard, 2000). Come è ben noto, una penalità di tipo *ridge* spinge i coefficienti dei previsori fortemente correlati ad assumere valori simili, mentre una penalità di tipo *lasso* tende a selezionare solo alcuni previsori, spingendo verso lo zero i coefficienti dei rimanenti (riducendo, pertanto, la sparsità). Una penalità di tipo elastic-net con $0 < \alpha < 1$ mescola i due tipi di comportamento, selezionando alcuni gruppi di previsori fortemente correlati e spingendo i coefficienti dei previsori nei gruppi rimanenti verso lo zero. Ovviamente, questo effetto raggiunge il suo livello massimo se scegliamo $\alpha = 0.5$.

La verosimiglianza penalizzata (22) dipende anche dal parametro globale di regolarizzazione λ, che può essere scelto mediante cross-validation ottimizzando per l'accuratezza previsiva. In linea di principio, anche per α può essere effettuata una ricerca a griglia all'interno della procedura di cross-validation, simultaneamente alla ricerca del valore ottimale di λ. Infine, per α e λ prefissati, la (22) può essere minimizzata utilizzando una classica tecnica di discesa del gradiente, minimizzando la funzione obiettivo un coefficiente alla volta mentre tutti gli altri sono mantenuti fissi, e ripetendo la procedura fino a quando la convergenza non è stata raggiunta. Questo è, per esempio, l'approccio implementato dalla libreria `glmnet`, disponibile sotto l'ambiente R (R Core Team, 2017).

3.2 Supervised Latent Dirichlet Allocation (SLDA)

Il processo generativo del modello SLDA è rappresentato in forma grafica nella Figura 3. Operativamente, la specificazione gerarchica del modello è la seguente dove i primi tre punti sono identici a quelli del modello LDA, mentre un quarto punto viene specificatamente introdotto per dotare il modello LDA di capacità di classificazione supervisionata (Blei, e McAuliffe, 2007; Blei, 2012):

- $\theta_d \mid \alpha \sim \text{Dirichlet}_K(\alpha)$, per $d = 1,\ldots,D$.
- $z_{d,n} \mid \theta_d \sim \text{Multinomial}_{|K|}(1;\theta_d)$, per $d = 1,\ldots,D$, $n = 1,\ldots,N_d$.
- $w_{d,n} \mid z_{d,n},\beta_{1:K} \sim \text{Multinomial}_{|V|}(1;\beta_{z_{d,n}})$, per $d = 1,\ldots,D$, $n = 1,\ldots,N_d$.
- $c \mid z_{d,1:N_d},\eta_{1:C} \sim \text{softmax}(\overline{z}_d,\eta_{1:C})$, per $c = 1,\ldots,C$.

Nell'ultima specificazione, definiamo:

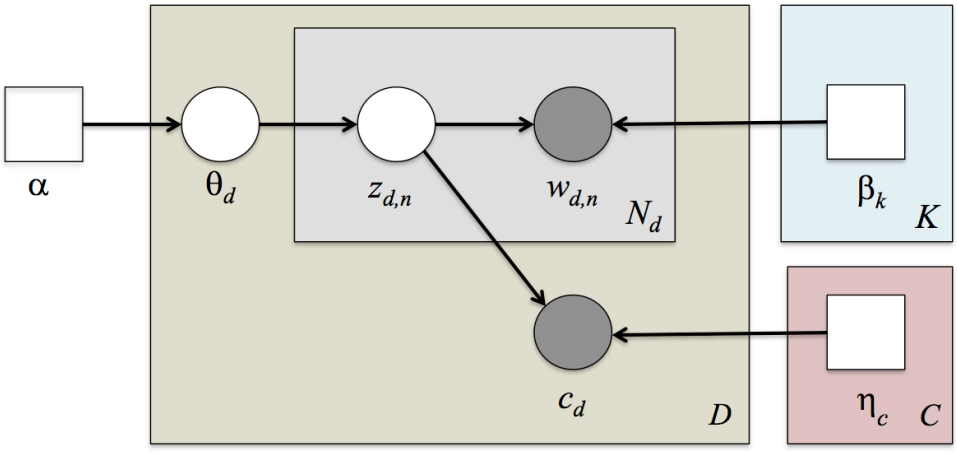
$$\overline{z}_d = \frac{1}{N_d}\sum_{d=1}^{N_d} z_{d,n}, \tag{23}$$

ossia $\bar{z}_d$ è il vettore delle frequenze empiriche di occorrenza dei topic latenti, mentre identicamente a quanto abbiamo scritto nella (21):

$$p(c_d \mid z_{d,1:N_d}, \eta_{1:C}) = \text{softmax}(z_{d,1:N_d}, \eta_{1:C}) = \frac{\exp(\eta_c^T \bar{z}_d)}{\sum_{\ell=1}^{C} \exp(\eta_\ell^T \bar{z}_d)} . \tag{24}$$

Si può verificare facilmente che per ciascun documento è valida una rappresentazione infinitamente scambiabile simile alla (5): dunque, anche per questo modello supervisionato i testi sono considerati come *bag-of-words*, nelle quali è irrilevante l'ordine di occorrenza dei token. Osserviamo inoltre, dalla Figura 3, che la variabile di risposta (etichetta) e i token hanno un arco generatore comune (ossia i topic latenti), e quindi non possono essere considerati condizionalmente indipendenti. In altre parole, i documenti sono generati come delle bag-of-words sotto l'ipotesi implicita di scambiabilità, e i topic sono poi utilizzati direttamente per prevedere l'etichetta associata a ciascuna documento.

Figura 7. *Rappresentazione del modello SLDA sotto forma di grafo orientato. I parametri* α, $\beta_{1:K}$ *ed* $\eta_{1:C}$ *sono trattati come iper-parametri incogniti da stimare sulla base dei dati, piuttosto che come nodi stocastici dotati di una distribuzione di probabilità a priori.*



In linea di principio, sono ovviamente sono possibili specificazioni alternative a quella rappresentata nella Figura 7. Per esempio, Blei e MacAuliffe (2007), sperimentano una soluzione nella quale l'etichetta è modellata tramite opportuna funzione non-lineare delle proporzioni dei topic contenute nel vettore θ_d. Tuttavia, gli autori concludono che questa specificazione è meno accurata in senso previsivo della precedente, come conseguenza del fatto che la massa probabilistica diffusa

sui topic latenti non è utilizzata per prevedere direttamente l'etichetta di risposta. Altri autori, in particolare Halpern et al. (2012), notano che la specificazione del modello SLDA tratta essenzialmente l'etichetta come un token aggiuntivo del documento, e può accadere che il contributo alla verosimiglianza del modello attribuibile ai token del vocabolario dei termini possa prevalere sul contributo attribuibile all'etichetta, con un conseguente degrado delle prestazioni su istanze future di test. Naturalmente queste considerazioni di carattere generale potrebbero non essere valide in domini specifici, e ritorneremo sulla questione nel prossimo Paragrafo.

Naturalmente, anche per il modello SLDA la distribuzione a posteriori delle variabili latenti non è disponibile in forma chiusa, ed anche in questo caso siamo costretti a ricorrere ad un algoritmo variazionale, i cui dettagli sono discussi in Blei e McAuliffe (2007). Sempre nell'ipotesi che la distribuzione variazionale q appresa in fase di training sia una buona approssimazione della distribuzione a posteriori delle variabili latenti, Wang et al. (2009) indicano come sia possibile utilizzare la distribuzione variazionale per ottenere le variabili di assegnazione $z_{1:D'}$ su un insieme di documenti di test D', e sulla base di ciò dimostrano che il classificatore MAP ottimale per un generico documento di test $d' \in D'$ può essere approssimato nel modo seguente:

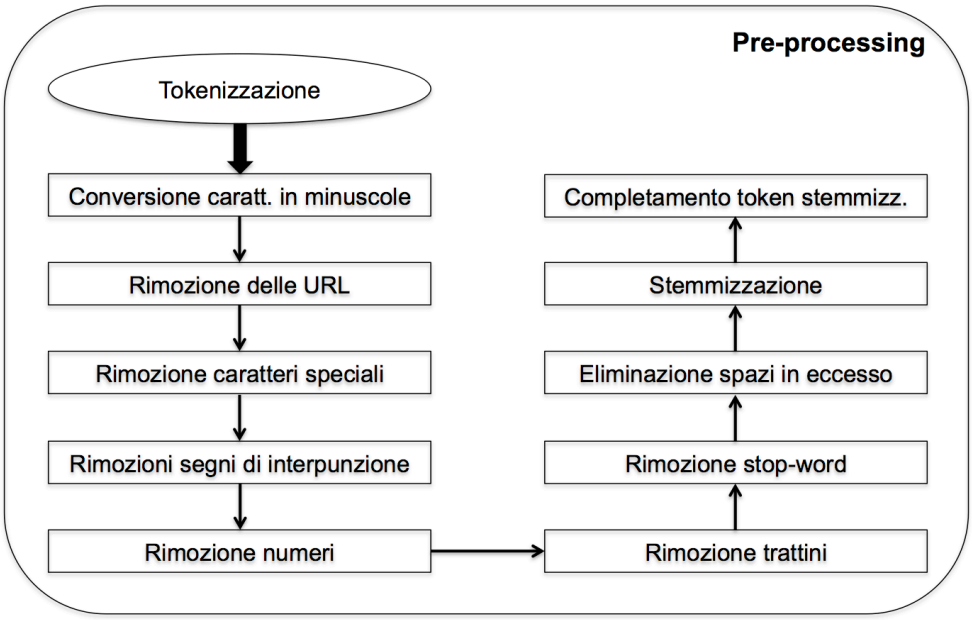
$$c_{d'}^{MAP} = \arg\max_{c \in \{1,\ldots,C\}} E_q\left[\hat{\eta}_c^T \bar{z}_{d'}\right] = \arg\max_{c \in \{1,\ldots,C\}} \hat{\eta}_c^T \hat{\bar{\phi}}_{d'}. \tag{25}$$

Nell'espressione che abbiamo appena scritto, $\hat{\eta}_c$ rappresenta la stima di η_c ottenuta in fase di training, ossia il coefficiente di regressione sulla scala del previsore lineare nella (24), $\bar{z}_{d'}$ è l'equivalente della (23) per il documento di test $d' \in D'$, mentre $\hat{\bar{\phi}}_{d'}$ è identica alla (23) per quanto attiene alla formula di calcolo, ma è ottenuta sulla base delle stime dei parametri variazionali $\phi_{d',n}$ per $d' \in D'$. Naturalmente, anche in questo caso abbiamo il problema di determinare il numero ottimale di componenti K. Per K prefissato, una volta classificati i documenti in D' siamo in grado di calcolare una stima dell'errore di generalizzazione. Quindi, la soluzione ovvia è quella di far variare K ottimizzando rispetto al tasso di incorretta classificazione: il valore ottimale di K è quello che ci permette di ottenere la classificazione più accurata dei documenti del test set.

4. Risultati

Come abbiamo già evidenziato nel Paragrafo precedente, la prima operazione da compiere consiste nel pre-processare opportunamente i documenti del corpus, in modo da renderli compatibili con una rappresentazione di tipo bag-of-words, come quella implicata dai modelli che abbiamo utilizzato in questo lavoro. Il flusso completo del pre-processing è rappresentato nella Figura 8, ed è stato ottenuto in R utilizzando un insieme di librerie dedicate appositamente al trattamento di dati testuali, supportate da una parte di codice proprietario scritto appositamente per processare il corpus dei Form 10-K (R Core team, 2017). Il significato di alcune operazioni è ovvio, mentre altre hanno uno scopo specifico che non è immediatamente evidente. Per esempio, l'operazione di *stemmizzazione* consiste nel ridurre due token apparentemente diversi alla stessa radice comune (come ad esempio: *computer* e *computing*, che condividono la radice comune *comput*; Manning et al., 2008). Dopo l'operazione di stemmizzazione, ciascuna radice stemmizzata viene ricompletata in base alla forma flessa che compare con la frequenza più elevata a livello di corpus. In questo modo si riduce grandemente la complessità del testo, eliminando la variabilità di natura morfologica.

Figura 8. *Procedimento mediante il quale un Form testuale grezzo viene trasformato in una rappresentazione del tipo 'bag-of-words'.*



Al termine del processo mostrato nella Figura 8, il testo di ciascun Form 10-K si trasforma secondo quanto mostrato nel secondo riquadro della Figura 9, ed è pronto per essere convertito in una matrice documenti-termini, all'interno della quale l'unica quantità rilevante è la frequenza di occorrenza di ciascun token per ciascun documento del corpus.

Figura 9. *Esempio di post-processing. Nel riquadro superiore è mostrata una parte di uno dei Form 10K inclusi nel corpus. Il riquadro inferiore mostra, invece, una parte dello stesso Form, trasformata in una bag-of-words tramite le operazioni di post-processing indicate in dettaglio nella Figura 8.*

```
> writeLines(as.character(corpus[[1]]))
Item 7.  Management's Discussion and Analysis of Financial Condition and          Results of Operations. A
AON engineers, manufactures and markets air-conditioning and heating equipment consisting of standardized
and custom rooftop units, chillers, air-handling units, make-up units, heat recovery units, condensing uni
ts and coils. AAON sells its products to property owners and contractors through a network of manufacturer
s' representatives and its internal sales force. Demand for the Company's products is influenced by nation
al and regional economic and demographic factors. The commercial and industrial new construction market is
subject to cyclical fluctuations in that it is generally tied to housing starts, but has a lag factor of 6
-18 months. Housing starts, in turn, are affected by such factors as interest rates, the state of the econ
omy, population growth and the relative age of the population. When new construction is down, the Company
emphasizes the replacement market. The principal components of cost of goods sold are labor, raw materials
, component costs, factory overhead, freight out and engineering expense. The principal raw materials used
in AAON's manufacturing processes are steel, copper and aluminum. The major component costs include compre
ssors, electric motors and electronic controls. Selling, general, and administrative ("SG&A") costs includ
e the Company's internal sales force, warranty costs, profit sharing and administrative expense. Warranty
expense is estimated based on historical trends and other factors. The Company's warranty on its products
is: for parts only, the earlier of one year from the date of first use or 14 months from date of shipment;
compressors (if applicable), an additional four years, on gas-fired heat exchangers (if applicable), 15 ye
ars, and on stainless steel heat exchangers (if applicable), 25 years. The Company's operations in Burling
```

```
> writeLines(as.character(corpus[[1]]))
item management discussion analysis financial conditions results operations aaon engine manufacturing mark
et air-conditioning heat equipment consisted standards customers rooftop units chillers air-handling units
make-up units heat recoveries units condensing units coil aaon selling products properties ownership contr
actors network manufacturing represent internal sales force demand companies products influenced national
region economic demographic factors commercial industrial new construction market subject cyclical fluctua
tions general tied housing starts lag factors months housing starts turn affected factors interest rate st
atements economic population growth related age population new construction companies emphasizes replaces
market principally component costs goodwill sold labor raw material component costs factori overhead freig
ht engine expense principally raw material used aaon manufacturing process steel copper aluminum majority
component costs included compressors electrical motors electronic control selling general administrative s
ga costs included companies internal sales force warranties costs profit share administrative expense warr
anties expense estimates based historical trends factors companies warranties products parts earlier one y
ear date first used months date shipments compressors applicable additional fourth year gas-fired heat exc
hange applicable year stainless steel heat exchange applicable year companies operations burlington ontari
o canada locations sumach driven consisted square fttx office manufacturing facility acrg tract land offic
e facility aaon income consisted square foot building square fttx manufacturing warehouse space square ftt
x office space locations sales yukon avenue tulsa oklahoma original facility square foot manufacturing war
ehouse building square foot office building expansion facility locations across street original facility s
```

Una precisazione necessaria riguarda l'individuazione del sottoinsieme di training. Anche in questo caso, per ridurre l'importanza del singolo insieme di test considerato, ciascun modello è stato appreso mediante una 5-fold cross-validation, ottimizzando per l'accuratezza nell'insieme di test. Per questa operazione è stato utilizzato il 90% dei documenti disponibili, per un totale di 2975 documenti utiliz-

zati in input nella fase di cross-validation. I rimanenti 331 documenti sono stati utilizzati come *insieme di validazione*. In altre parole, abbiamo ricalcolato le metriche di accuratezza sui nuovi documenti utilizzando il modello ottimale appreso durante la fase precedente, per verificare se l'accuratezza stimata fosse confermata o meno in un insieme di nuovi documenti, mai utilizzati nella fase di apprendimento.

La creazione del vocabolario dei termini attraverso il pre-processing è stata effettuata *prima* della suddivisione del corpus in un insieme di training e un insieme di validazione. Ciò implica una forma di apprendimento semi-supervisionato, nel quale parte dell'informazione 'futura' (ossia i termini che appaiono nei fold di test e nell'insieme di validazione) contribuisce a determinare le frequenze di occorrenza dei documenti di apprendimento. Ciò implica anche che i risultati che presentiamo saranno generalizzabili ad un sistema di previsione nel quale, ogni volta che si dovessero presentare nuovi documenti da classificare, l'intera matrice documenti-termini sia ricostruita sulla base dell'intero corpus disponibile (tenendo conto di questi nuovi documenti non classificati), e i modelli di classificazione vengano riaddestrati sulla base della nuova matrice documenti-termini. Questo approccio è ovviamente differente dal costruire una volta per tutte la matrice documenti-termini sull'insieme di addestramento, e poi calcolare le frequenze di occorrenza, nei documenti di test e di validazione, per i soli token che compaiono nel vocabolario dei termini dell'insieme di addestramento.

Per quanto riguarda la discretizzazione della volatilità, abbiamo usato una suddivisione naturale basata sui quantili. Per effettuare almeno una prima analisi di sensibilità minimale, abbiamo considerato due suddivisioni distinte. La prima prevede due classi ("BASSA" ed "ALTA"), ed è ottenuta suddividendo la distribuzione empirica della log-volatilità in due classi in base alla mediana. La seconda discretizzazione prevede invece tre classi, ossia "BASSA", "MEDIA" ed "ALTA", ottenute in base al primo e al terzo quartile della medesima distribuzione empirica.

Nella Figura 10 sono mostrate le curve di cross-validation per il modello di regressione logistica penalizzato la cui verosimiglianza è espressa nella (22). Tali curve sono utilizzate per determinare il valore ottimale di λ ottimizzando sull'accuratezza, e sono state ottenute prendendo $\alpha = 0.5$ che, come sappiamo, corrisponde ad un miscuglio tra una penalità di tipo ridge ed una di tipo lasso. Ovviamente, nel caso di due sole classi target per la previsione, una volta che le etichette previste nel test set (o nel validation set) dall'algoritmo di classificazione e le relative true label siano state disposte in una matrice di confusione 2×2, l'accuratezza può essere espressa come:

$$\text{accuratezza} = \frac{TP+TN}{TP+TN+FP+FN} = \frac{TP+TN}{N_{\text{test}}}, \tag{26}$$

dove, come è ben noto TP sta per *true positives*, e rappresenta il numero di istanze di test nella positive class che sono state correttamente classificate, e così via. Nel caso di tre classi, l'accuratezza è semplice definibile come la percentuale di istanze correttamente classificate nell'insieme di test (o di validazione). Si noti, tra l'altro, che le curve riportate nella Figura 10 basate, per costruzione, sul tasso di errata classificazione, che è il complemento ad 1 della (26).

Figura 10. *Curve per la determinazione del valore ottimale di* λ *nel modello di regressione logistica penalizzato, la cui verosimiglianza è espressa nella* (22). *In entrambi i casi è stato utilizzato* $\alpha = 0.5$, *corrispondente ad un miscuglio tra la penalizzazione di tipo ridge e una penalizzazione di tipo lasso.*

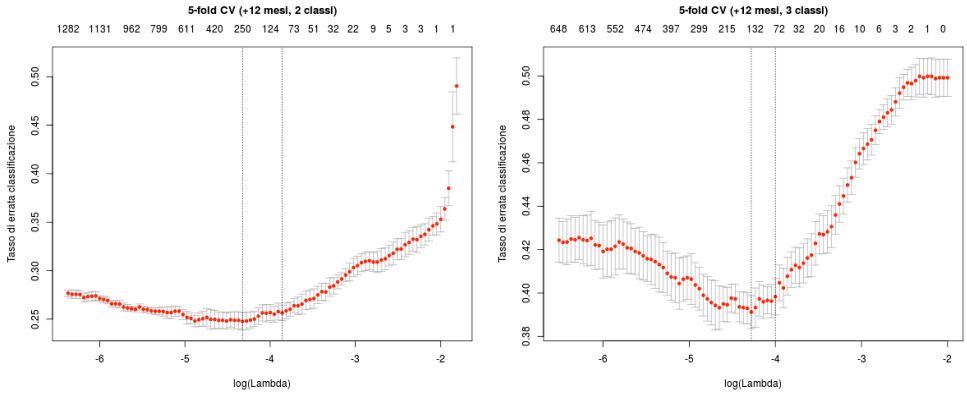
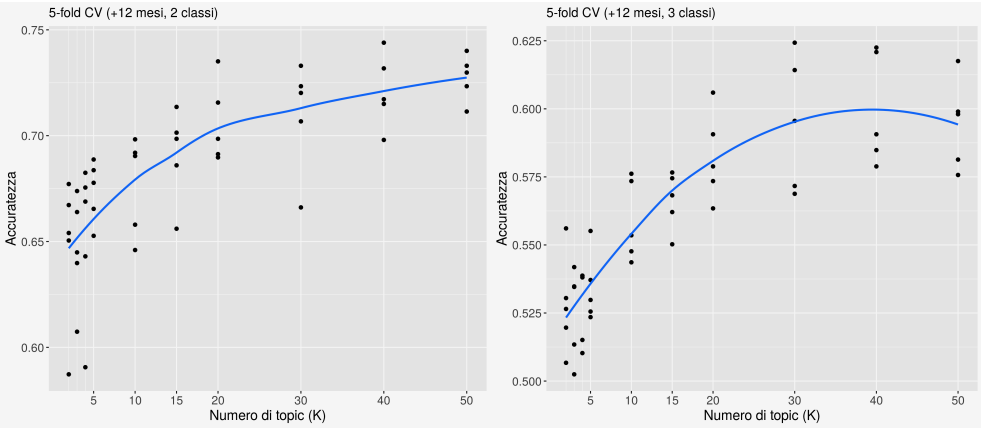


Figura 11. *Curva per la determinazione del valore ottimale di* K *nel modello SLDA, ottenuta inserendo il calcolo dell'accuratezza in una 5-fold cross-validation.*



Allo stesso modo, la Figura 11 riporta la curva di cross-validation utilizzata per determinare il valore ottimale di K (ottimizzando, anche in questo caso rispetto all'accuratezza nell'insieme di test). Si noti che, nel caso di due classi, non esiste un valore massimo chiaramente determinato per la media delle cinque accuratezze (calcolate al variare di K). Anche in questo caso (come già visto nell'esempio basato sul modello non-supervisionato), l'accuratezza media è praticamente identica per $K = 40$ e $K = 50$, e quindi conveniamo di scegliere il modello meno parametrizzato.

I risultati globali del nostro esercizio di previsione sono stati riportati nella Tabella 1. Nel caso di due classi, è possibile calcolare di metriche di accuratezza alternative, che in alcuni casi possono risultare più espressive della semplice accuratezza, e in particolare (Parikh et al., 2008):

$$\begin{aligned} \text{sensibilità} &= \frac{TP}{TP+FN} \\ \text{specificità} &= \frac{TN}{TN+FP} \end{aligned} \quad . \tag{27}$$

La *sensibilità*, nota anche come *recall* o *true positive rate* (TPR) misura la proporzione delle istanze della *positive class* che sono correttamente classificate. Se indichiamo come positive class l'etichetta "ALTA" (ossia siamo maggiormente interessati alle fasi di volatilità elevata), la sensibilità misura la proporzione delle imprese per le quali viene previsto che nei prossimi 12 mesi osserveremo una fase di volatilità elevata, che si è poi effettivamente realizzata. Invece, la specificità è nota anche come *true negative rate* (TNR), e misura la percentuale di istanze della negative class che sono correttamente classificate. Dunque, la specificità misura la capacità del classificatore di individuare correttamente le imprese che non entreranno in una fase di volatilità elevata nei prossimi 12 mesi. Ottimizzare sul test set per la sensibilità o la specificità significa evidentemente perseguire obiettivi differenti, ed esiste un trade-off tra le due misure, nel senso che ottimizzare per una delle due significa, generalmente, ridurre il valore dell'altra.

Anche il coefficiente AUC (*Area Under the Curve*) è particolarmente espressivo, in quanto misura l'area compresa sotto la curva ROC (*Receiver Operating Characteristics*; Fawcett, 2006; Liu, 2011). Tale coefficiente assume sempre un valore compreso tra 0.5 ed 1. Nell'estremo inferiore, ossia quando AUC assume un valore pari a 0.5, il classificatore attuale è indistinguibile da un classificatore casuale, nel quale le istanze di test vengono attribuite alle due classi sulla base dei risul-

tati ottenuti tramite il lancio di una moneta non truccata. Invece, l'estremo superiore, quando AUC assume valore pari ad 1 corrisponde ad al classificatore *perfetto*, nel quale ogni istanza di test viene classificata senza errore. I classificatori empirici reali si situano nel mezzo di questi due estremi.

Tabella 1. *Metriche di accuratezza previsiva relative ai due classificatori utilizzati (Regressione logistica penalizzata e Supervised Latent Dirichlet Allocation). Per ciascun classificatore, i parametri di controllo ottimali sono stati determinati mediante una 5-fold cross-validation ottimizzando rispetto all'accuratezza. Abbiamo poi ricalcolato un insieme di metriche sull'insieme di validazione, utilizzando il modello ottimale appreso durante la fase di cross-validation.*

Regr. logist. penalizzata	**Acc.**	**AUC**	**Sens.**	**Spec.**
5-fold CV (training)				
2 classi	0.75			
3 classi	0.61			
Validazione				
2 classi	0.76	0.85	0.86	0.68
3 classi	0.60			
SLDA	**Acc.**	**AUC**	**Sens.**	**Spec.**
5-fold CV (training)				
2 classi ($K_{opt} = 40$)	0.72			
3 classi ($K_{opt} = 40$)	0.58			
Validazione				
2 classi	0.73	0.73	0.81	0.66
3 classi	0.60			

Per il modello SLDA, i risultati sono stati ottenuti scegliendo l'iper-parametro α della proporzione dei topic θ (ovviamente dal non confondere con il parametro di regolarizzazione α nel modello di regressione logistica penalizzato) in base ad un setting standard, e cioè ponendo $\alpha = 50 / K$. Quanto riportato nella Tabella 1 sembra indicare che il modello di regressione logistica penalizzato ottiene performance sistematicamente superiori rispetto a quelle del modello SLDA, in termini di accuratezza previsiva, al modello SLDA (sia nel caso di 2 classi, che nel caso di 3 classi). Nel prossimo paragrafo, conclusivo, andiamo ad analizzare il significato di questo risultato in maggior dettaglio.

5. Discussione e conclusioni

In questo lavoro abbiamo presentato due approcci alternativi alla stima della volatilità futura di un insieme di asset, basandoci su due diversi classificatori testuali, ed utilizzando come input le informazioni contenute in un corpus di documenti noti come Form-10K, che ciascuna delle aziende oggetto di analisi è stata obbligata a compilare da parte della US-SEC.

Il primo dei due classificatori testuali analizzati è la regressione logistica multinomiale, ossia un classificatore discriminativo nel quale le probabilità a posteriori delle etichette sono apprese, documento per documento, direttamente dal modello sulla base di una opportuna funzione di un insieme di variabili previsive. Nello specifico, per ciascun documento, il feature vector contiene le relative frequenze di occorrenza di ciascun token del vocabolario dei termini. Di questo modello abbiamo utilizzato una versione opportunamente penalizzata, per ridurre la sparsità nel stime e il conseguente overfitting. Il secondo classificatore è invece un classificatore generativo molto più elaborato, noto come Supervised Latent Dirichlet Allocation (SLDA). In quest'ultimo modello, la previsione della volatilità discretizzata è ottenuta non solo attraverso le frequenze di occorrenza dei token, bensì anche attraverso il contenuto semantico latente dei documenti, che non è ristretto ad essere univoco come nel modello di regressione logistica (il modello SLDA, infatti, permette di considerare, più realisticamente, ciascun documento come un miscuglio di topic). A dispetto della sua immensa popolarità, i risultati ottenuti in questo lavoro indicano che il modello SLDA non ottiene una performance superiore, in termini di accuratezza previsiva, alla performance ottenuta dal modello di regressione logistica multinomiale penalizzato. Questo risultato è ancora più preoccupante, se si considera che i tempi di calcolo del modello SLDA sono all'incirca tre ordini di grandezza più elevati di quelli necessari ad ottenere le stime dei parametri del modello di regressione logistica.

Ovviamente, questo risultato deve essere opportunamente valutato. Come in ogni modello bayesiano gerarchico complesso, la specificazione delle distribuzioni a-priori assume un ruolo fondamentale anche per il modello SLDA. In particolare, sarà necessario condurre un'analisi di sensibilità approfondita per il parametro che governa la sparsità dei topic, ossia α, che in questo lavoro è stato settato come $\alpha = 50/K$. Valutare l'accuratezza corrispondente ad una scelta del tipo $\alpha << 1$ (ossia per una distribuzione della proporzione dei topic molto più sparsa) sarà il primo obiettivo da prendere in considerazione nei prossimi lavori. La nostra aspettativa è che con una scelta di questo tipo, l'accuratezza potrebbe migliorare note-

volmente (anche se questa aspettativa dovrà essere opportunamente testata dal punto di vista empirico).

Un'ulteriore alternativa è quella di prendere in considerazione una versione puramente discriminativa del modello SLDA, che non consideri l'etichetta come un ulteriore token del documento, ma che apprenda direttamente quest'ultima sulla base della struttura dei topic (esattamente come nel modello di regressione logistica). Da questo punto di vista, il modello MedLDA (*Maximum Entropy Discimination LDA*), proposto in Zhu et al. (2012), potrebbe costituire un'alternativa promettente. Anche il ruolo del pre-processing (che è governato da una serie di parametri di controllo impliciti) andrebbe analizzato più approfonditamente (Boyd-Graber et. al, 2014). Nel complesso, riteniamo che la ricerca sulle capacità previsive dei modelli a topic latenti debba essere sicuramente approfondita, poiché questi ultimi costituiscono uno strumento, al momento insuperato, per sintetizzare grandi masse di documenti, riducendone la dimensionalità lungo alcune dimensioni facilmente interpretabili. Le modalità in base alla quale è possibile migliorare la capacità previsiva di queste dimensioni non sono al momento note con esattezza, ma questo problema specifico può sicuramente costituire un argomento di ricerca interessante nei prossimi anni.

Riferimenti bibliografici

Awasthi, P., Risteski, A. (2015) On some provably correct cases of variational inference for topic models. In *Advances in Neural Information Processing Systems 28*: 2098–2106.

Blei, D. M. (2012) Probabilistic topic models. *Communications of the ACM*, 55(4): 77–84.

Blei, D.M., Ng, A.Y., Jordan, M.I. (2003) Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022.

Blei, D. M., McAuliffe, J. D (2007) Supervised Topic Models. In *Advances in Neural Information Processing Systems 20*, Proceedings of the Twenty-First Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 3-6, 2007: 121–128.

Blei, D.M., Kucukelbir, A., McAuliffe, J.D. (2017) Variational Inference: A Review for Statisticians. *Journal of the American Statistical Association*, 112(518): 859–877.

Bouchaud, J.P., Potters, M. (2003). *Theory of Financial Risks and Derivatives Pricing: From Statistical Physics to Risk Management*, 2nd ed. Cambridge: Cambridge University Press.

Boyd-Graber, J., Mimno, D., Newman, D. (2014). Care and Feeding of Topic Models: Problems, Diagnostics, and Improvements. In *Handbook of Mixed Membership Models and Their Applications*, edited by Edoardo M Airoldi, David Blei, Elena A Erosheva, and Stephen E Fienberg. CRC Handbooks of Modern Statistical Methods. Boca Raton, Florida: CRC Press.

Chang, J., Boyd-Graber, J., Gerrish, S., Wang, C., Blei, D. M. (2009). Reading Tea Leaves: How Humans Interpret Topic Models. In *Proceedings of the 22Nd International Conference on Neural Information Processing* Systems: 288–296.

Chen, J., Li, K., Zhu, J., Chen, W. (2015) WarpLDA: a Cache Efficient O(1) Algorithm for Latent Dirichlet Allocation. *Eprint arXiv:1510.08628*. Retrived from: https://arxiv.org/abs/1510.08628.

Cumby, R., Figlewski, S., Hasbrouck, J. (1993). Forecasting Volatility and Correlations with EGARCH models. *Journal of Derivatives*, 1(2): 51–63.

Fama, E. (1970). Efficient Capital Markets: A Review of Theory and Empirical Work. *Journal of Finance*, 25 (2): 383–417.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, *27*(8): 861–874.

Foye, J. and Mramor, D. Pahor, M. (2013). The Persistence of Pricing Inefficiencies in the Stock Markets of the Eastern European EU Nations. *Economic and Business Review*, 15(2): 113–133.

Friedman, J., Hastie, T., Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software*, 33(1): 1–22.

Griffiths, T.L., Steyvers, M. (2004) Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(suppl 1), 5228–5235.

Grün, B., Hornik, K. (2011). Topicmodels: An R Package for Fitting Topic Models. *Journal of Statistical Software*, 40(13): 1–30.

Halpern, Y., Horng, S., Nathanson, L.A., Shapiro, N.I., Sontag, D. (2012). A Comparison of Dimensionality Reduction Techniques for Unstructured Clinical Text. *ICML 2012 Workshop on Clinical Data Analysis*.

Hastie, T., Tibshirani, R., Friedman, J. (2009). *The Elements of Statistical Learning* (2nd ed.). New York, NY: Springer New York.

Hoerl, A.E, Kennard, R.W. (2000). Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics* 42 (1): 80–86.

James, G., Witten, D., Hastie, T., Tibshirani, R. (2013). *An Introduction to Statistical Learning*. Springer. New York.

Kogan, S., Levin, D., Routledge, B.R., Sagi, J.S., Smith, N.A. (2009). In: *NAACL '09 Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Boulder, Colorado: 272–280.

Liu B. (2011). *Web data mining exploring hyperlinks, contents, and usage data, 2nd Edition*. Springer, Berlin.

Liu, Z., Zhang, Y., Chang, E.Y., Sun, M. (2011). PLDA+: Parallel Latent Dirichlet Allocation with Data Placement and Pipeline Processing. ACM Transactions on Intelligent Systems and Technology, 2(3):1–26.

Manning, C. D., Raghavan, P., Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press.

Nigam, K., Mccallum, A. K., Thrun, S., Mitchell, T. (2000). Text Classification from Labeled and Unlabeled Documents using EM. *Machine Learning*, 39(2/3), 103–134.

Ng, A.Y., Jordan, M.I. (2001). On Discriminative vs. Generative Classifiers: A comparison of logistic regression and naive Bayes. In *Advances in Neural Information Processing Systems 14*: 841–848.

Parikh R., Mathai A., Parikh S., Chandra Sekhar G., Thomas, R. (2008). Understanding and using sensitivity, specificity and predictive values. *Indian Journal of Ophthalmology*, *56*(1): 45–50.

Poon, S., Granger, C.W.J. (2003). Forecasting Volatility in Financial Markets: A Review. *Journal of Economic Literature*, 41(2): 478–539.

Porteous, I., Newman, D., Ihler, A., Asuncion, A., Smyth, P., Welling, M. (2008). Fast collapsed Gibbs sampling for Latent Dirichlet Allocation. In *Proceeding of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD 08*: 569–577.

R Core Team. (2017). *R: A Language and Environment for Statistical Computing*. Vienna, Austria. https://www.r-project.org/.

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1): 1–47.

Sontag, D., Roy, D. (2011). Complexity of Inference in Latent Dirichlet Allocation. In *Advances in Neural Information Processing Systems 24*: 25th Annual Conference on Neural Information Processing Systems 2011. Proceedings of a meeting held 12-14 December 2011, Granada, Spain: 1008–1016.

Speh, J., Muhic, A., Rupnik, J. (2013). Algorithms of the LDA model. *Eprint arXiv:1307.0317*. Retrieved from http://arxiv.org/abs/1307.0317

Steyvers, M., Griffiths, T. (2007). Probabilistic Topic Models. In T. Landauer, D. Mcnamara, S. Dennis, & W. Kintsch (Eds.), *Handbook of Latent Semantic Analysis*. Lawrence Erlbaum Associates: 427–448.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society B*, 58(1), pages 267–288.

Tzikas, D., Likas, A., Galatsanos, N. (2008). The variational approximation for Bayesian inference. *IEEE Signal Processing Magazine*, 25(6): 131–146.

Wainwright, M.J., Jordan, M.I. (2007). Graphical Models, Exponential Families, and Variational Inference. *Foundations and Trends® in Machine Learning*, 1(1–2): 1–305.

Wallach, H.M., Mimno, D.M., McCallum, A. (2009a). Rethinking LDA: Why Priors Matter. In: *Advances in Neural Information Processing Systems 22*: 1973–1981.

Wallach, H.M., Murray, I., Salakhutdinov, R., Mimno, D. (2009b). Evaluation Methods for Topic Models. In *Proceedings of the 26th Annual International Conference on Machine Learning*: 1105–1112.

Wang, C., Blei, D.M., Li, F.F. (2009). Simultaneous image classification and annotation. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*: 1903–1910.

Zhang, C., Kjellström, H. (2015). How to Supervise Topic Models. In L. Agapito, M. M. Bronstein, C. Rother (Eds.), *Computer Vision - ECCv 2014 Workshops*, 500–515. Springer International Publishing.

Zhu, J., Ahmed, A., Xing, X.P. (2012). MedLDA: Maximum Margin Supervised Topic Models. *Journal of Machine Learning Research* 13: 2237–78.

Zou, H., Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B*, 67(2): 301–320.

UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO
Metodi e analisi statistiche 2017
DIPARTIMENTO DI ECONOMIA E FINANZA
ISBN: 978-88-6629-013-1
pp. 103-126

Studio di relazioni tra serie storiche tramite analisi co-spettrale

Francesco D. d'Ovidio[1]*, Najada Firza[2], Ernesto Toma[1]

[1] *Università degli studi di Bari Aldo Moro (Italy),*
[2] *Università Nostra Signora del Buon Consiglio (Tirana, Albania)*

Riassunto: In questo lavoro si presenta una tecnica esplorativa per identificare relazioni tra serie storiche in grado di fornire informazioni su eventuali situazioni di antecedenza tra le medesime, che potrebbero costituire premessa per relazioni di dipendenza che il classico studio delle correlazioni incrociate non è in grado di identificare. La soluzione qui proposta si basa su una trasformazione dei dati osservati in serie di Fourier e sull'analisi congiunta di due serie tramite analisi co-spettrale. Viene portato un esempio applicato allo studio di serie finanziarie che mostra le particolarità e l'utilità di tale tecnica a fini di successive analisi.

Keywords: Serie storiche, Correlazioni incrociate, Trasformazione di Fourier, Spettrogramma, Analisi co-spettrale, Guadagno incrociato.

1. Introduzione

Nelle analisi delle serie temporali è usuale valutare le mutue relazioni fra esse tramite la tecnica della correlazione incrociata (o *cross-correlation*), che rappresenta appunto la misura di similitudine di due serie di dati osservati nel tempo come funzione di uno spostamento o traslazione temporale applicata ad uno di essi.

Tuttavia, quando l'obiettivo dell'analisi non è la generica inter-relazione tra due (o più) serie di dati, bensì la ricerca di quale tra queste possa essere "antecedente" e quale invece "conseguente" (implicando quindi concetti che spesso sono propedeutici alla causalità, pur non implicandola automaticamente), lo studio delle correla-

* Autore corrispondente: francescodomenico.dovidio@uniba.it
Il lavoro qui presentato è frutto di un progetto comune, ma E. Toma ha curato l'Introduzione, F. D. d'Ovidio ha provveduto alla redazione dei paragrafi 2-3, mentre N. Firza ha redatto i paragrafi 4-5.

zioni incrociate si rivela strumento certamente necessario, ma non sufficiente a fornire le informazioni richieste (utilizzabili direttamente o come *esplorazione* dei fenomeni studiati). Dovendo appunto affrontare il problema di identificare serie temporali antecedenti e conseguenti al fine di successive analisi mirate, si è deciso di seguire i suggerimenti di un lavoro di Delvecchio (1974), ponendo attenzione all'analisi co-spettrale, ossia la trasformazione dei dati osservati in serie di Fourier e sulla loro analisi congiunta a due a due con tecniche operanti nel dominio frequentistico invece che temporale, ove agisce invece l'analisi delle correlazioni incrociate: tecniche ben note, soprattutto in Teoria dei segnali, ma di cui forse si sottovaluta la valenza esplorativa nelle scienze economiche e (soprattutto) sociali.

In questo studio sono esposti dapprima i fondamenti matematici e metodologici dell'analisi spettrale (§ 2), per poi approfondire le specificità dell'analisi del co-spettro (§ 3); segue una dettagliata disamina dell'applicazione di tale tecnica allo studio di serie finanziarie (§ 4) e, infine, alcune considerazioni conclusive (§ 5).

2. Analisi delle serie temporali e trasformazione in serie di Fourier

Per verificare la rilevanza statistica delle variazioni cicliche in una serie temporale stazionaria[1] è possibile fare ricorso, in prima istanza, all'analisi del correlogramma, ossia della serie dei coefficienti di autocorrelazione $r_h = C_h/C_0$, in cui le C_h (con $h = 0, 1, 2, ..., m$) sono le autocovarianze di slittamento della serie detrendizzata, supposta stazionaria. La migliore determinazione di dette autocovarianze, per ogni ritardo (o *lag*) h, in caso di serie temporali stazionarie con media nulla (cfr., ad es., Bendat & Piersol, 1966, 2004), può essere fornita da:

$$C_h = \frac{1}{s-h}\sum_{t=1}^{s-h}\left(x_t - \bar{x}\right)\left(x_{t+h} - \bar{x}\right) = \frac{1}{s-h}\sum_{t=1}^{s-h} x_t x_{t+h} \quad .$$

Per ottenere serie di dati stazionari, in genere, è necessario intervenire sul trend tramite differenza rispetto a una funzione analitica o tramite differenze prime. La

[1] Una serie storica è detta "stazionaria in senso forte" (o *in senso stretto*) quando sono indipendenti dal tempo la media, la varianza e tutti i momenti di ordine superiore al secondo; è invece detta "stazionaria in senso debole" o *largo* (o anche "stazionaria del 2° ordine"), se la sua media è indipendente dal tempo, e dunque è costante, e se l'autocovarianza (o, del pari, l'autocorrelazione) dipende solo dalla differenza fra i tempi in cui è misurata (cfr., ad es., Oppenheim & Verghese, 2010), ossia:

$$E[x(t)] = E(x); \qquad C_{xx}(t_i, t_k) = C_{xx}(t_i - t_k) \qquad \forall\ t,\ t_i,\ t_k.$$

Complementare al precedente (ma non del tutto opposto) è il concetto di *integrazione* di una serie storica: una serie $\{X_t\}$ non stazionaria si dice integrata di ordine 1 se è stazionaria la serie temporale definita dalle differenze prime $\delta X_t = X_t - X_{t-1}$.

seconda tecnica risulta qui preferibile in quanto consente di ottenere una serie autostazionaria in media senza complesse analisi funzionali e senza alterare la struttura di serie evolutive di tipo polinomiale (cfr., ad es., Yaglom, 1958)[2]; inoltre, ad essa si riallacciano molti concetti e metodi di analisi delle serie temporale, come, per esempio, il concetto di integrazione di una serie.

2.1 Verifica della stazionarietà delle serie temporali

Per quanto riguarda la stazionarietà (condizione necessaria per l'analisi) e il suo quasi-contrario, l'integrazione, ad affiancare l'osservazione grafica delle serie differenziate troviamo vari test, due dei quali vengono utilizzati in questo studio.

Il più noto (essendo stato anche il primo test proposto su questo argomento[3]) è il test di Dickey-Fuller aumentato (ADF), il quale, dopo aver escluso le componenti deterministiche, confronta un processo stocastico (*random walk*) contro un AR(p) stazionario. Per applicare il test ADF, basato sulla statistica T di una regressione ausiliaria (a un prefissato livello α di significatività), bisogna prestabilire il numero k di ritardi della variabile differenziata da porre come limite della sommatoria di termini differenziati nella funzione di regressione

$$\Delta y_t = \mu + \beta t + \varphi y_{t-1} + \sum_{j=1}^{k} \delta_j \Delta y_{t-j} + \varepsilon_t \ ,$$

ove μ è una costante, β il coefficiente di un ipotetico trend residuo, φ sarebbe il coefficiente della radice unitaria della serie originale mentre i coefficienti δ_j rappresentano l'autoregressione della serie differenziata. La selezione del numero dei ritardi k può essere fatta per mezzo del criterio di informazione di Akaike, di Schwarz oppure di Hannan-Quinn, stimando diversi modelli e scegliendo il valore di k che fa rilevare il minimo valore della statistica prescelta[4].

Il test ADF è unidirezionale, con ipotesi di base H_0: $\varphi = 0$ e ipotesi alternativa H_1: $\varphi < 0$. Sotto l'ipotesi nulla, y_t deve essere differenziata almeno una volta per poter essere considerata stazionaria. Sotto l'ipotesi alternativa, y_t è già stazionaria

---

[2] La qual cosa si risolve spesso in mancanza di stazionarietà sul medio-lungo termine. Si tenga conto, peraltro, che i filtri alle differenze prime, pur lasciando invariate eventuali periodicità delle serie (non distorsione in frequenza), riducono l'ampiezza di tutte ciclicità di lungo periodo ed esaltano quelle con periodicità breve (distorsione in ampiezza); cfr., ad es., Battaglia, 2007.

[3] Inizialmente limitato, come test DF, all'ipotesi base che una serie storica sia espressione di un pro–cesso integrato del 1° ordine (Dickey e Fuller, 1979); la metodologia è stata poi estesa anche agli ordini superiori, ossia ai ritardi successivi, che è appunto il test DF "aumentato" (Said e Dickey, 1984).

[4] Il software GRETL, a cui si fa riferimento per i test di stazionarietà in questo studio, effettua tale valutazione in modo automatico.

e non richiede ulteriori differenziazioni[5]: di conseguenza, valori ampiamente e significativamente *negativi* della statistica test DF_τ (e dunque valori $p < \alpha$) implicano il rifiuto dell'ipotesi di base. Va da sé che, per il confronto con i valori critici tabulati da Dickey e Fuller tramite tecniche di simulazione, la statistica test viene standardizzata tramite rapporto con il suo s.e.: $DF_\tau = \hat{\varphi}/\text{s.e.}(\hat{\varphi})$.

Questo è un test molto flessibile, in quanto la forma della sua distribuzione (e quindi l'insieme dei valori critici per il test) dipende dai vincoli che vengono posti su alcuni dei parametri della regressione ausiliaria (cfr., ad es., Hamilton, 1994).

Il modello completo descritto nella precedente equazione è da utilizzare quando si presume che la serie storica non sia stazionaria in media e in varianza (abbia cioè un trend con variazioni cicliche o congiunturali abbastanza differenti); se si presume una non stazionarietà in varianza con media stazionaria non nulla, è invece opportuno fissare a zero il parametro β; ove la media sia invece nulla (e si presume non stazionarietà in varianza), sarà pari a zero anche il parametro μ. Si noti che, fissando a zero tutti i parametri δ_j (ovvero ponendo $k = 0$, cioè nessun ritardo), il test ADF si trasforma nel test Dickey-Fuller per l'integrazione del 1° ordine.

Il test KPSS (Kwiatkowski *et al.*, 1992), seppur meno potente del precedente, ne risolve qualche incertezza applicativa (invero, ADF fornisce risultati inattendibili se la serie temporale non è stazionaria né integrata, ossia se anche la serie differenziata è non stazionaria). Questo test, infatti, ha come ipotesi di base la *stazionarietà* della serie studiata e non la sua integrazione: la sua *ratio* è che, esprimendo gli elementi della serie come $y_t = \mu_y + u_t$, *se e solo se* u_t rappresenta un processo stazionario a media nulla, allora la media campionaria di y_t è uno stimatore consistente di μ_y e, inoltre, la varianza di lungo periodo di u_t è un numero finito.

In pratica, stimate le u_t con $e_t = y_t - \bar{y}$, se ne calcola innanzitutto le autocovarianze empiriche $\hat{c}$ dall'ordine $-m$ all'ordine m, ove m (detto *larghezza di banda*) deve essere abbastanza grande da consentire la persistenza a breve termine delle e_t ma non troppo grande rispetto alla lunghezza T della serie: gli autori del software GRETL consigliano valori di m pari a $k = \text{INT}[(4\times T/100)^{0,25}]$ o non molto maggiori (Cottrell e Lucchetti, 2017). In questo lavoro, è stato scelto $m = 2\times k = 16$.

Partendo dalle autocovarianze, si stima la varianza di lungo periodo con

$$\bar{\sigma}^2 = \sum_{i=-m}^{m} \left(1 - \frac{|i|}{m+1}\right) \cdot \hat{c}_i \,.$$

[5] La *ratio* del test ADF è che, se la serie è integrata, allora il livello di ritardo della serie $\{y_{t-1}\}$ non fornirà informazioni rilevanti per prevedere i cambiamenti in y_t oltre a quelle fornite dalle differenze ritardate $\{\Delta y_{t-j}\}$. In questo caso, dunque, potrà porsi $\varphi = 0$, che è appunto l'ipotesi di base H_0.

La statistica test KPSS è:

$$\eta = \frac{\sum_{t=1}^{T}\left(\sum_{i=1}^{t} e_i\right)^2}{T^2\,\overline{\sigma}^2}$$

e, sotto l'ipotesi di base, ha una distribuzione asintotica indipendente da parametri di disturbo, definita dagli Autori con metodi di simulazione. L'ipotesi di stazionarietà va rigettata se il valore empirico del test risulta superiore al valore critico asintotico tabulato per il livello di significatività prefissato ($\eta_{0,10}$=0,347; $\eta_{0,05}$=0,463; $\eta_{0,025}$=0,574; $\eta_{0,01}$=0,739).

2.2 Correlogrammi e spettrogrammi

Nell'analisi di autocorrelazione, benché in serie di cospicua numerosità possano essere posti anche slittamenti pari a metà della lunghezza della serie medesima, per assicurare una buona precisione il massimo slittamento m non dovrebbe essere superiore a $s/3$, ove s è la numerosità dei termini della serie temporale (Malinvaud, 1971).

I coefficienti empirici di autocorrelazione godono di una proprietà molto utile: oltre ad essere, com'è noto, invarianti (non dipendendo, quindi, né dalle unità di misura delle variabili considerate, né dal punto di origine), essi seguono in caso di incorrelazione una legge di distribuzione approssimativamente normale (Hannan, 1960), per cui è possibile tracciare gli intervalli di confidenza (dati da $\pm 2\hat{\sigma}$ o da $\pm 3\hat{\sigma}$)[6] entro cui dovrebbero essere compresi detti coefficienti quando i residui si distribuiscono casualmente (per esempio, una serie autostazionaria dovrebbe presentare coefficienti tutti non significativi, ad eccezione eventualmente del primo).

L'analisi dei correlogrammi è di buon ausilio nell'analisi dei cicli[7], ma spesso essa non permette di andare oltre semplici indicazioni di massima. Infatti, in presenza di cicli temporali di diverse frequenze, ciascun coefficiente di autocorrelazione è influenzato simultaneamente da tutti i cicli esistenti e non sempre l'analisi della funzione di autocorrelazione parziale (PACF) è in grado di risolvere il pro-

[6] Ove $\hat{\sigma} = 1/\sqrt{s-h}$ sotto l'ipotesi che le serie siano assimilabili a rumore bianco, con correlazione nulla fra i termini della serie slittati di un intervallo h>0 (cfr. Kendall, 1973; Kendall & Stuart, 1976).

[7] In effetti, il correlogramma (il cui uso nell'analisi dei cicli è citato anche in Kendall & Stuart, 1976) può essere considerato uno strumento abbastanza efficace per individuare le componenti periodiche presenti in una serie storica stazionaria, in quanto elevati valori positivi dei coefficienti di autocorrelazione per un *lag*>1 possono indicare una componente periodica di periodo pari al *lag* medesimo. Eventuali valori negativi dei coefficienti di autocorrelazione lasciano presumere, invece, l'esistenza di componenti cicliche con semiperiodo pari a detto *lag*. Com'è evidente, infatti, un ciclo di periodo *t* implica coefficienti di autocorrelazione positivi per *lag*=t, *lag*=2t, ecc., ma anche coefficienti negativi per *lag*=t/2, *lag*=3t/2 e così via (Cfr., ad esempio, Malinvaud, 1971).

blema[8]. Inoltre, in caso di serie con un numero di termini ridotto o elevatissimo la significatività dei coefficienti può risultare rispettivamente sottostimata o sovrastimata. Tali difetti possono rendere considerevolmente meno utili (o comunque meno chiari) i risultati ottenuti, ragion per cui occorre valutare volta per volta l'applicabilità del metodo.

Esistono, tuttavia, metodologie statistiche che forniscono informazioni più robuste, fra cui l'analisi dello spettrogramma (o analisi spettrale)[9]. Si tratta di tecniche note, facilmente applicabili al fenomeno oggetto della presente nota.

Periodogramma e spettrogramma permettono di stimare l'ammontare della varianza della serie spiegata da vari cicli di differente frequenza, la cui combinazione genera la serie medesima; essi sono definiti all'interno del cosiddetto "dominio frequenziale", ben diverso come proprietà e sviluppi dal "dominio temporale" in cui sono definiti i dati di partenza e anche le funzioni di autocorrelazione.

Vi è certamente un preciso rapporto fra queste ultime e la rappresentazione delle varie densità spettrali, cosicché correlogramma e spettrogramma possono fornire, al limite, le medesime informazioni. Tuttavia, quando sono presenti nella serie più componenti cicliche o quando vi sia una considerevole componente erratica, l'analisi spettrale risulta molto fruttuosa, soprattutto se utilizzata complementarmente all'analisi delle autocorrelazioni.

Innanzitutto, la robustezza delle soluzioni identificate con l'analisi spettrale *non dipende* dalla numerosità dei dati disponibili[10], poiché detta analisi è basata su trasformazioni ed elaborazioni puramente matematiche e non su ipotesi circa i processi

---

[8] La funzione di autocorrelazione parziale ha lo scopo di valutare ogni coefficiente autoregressivo al netto dell'effetto dei precedenti, riducendo così l'eventualità di considerare significative componenti di "autoregressione spuria", ossia dovuta alla combinazione di due o più componenti autoregressive; si pensi a una serie che presenti componenti autoregressive significative per *lag* 2 e per *lag* 5, ma anche per *lag* 10: ora, il problema identificare quanta parte della componente con *lag* 10 è dovuta al "battimento" tra la componente con lag 2 e quella con lag 5.

[9] Detta forma di analisi si basa sul principio, dimostrato da Fourier nel 1822, che qualsiasi funzione periodica (ossia con oscillazioni più o meno ricorrenti attorno ad un determinato livello medio) può essere espressa come combinazione lineare di un numero infinito di funzioni sinusoidali con differenti frequenze, dette *armoniche*. Analogamente, come ha dimostrato Wold (1954), l'analisi della periodicità di una serie discreta con numero di termini s può essere effettuata esprimendo la serie medesima come combinazione lineare di p armoniche sinusoidali, ove $p=s/2 \in \mathrm{N}$. Le varie armoniche avranno frequenze $\varphi_k=\frac{1}{s}, \frac{2}{s}, \ldots, \frac{p}{s} \simeq 0{,}5$ e, rispettivamente, periodi $\psi_k= s, \frac{s}{2}, \ldots, \frac{s}{p} \simeq 2$, ove si ha il segno di uguaglianza solo se s è pari. Cfr., ad es., Malinvaud, 1971; Vajani, 1980.

[10] Dalla lunghezza delle serie di dati dipende, invece, il numero di soluzioni identificate: il massimo periodo determinabile (ossia l'armonica più lunga e con minor frequenza) non supera infatti la metà di tale lunghezza, come si evince dalla nota precedente. Inoltre, se i dati di una serie sono troppo pochi, i test statistici per verificarne preventivamente la stazionarietà (anche solo del 2° ordine, per non parlare di quelli più articolati come ADF e KPSS) possono rivelare una potenza insufficiente.

generatori della serie, e dunque è quasi interamente non parametrica (IBM Corp., 2012). Inoltre, dati i presupposti dell'analisi di Fourier, la quale stabilisce che la correlazione fra le funzioni sinusoidali utilizzate è sempre nulla, i singoli termini di un periodogramma sono matematicamente (e dunque anche statisticamente) indipendenti fra loro. Infine, è sempre possibile ricostruire la serie data a partire dai coefficienti spettrali, il che implica che si tratta di una analisi senza perdita di informazioni[11].

Per quanto riguarda i riferimenti metodologici dell'analisi spettrale, basti tenere presente che, a partire da un processo stocastico stazionario $\{X_t\}$, come possono essere giudicate le serie residuali considerate nel presente studio, si definisce con la notazione $I_k = a_k^2 + b_k^2$ il k.mo termine del periodogramma del processo[12].

Tuttavia, è generalmente poco opportuno attribuire significato ad ogni picco del periodogramma, per cui, allo scopo di ridurre la varianza e le componenti di disturbo, si preferisce applicare al periodogramma determinate funzioni di spianamento. Tali trasformazioni sono chiamate "finestre", e corrispondono, in genere, a medie mobili di tre o più termini del periodogramma.

Il risultato della trasformazione del periodogramma nel punto k è detto *funzione di densità spettrale*[13], e la sua rappresentazione cartesiana è appunto lo *spettro-*

---

[11] Come enuncia il teorema di Stuart (1961), "gli spettri si compongono linearmente"; inoltre i coefficienti spettrali sono *unici*, ossia esiste una ed una sola combinazione di essi che comprende tutte le informazioni sulla serie data.

[12] I coefficienti che costituiscono I_k, nella lettura proposta da Malinvaud (1971), sono forniti da: $a_k = \sqrt{\frac{2}{s}} \sum_{t=1}^{s} x_t \cos\frac{2\pi t}{s}$, $b_k = \sqrt{\frac{2}{s}} \sum_{t=1}^{s} x_t \operatorname{sen}\frac{2\pi t}{s}$; in tale contesto, si avrà:

$$a_0 = \frac{1}{\sqrt{s}} \sum_{t=1}^{s} x_t , \qquad a_p = 1/\sqrt{s} \cdot \sum_{t=1}^{s} (-1)^t x_t , \qquad b_0 = b_p = 0 .$$

I_k rappresenta, inoltre, il contributo che la componente sinusoidale di periodo s/k apporta alla somma dei quadrati delle x_t. Sviluppandone l'espressione, nell'ipotesi che la serie osservata abbia media zero, si ottiene: $I_k = 2 \sum_{h=-s}^{s} C_h \cos\frac{2\pi kh}{s}$, ove le C_h sono le autocovarianze di slittamento della serie. Detta equivalenza non si applica a I_o (nullo per ipotesi) né, se il numero di termini della serie è pari, al p.mo termine del periodogramma, ossia $I_p = \sum_{h=-s}^{s} C_h \cos(\pi h)$.

[13] Una stima diretta di tale funzione è fornita da $L_k = \frac{1}{2\pi}\left[C_0 + \sum_{h=1}^{m} C_h \cos(\omega_k h)\right]$, ma per ridurre l'influenza di frequenze estranee alla banda di spettro considerata è utile moltiplicare nella [2] le C_h per un filtro "passa banda" λ_h (Malinvaud, 1971). Utilizzando ad esempio la *finestra di Tukey-Hanning*, porremo $\lambda_h = \frac{1}{4\pi}\left(1 + \cos\frac{\pi h}{m}\right)$, ottenendo una buona perequazione dei termini dello spettro senza eccessiva distorsione. Detta stima della densità spettrale equivale, in pratica, a calcolare la media mobile centrata delle L_k, ossia $f(\omega_k) = 0{,}25\ L_{k-1} + 0{,}50\ L_k + 0{,}25\ L_{k+1}$, ove $L_{-1} = L_1$ e $L_{m+1} = L_{m-1}$ (cfr. Bendat e Piersol, 1966, 2004). Si tenga presente che le serie livellate sono sempre caratterizzate da una certa variazione di bassa frequenza, ampiamente compensata, tuttavia, dalla riduzione della variazione casuale, o "rumore bianco".

gramma. Ricordando che la densità spettrale è il risultato di una funzione di spianamento del k-mo termine del periodogramma, assimilabile ad una media mobile centrata, è evidente che di tutti gli spettri di sufficiente entità vanno presi in considerazione, a fini interpretativi, soprattutto quelli che rappresentano punti di massimo relativo dello spettrogramma.

Nel seguito, i grafici riportano anche gli spettri in corrispondenza della frequenza nulla (a cui corrisponderebbe, matematicamente, un periodo di tempo non quantificabile ma tendenzialmente infinito), al solo scopo di evidenziare l'andamento complessivo dello spettrogramma e, soprattutto, l'effettivo contributo fornito dalla singola componente alla variabilità della serie residua. Il valore dell'ordinata dello spettro che misura detto contributo è espresso, per maggiore chiarezza, in termini di devianza della serie, cosicché si avrà, in base al teorema di Stuart:

$$\frac{f(0)}{2} + \sum_{k=1}^{m-1} f(\omega_k) + \frac{f(\omega_m)}{2} = DEV(X) .$$

3. Correlazione incrociata e analisi co-spettrale

In probabilità e statistica, dati due processi stocastici $X=\{X_t\}$ e $Y=\{Y_t\}$, la covarianza incrociata (*cross-covariance*) è una funzione che restituisce la covarianza di ciascun processo con l'altro a coppie di punti temporali (t, s): ossia, se $E(X_t) = \mu_{x(t)}$ e $E(Y_t) = \mu_{y(t)}$, allora la covarianza incrociata è data da

$$Cov(X_t , Y_s) = C_{xy}(t , s) = E[(X_t - \mu_{x(t)}) (Y_s - \mu_{y(s)})] = E(X_t Y_s) - \mu_{x(t)} \mu_{y(s)} ,$$

e, ove X e Y siano processi stocastici stazionari in senso debole, sarà senz'altro

$$E(X_t)=\mu_x , E(Y_s)=\mu_y , Cov(X_t, X_s)=E(X_t X_s)-\mu^2_x , Cov(Y_t, Y_s)=E(Y_t Y_s)-\mu^2_y ;$$

ma, per parlare di *processo stocastico bivariato stazionario* (al secondo ordine), occorre verificare che anche i momenti incrociati $Cov(X_t , Y_s)$ dipendano solo da (t–s), condizione per cui vale la proprietà: $Cov(X_t , Y_s) = E(X_t Y_s) - \mu_x \mu_y$, onde evitare distorsioni dei risultati[14].

[14] Una delle condizioni aggiuntive che è necessario sottoporre a verifica è dunque la *cointegrazione* tra le serie (ossia l'esistenza di almeno una loro combinazione lineare *non banale* che risulti stazionaria), perché, se ciò non avviene, l'errore di regressione tra esse sarà integrato e avrà varianza crescente al crescere della componente temporale, portando a rifiutare l'ipotesi di assenza di relazione lineare, al crescere di *n*, anche quando tale relazione non sussiste: a tale scopo, Engle e Granger (1987) proposero una semplice procedura (stimare una regressione tramite OLS e di applicare il test ADF sui residui della regressione per verificarne l'integrazione) che ovviamente in caso di correlazione incrociata dovrebbe essere replicata sia facendo regredire la v.c. Y dalla X che viceversa.

Si tenga conto che la funzione di covarianza incrociata non è una *funzione pari*, essendo generalmente $\mathrm{Cov}(X_t, Y_s) \neq \mathrm{Cov}(X_s, Y_t)$, *non* è semidefinita positiva e, infine, non presenta (come avviene invece per l'autocovarianza) un massimo per il lag nullo ($h = t–s = 0$), ma il suo valore massimo può presentarsi in corrispondenza di qualsiasi ritardo (Battaglia, 2007). Se, inoltre, si vogliono confrontare le relazioni tra diverse coppie di processi stocastici stazionari occorre costruire un indice standardizzato rispetto al massimo, analogo al coefficiente di correlazione, che misura quindi la correlazione incrociata (*cross-correlation*):

$$r_{xy}(h) = \frac{\mathrm{Cov}(X_t, Y_{t+h})}{\sqrt{\mathrm{Var}(X_t) \cdot \mathrm{Var}(Y_{t+h})}} = \frac{C_{xy}(h)}{\sqrt{\sigma_x^2 \cdot \sigma_y^2}} \ .$$

La correlazione incrociata è dunque una misura della somiglianza di due serie temporali, ed è funzione del tempo tra le singole osservazioni delle serie. Si noti che i valori di ampiezza della *cross-correlation* non sono pienamente normalizzati[15], ossia la correlazione massima (positiva o negativa) sarà inferiore a 1 in valore assoluto.

Peraltro, il risultato della funzione di correlazione incrociata mostra solo metà dell'insieme di coefficienti di correlazione che si vuole studiare (tempi t, t+1, t+2,…, t+h); per esplorare l'altro lato, è sufficiente commutare l'ordine delle serie (correlazione B incrociata con A invece di A incrociata con B).

Ove la *cross-correlation* risulti inapplicabile (ad esempio per fallimento del test di cointegrazione, oppure per la ridotta numerosità delle serie e conseguente carenza di potenza dei test statistici applicabili), è possibile acquisire affidabili informazioni aggiuntive dall'analisi co-spettrale (*cross-spectral analysis*), che, come quella spettrale, agisce nel dominio frequentistico ed è dunque una tecnica quasi non parametrica.

Un altro metodo per verificare la cointegrazione, più rigoroso e senza necessità di replicazione con termini invertiti, è stato proposto da Johansen (1995), ma esso ha una formulazione molto complessa, al contrario del suo utilizzo, per cui se ne demanda la descrizione ai testi in letteratura (Harris, 1995; Johansen, 1995, 2000). Basti sapere, come ben chiariscono gli sviluppatori di Gretl (Cottrell e Lucchetti, 2017), che la procedura di Johansen, per stabilire il numero di vettori di cointegrazione del sistema, fa un uso congiunto di due test distinti: il test "*λ-max*", per le ipotesi sui singoli autovalori (ordinati dal maggiore al minore, avendo come ipotesi di base H_0 che vi sia un autovalore $\lambda_i = 0$, ossia *non stazionarietà delle singole serie*), e il test "*traccia*" per le ipotesi congiunte ($\lambda_j = 0$ per ogni $j \geq i$, ossia *non stazionarietà di una loro combinazione lineare*). La distribuzione asintotica dei test varia a seconda dei vincoli posti sulle componenti deterministiche, come nel test ADF.

[15] I coefficienti di correlazione incrociata, infatti, sono proporzionali alle relative covarianze, e dunque, non essendo funzioni pari, il loro massimo funzionale è $|r_{xy}(h)|<1$. Inoltre, in corrispondenza del *lag* nullo la correlazione incrociata sarà esattamente pari al coefficiente di correlazione lineare tra le due variabili: $r_{xy}(0) = r_{xy}$.

Infatti, tenendo conto che gli spettrogrammi (univariati) di due processi $\{X_t\}$ e $\{Y_t\}$ descrivono ciascuna di dette serie come combinazione di componenti cicliche (di frequenza ω compresa tra 0 e π) *incorrelate* tra loro, si dimostra che eventuali relazioni lineari tra $\{X_t\}$ e $\{Y_t\}$ agiscono solo tra le componenti di egual frequenza (Battaglia, 2007; Bendat e Piersol, 1966), per cui le ampiezze delle armoniche di $\{X_t\}$ e $\{Y_t\}$ sono mutualmente incorrelate per tutte le componenti di frequenza differente.

La funzione di densità spettrale congiunta di $\{X_t\}$ e $\{Y_t\}$ può essere ricavata come trasformata di Fourier della covarianza incrociata, ma in genere ha valori complessi e quindi va rappresentata separando la parte reale e quella immaginaria:

$$S_{xy}(\omega) = c(\omega) + {}_i q(\omega).$$

La parte reale $c(\omega)$ è una funzione *pari*, ossia $f(x) = f(-x)$, e viene chiamata *densità co-spettrale* (*cross-spectrum*), mentre la parte immaginaria $q(\omega)$ è una funzione *dispari*, ossia $f(x) = -f(-x)$, e viene detta *spettro di quadratura*[16]. La prima misura la correlazione delle componenti di frequenza in fase delle due serie, mentre la seconda corrisponde alla correlazione delle componenti sfasate.

L'analisi co-spettrale fornisce altre informazioni sulle relazioni tra le serie nel dominio frequentistico:

- innanzitutto la *coerenza*, una forma normalizzata del cross-spettro che misura la corrispondenza globale tra gli spettri delle due serie poste in relazione (in pratica la "correlazione tra spettri"):

$$K_{xy}(\Omega) = \frac{S_{xy}(\Omega)}{\sqrt{S_x(\Omega) \cdot S_y(\Omega)}} \ .$$

  Il valore $K_{xy}(\omega) = 1$ significa che la componente di frequenza ω è molto simile in entrambi i segnali (con perfetta relazione lineare se ciò avviene per ogni ω), mentre un valore nullo $K_{xy}(\omega) = 0$ significa che non esiste alcuna somiglianza e, se ciò avviene per ogni ω le serie sono del tutto incorrelati e quindi sarà $r_{xy}(h) = 0$ per ogni h[17]. A volte viene calcolato il quadrato di tale rapporto, ossia la *coerenza quadrata*, che può essere interpretato in modo si-

[16] Se le due serie sono incorrelate tra loro, ossia $r_{xy}(h) = 0$ per ogni h, allora $c(\omega) = q(\omega) = 0$; se invece $Y_t = X_t$, allora $S_{xy}(\omega) = S_x(\omega)$ e dunque si ricade nel caso univariato, essendo $c(\omega) = S_x(\omega)$ e $q(\omega) = 0$.

[17] Ciò implica che si può utilizzare questo risultato dell'analisi co-spettrale (robusta perché quasi non parametrica) come informazione confermativa di una correlazione incrociata poco affidabile per carenza di cointegrazione o per scarsa numerosità delle serie. Peraltro, dato che $K_{xy}(\omega) = K_{yx}(\omega)$, è generalmente preferibile omettere gli indici, e dunque scrivere semplicemente $K(\omega)$.

mile al noto *indice di determinazione* (quadrato del coefficiente di correlazione lineare). Coerenza e coerenza quadrata sono logicamente insensibili alle trasformazioni lineari dei processi $\{X_t\}$ e $\{Y_t\}$, di conseguenza le informazioni che esse forniscono valgono non solo per le serie stazionarie che li descrivono, ma, più generalmente, anche per le serie (non stazionarie) di cui le prime costituiscono una trasformazione lineare.

Tuttavia, è sconsigliabile interpretare i valori di coerenza in modo autonomo; infatti, per esempio, quando le stime di densità spettrale in entrambe le serie sono molto piccole, possono generare grandi valori di coerenza (il divisore nel calcolo dei valori di coerenza sarà molto piccolo), anche se non esistono forti componenti ciclici in entrambe le serie nelle rispettive frequenze.

- In seconda istanza, l'*ampiezza incrociata* (detta anche "ampiezza di fase")

$$A_{xy}(\omega) = \sqrt{c(\omega)^2 + {}_i q(\omega)^2} \; ;$$

l'ampiezza di fase misura quanto ciascuna componente di frequenza di una determinata serie viene influenzata dalle componenti dell'altra serie.

- Una misura molto interessante è il *guadagno spettrale* delle serie, dato dal valore di ampiezza incrociata rapportato alla densità spettrale stimata per una delle due serie nell'analisi. Di conseguenza, vanno calcolati *due* distinti valori di guadagno $A_{xy}(\omega)/S_x(\omega)$ e $A_{xy}(\omega)/S_y(\omega)$, che possono essere interpretati come coefficienti di regressione OLS delle rispettive frequenze delle serie, potendo dunque fungere da *conferma* (nel dominio frequentistico) dei coefficienti di correlazione incrociata nel dominio temporale[18].

[18] Infatti, se la funzione di guadagno di una serie, considerata conseguente all'altra, assume valori rilevanti mentre il guadagno della seconda dalla prima risulta irrilevante in senso assoluto o anche relativo (ad esempio, con valori mediamente inferiori alla metà dei primi), è giustificato ritenere che esista una relazione unidirezionale, la quale nel dominio temporale può essere interpretata come *dipendenza*; se entrambe le funzioni presentano valori poco rilevanti, logicamente, se ne dovrebbe ammettere l'indipendenza nel dominio temporale, mentre se entrambe sono cospicue è presumibile l'esistenza di una relazione mutua tra le serie: una presumibile *correlazione* nel dominio temporale, che però in pratica risulta di difficile interpretazione. Invero, semplificando al massimo, con una serie di 100 osservazioni potremmo rilevare nel dominio frequentistico un guadagno elevato della serie X dalla Y per la frequenza 0,333 (corrispondente a un periodo pari a 3 unità di tempo), e un guadagno altrettanto elevato della serie Y dalla X per la frequenza 0,25 (periodo = 4). Trasponendo queste relazioni nel dominio temporale, si dovrebbe ipotizzare che la serie Y sarebbe influenzata dalla serie X di 4 periodi temporali precedenti, la quale però sarebbe a sua volta influenzata dalla Y di 3 periodi ancora precedenti e così via; dunque si avrebbe una autocorrelazione spuria della Y da se stessa, con lag dipendenti dallo sfasamento fra le serie. La stessa cosa, naturalmente, avverrebbe per quanto riguarda la serie X, seppure con ordine di sfasamento invertito; dunque, ogni 12 periodi di tempo si avrebbe una relazione di autocorrelazione circolare per ambo le serie.

- Infine, le stime di *sfasamento* (o *spettro di fase*), normalmente indicate dalla lettera greca φ, sono calcolate come arcotangente del rapporto tra lo spettro di quadratura stimato e la densità co-spettrale: $\varphi(\omega) = \arctan[q(\omega)/c(\omega)]$.[19] Lo spostamento di fase misura quanto ciascuna componente di frequenza di una serie precede o segue quelle dell'altra serie.

Come conseguenza delle considerazioni esposte in questo paragrafo, sembra lecito affermare che i risultati dell'analisi co-spettrale, ed in particolare i valori di *guadagno* e di *sfasamento* delle serie studiate, oltre al loro valore intrinseco in termini di analisi nel dominio frequentistico, siano ottimi complementi all'analisi delle correlazioni incrociate (la quale fornisce, comunque, risultati chiari e immediatamente identificabili nel dominio temporale), non soffrendo di alcuni dei vincoli metodologici di queste ultime, quali le condizioni necessarie di *cointegrazione* e di un *consistente numero di termini* nelle serie di dati.

Inoltre, così come nel correlogramma alcuni termini ritardati possono assumere significatività statistica pur essendo risultato di "autoregressione spuria", anche tra le correlazioni incrociate potrebbero esservi alcune componenti che risultano significative solo per l'interazione di componenti con ritardi minori (*battimento*), e che non risulta semplice identificare in assenza di informazioni sulla correlazione incrociata parziale.

Tale serie di constatazioni si estrinseca in una procedura di analisi che può essere applicata a serie temporali di varia natura, sia economiche che sociali, finanziarie o fisico-tecniche: come esempio di applicazione di detta procedura a serie finanziarie, si veda il paragrafo seguente.

3. Applicazione della tecnica proposta al caso di serie finanziarie

Le serie finanziarie qui considerate, derivanti da una precedente ricerca svolta in collaborazione con una società di consulenza finanziaria, dove l'obiettivo preposto enfatizzava la costruzione di un portafoglio di investimento strategicamente allocato con 4 o 5 *asset class*, ripercorrono l'arco temporale compreso tra gennaio 2008 e settembre 2014.

I dati disponibili riguardano le quotazioni di apertura, massimo e minimo e quelle di chiusura giornaliera, entro ciascuna settimana lavorativa (5gg), dei seguenti asset:

[19] Se $c(\omega) = 0$, però, si pone $\varphi(\omega) = \pm\pi/2$ a seconda che $q(\omega)$ sia positivo o negativo; se invece $q(\omega) = 0$, si pone $\varphi(\omega) = 0$ ove $c(\omega) > 0$ e $\varphi(\omega) = \pi$ altrimenti.

- *Gold Future*[20], strumento privilegiato dell'investitore "tradizionale", pur se ultimamente ha fatto registrare periodi di grande interesse al rialzo alternati a periodi di crisi ribassista, anche lunghi;
- *Copper Spot*, riguardante le quotazioni del rame, bene di *consumo* e non di *investimento* come il precedente, e per questo molto legato all'andamento dell'economia reale e del settore industriale;
- *Thomson-Reuters-Jefferies CRB*, indice delle materie prime che ha sostituito il precedente CRBMET (sempre meno rappresentativo, e dunque corretto, dal 2010 in poi),
- *U.S. 10-yrs Treasure Note*, asset che rappresenta il rendimento del principale titolo di stato a medio-lungo termine statunitense, con scadenza decennale (molto utile per descrivere lo stato di salute dell'economia americana, e che potrebbe dunque influenzare le scelte dei grandi investitori);
- *indice Standard&Poor's 500*, principale indice del mercato azionario mondiale, pur riguardando soltanto aziende statunitensi[21];
- *indice di volatilità dell'indice S&P 500 con scadenza a 30 giorni* (rappresenta il valore medio del "premio di rischio" che gli investitori sono disposti a pagare per le opzioni sullo S&P 500, offrendo una previsione della variabilità del mercato azionario nei successivi 30 giorni);
- *indice di volatilità dell'indice S&P 500 con scadenze a 60 e 90 giorni* (offre una previsione della variabilità del mercato nei successivi 2-3 mesi).

Le serie descritte hanno una copertura molto buona nel periodo considerato, con pochissime interruzioni infrasettimanali (generalmente per festività, che nel mercato statunitense a cui esse si riferiscono sono molto meno numerose rispetto all'Italia e ad altri Paesi).

I dati mancanti, in sequenze non maggiori di 2-3 giorni, sono quindi stimati tramite interpolazione rettilinea dei valori contigui.

[20] I *futures* sull'oro sono contratti tramite i quali si effettuano compravendite di oro secondo termini decisi al momento della stipula ma che avranno validità futura in un giorno di scadenza prestabilito, in genere tre mesi dopo. Ciò significa che il compratore non paga immediatamente (o almeno non paga l'intero importo pattuito) e che il venditore non ha l'obbligo di consegnare l'oro immediatamente. Lo scambio "fisico" del bene avviene invece alla scadenza, quando il compratore paga e il venditore consegna l'oro.

[21] L'indice S&P 500 è stato realizzato dalla società di consulenza finanziaria Standard & Poor's nel 1957 sulla base dell'andamento di un paniere azionario formato dalle 500 grandi aziende statunitensi a maggiore capitalizzazione, scelte tra quelle contrattate al *New York Stock Exchange* (Nyse), all'*American Stock Exchange* (Amex) e al *National Association of Securities Dealers Automated Quotation* (Nasdaq). Il peso attribuito a ciascuna azienda è direttamente proporzionale al valore di mercato.

Figura 1. *Chiusure giornaliere di Gold Future e Rame Spot.*

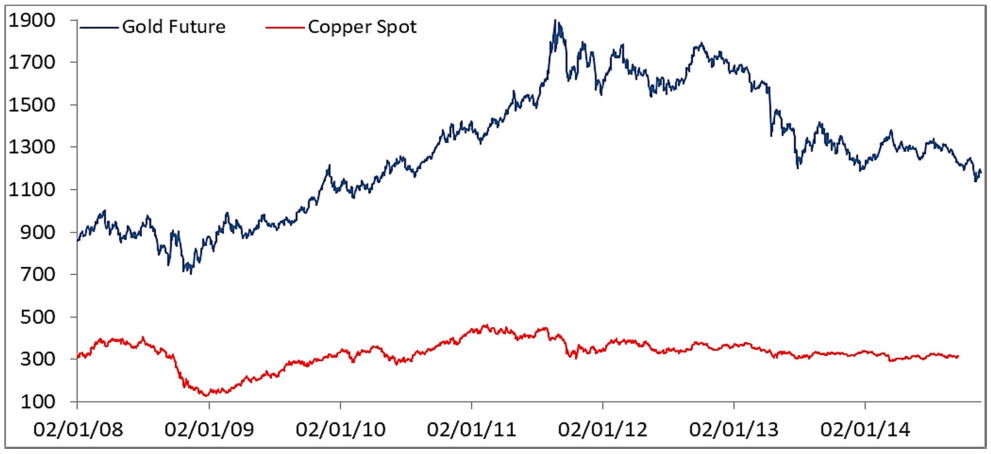


Figura 2. *Chiusure giornaliere degli indici Standard&Poor's 500 e Thomson-Reuters-Jefferies CRB.*

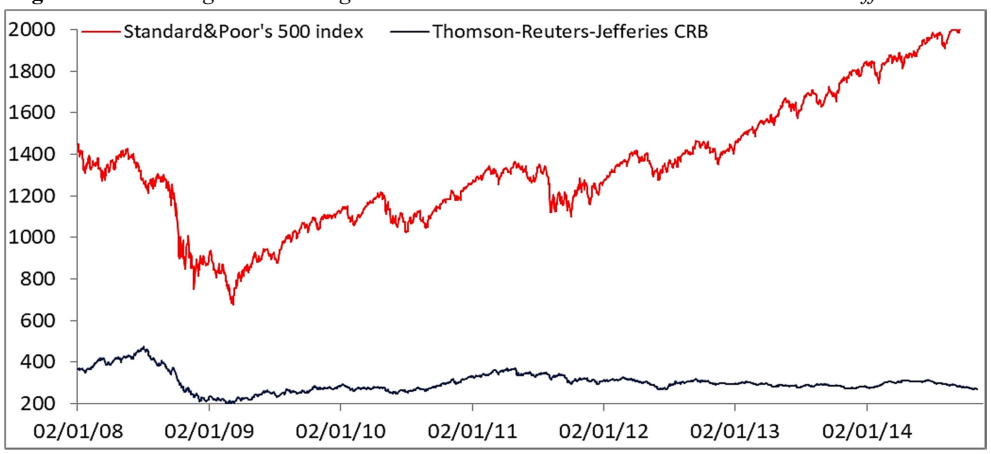
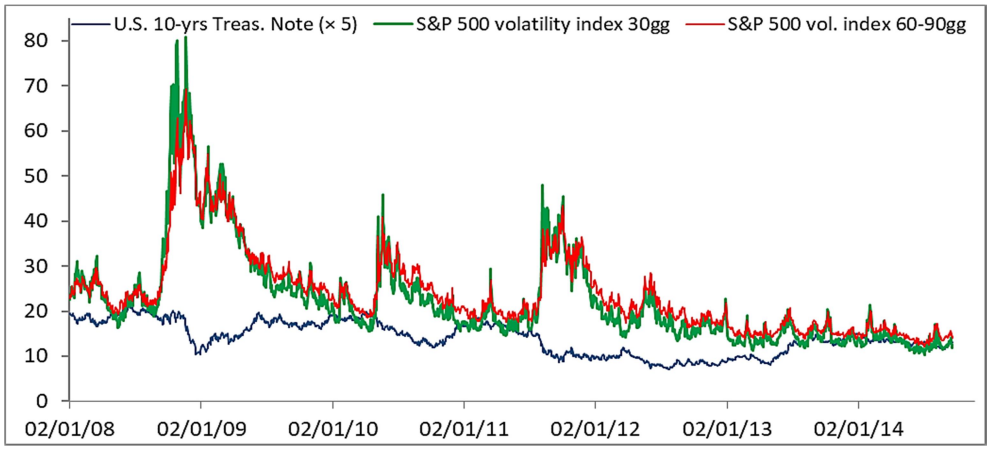


Figura 3. *Chiusure giornaliere di U.S. 10-Year Treas.Note e di S&P500 volatility index a 30gg e 60-90gg.*

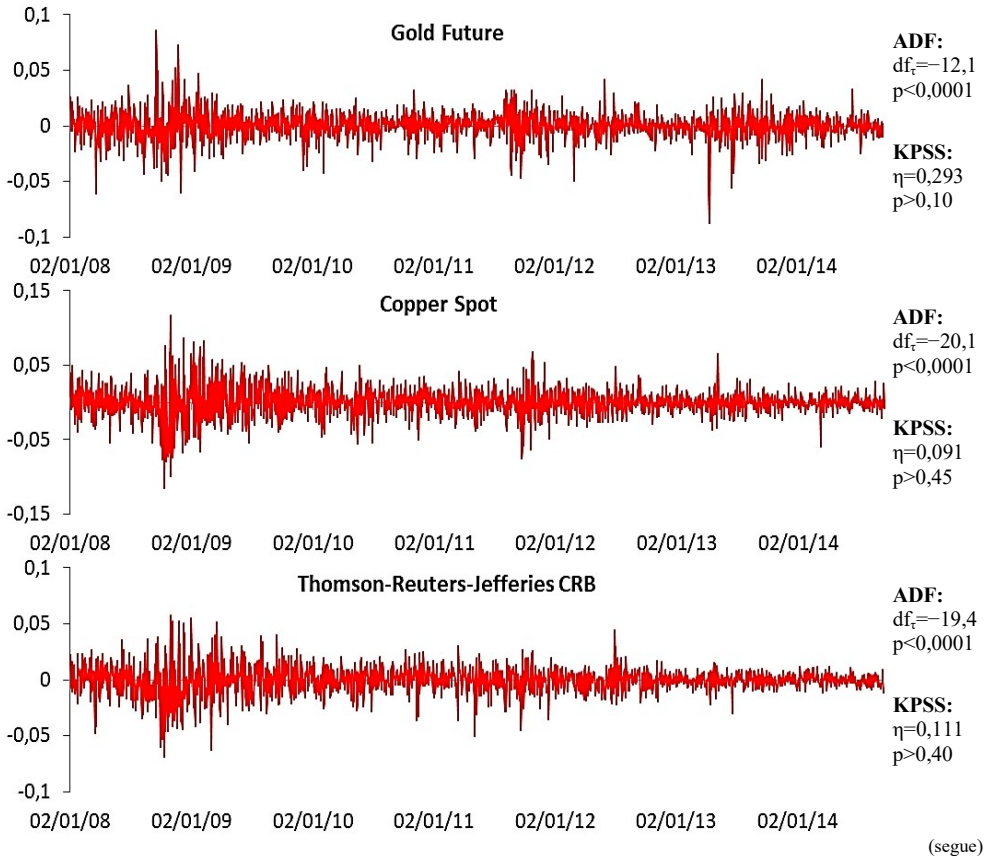


L'osservazione grafica delle serie finanziarie evidenzia varie forme di trend (a volte crescente e a volte decrescente, comprendendo vari cicli interni di diversa cadenza), ossia non stazionarietà in media, ma anche forti differenze di variabilità nel tempo, soprattutto per gli asset *Gold Future*, *Standard&Poor's 500*, *S&P 500 volatility index* (a entrambe le scadenze descritte), ossia non stazionarietà in varianza. Per ottenere serie di dati stazionari (evitando così l'insorgenza di relazioni spurie tra le serie, legate tra loro dalla relazione con il tempo) è quindi opportuno innanzitutto far ricorso a una trasformazione logaritmica dei dati (Box-Cox, 1964) per poi calcolarne le differenze prime. In definitiva, la trasformazione utilizzata è la seguente:

$$y_t = \ln(x_t) - \ln(x_{t-1}) = \ln[x_t/(x_{t-1})] \, .$$

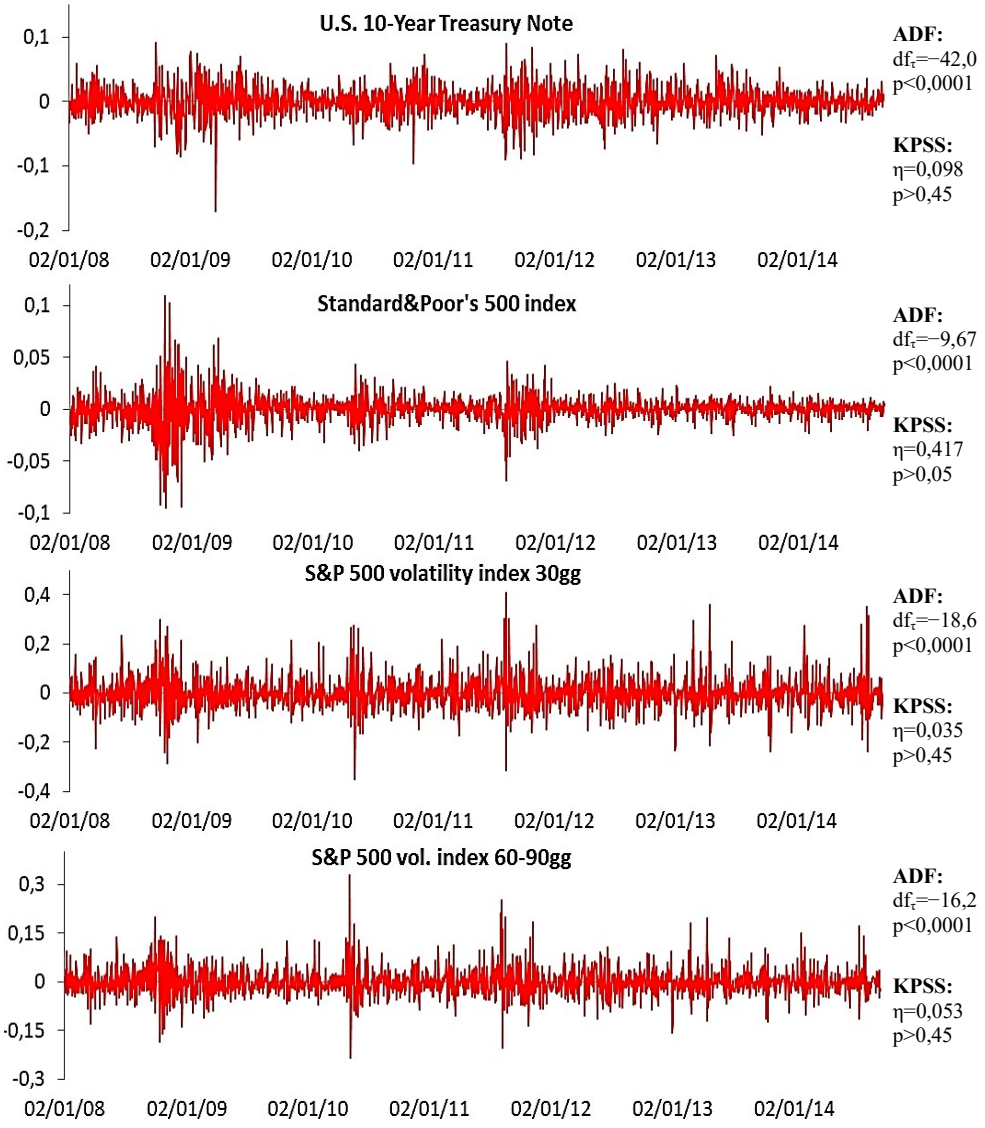
Come risultato di tale operazione, le serie trasformate risultano assimilabili a processi stazionari del 2° ordine (Fig. 4).

Figura 4. *Serie logaritmiche differenziate delle chiusure giornaliere degli asset. Significatività statistica dei test DW, ADF e KPSS per ogni asset.*



(segue)

Figura 4. (*Continuazione).*



Tale assunzione è stata comunque verificata per ogni serie sia con test di radice unitaria ADF (ponendo un livello di significatività $\alpha = 0{,}01$) e sia tramite KPSS con 16 periodi di ritardo, ponendo come valore critico $\eta_{0,025} = 0{,}574$. Questa scelta di un livello di significatività α=0,025 anziché il più critico α=0,01 è motivata dalla minor potenza del test, e quindi dal maggior rischio di accettare una ipotesi di base falsa se si fissa un livello di significatività troppo esiguo. La ridondanza delle statistiche utilizzate è invece giustificata dal fatto, già discusso nel par. 2, che la stazio-

narietà debole delle serie è l'unico vincolo al corretto utilizzo dell'analisi nel dominio frequentistico, più che in quello temporale, e che il test ADF presenta il vincolo dell'integrazione delle serie.

Dopo aver verificato la stazionarietà delle singole serie con risultati confortanti (soltanto la serie dell'indice Standard&Poor's 500 presenta un valore empirico del test KPSS che si avvicina a α=0,05, ma non abbastanza da indurre dubbi, tenendo conto della sua minor potenza), si è proceduto all'analisi della cointegrazione a coppie, utilizzando anche in questo caso ambo le tecniche disponibili in Gretl. Per le combinazioni ottenute ponendo come termine fisso la serie *Gold Future*, la quale fin dall'inizio è apparsa di primario interesse, i risultati del test di Engle-Granger e quelli del test di Johansen generalmente coincidono perfettamente, rifiutando l'ipotesi di base (integrazione dei residui della combinazione lineare il primo, presenza di autovalori non nulli l'altro) con valori $p<0,001$. Fa parziale eccezione, anche in questo caso, la relazione tra indici *Gold Future* e indici *Standard&Poor's*, per la quale il *test traccia* di Johansen fornisce $0,010 < p < 0,015$, dunque conducendo (per un livello di significatività $\alpha = 0,01$) ad accettare l'ipotesi di base che almeno una combinazione lineare delle due serie non sia stazionaria[22].

Limitando la presente analisi alla ricerca delle eventuali influenze dei vari indici considerati sull'indice relativo al *Gold Future* (inteso qui come semplice antecedenza/conseguenza dei dati, non potendo ancora ipotizzare relazioni causali), si utilizza qui innanzitutto l'analisi co-spettrale, e in particolare la disamina della funzione di guadagno, con finestra di Tukey-Hamming a 5 termini, di quest'ultimo indice rispetto a ciascuno degli altri indicatori finanziari (e viceversa, a coppie). Ove la funzione di guadagno indichi una relazione di antecedenza, in un senso o nell'altro, ad essa viene affiancata l'analisi dell'opportuna funzione di correlazione incrociata, con lag fino a 20 giorni[23].

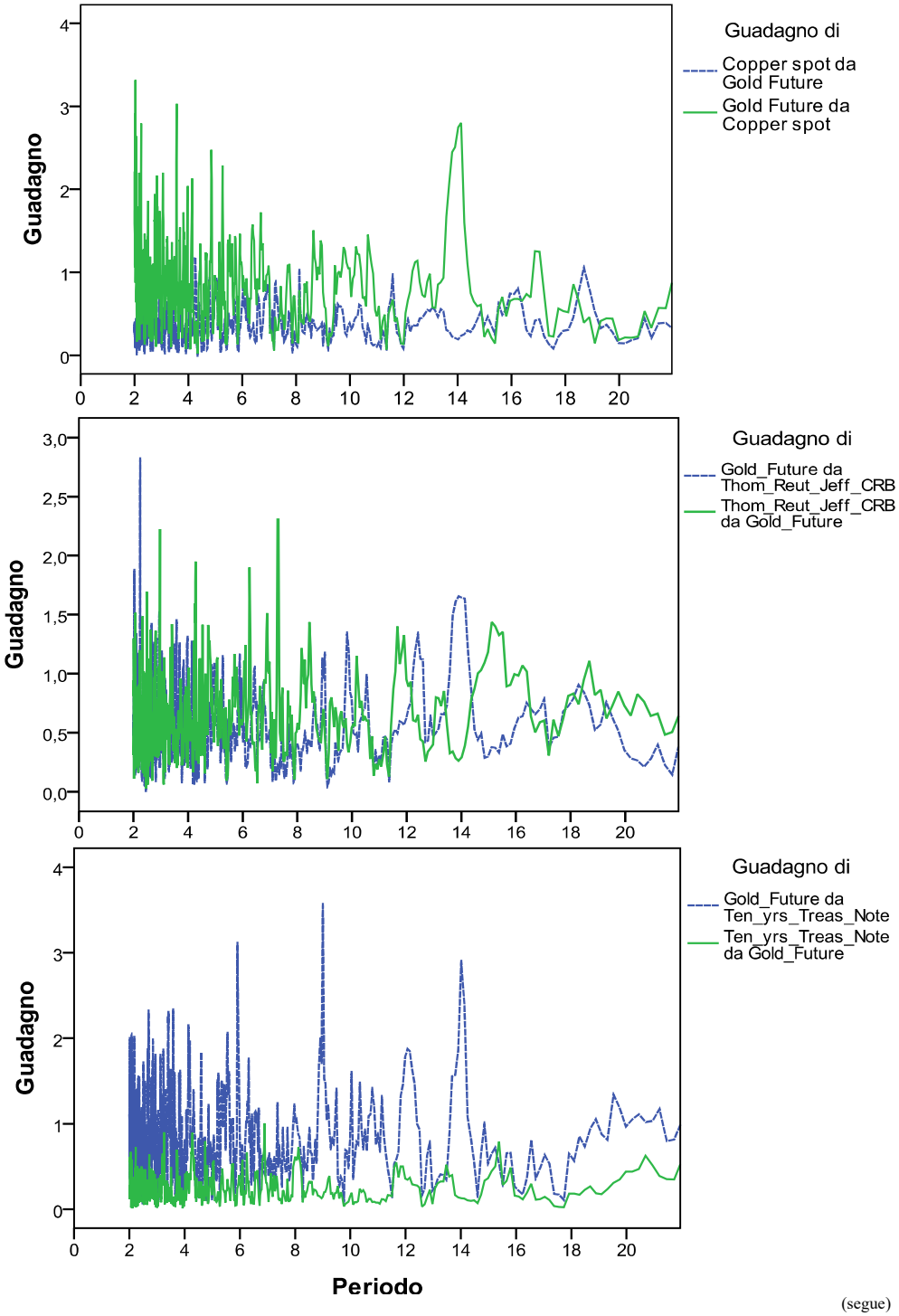
La Fig. 5 mostra dunque le funzioni di guadagno tratte dall'analisi co-spettrale (bivariata) tra l'indice Gold Future e ciascuno degli altri indici finanziari descritti in precedenza.

---

[22] Anche altre combinazioni di serie hanno presentato risultati interessanti, ma in questa sede è opportuno, per maggior chiarezza di trattazione, limitare la discussione alla serie suddetta.

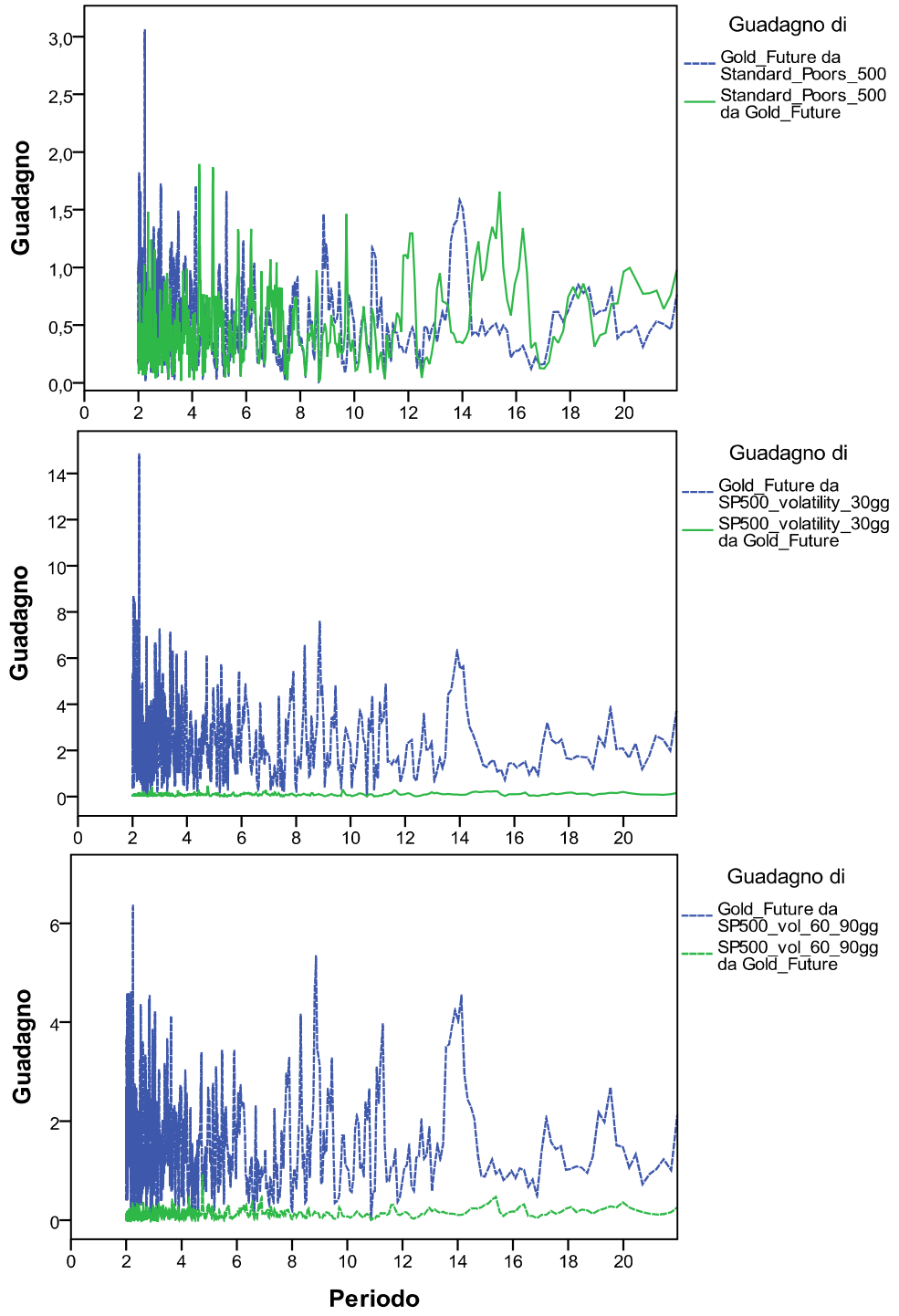
[23] Un ritardo di 20 giorni corrisponde, per le serie finanziarie rilevate su "settimane lavorative" di 5 giorni, a quattro settimane piene, sufficienti a rilevare la maggior parte delle regolarità di breve periodo senza eccessivo carico di informazioni, che potrebbe andare a detrimento dell'interpretazione dei risultati. Per gli stessi motivi di chiarezza dei risultati, se la relazione identificata appare unidirezionale, nei grafici di correlazione incrociata viene riportato solo il ramo positivo; inoltre, i grafici delle funzioni di guadagno sono espressi in termini di periodo invece che di frequenza, e riportano solamente i primi 20 periodi (a partire dal periodo minimo valutabile, ossia 2 unità temporali).

Figura 5. *Funzioni di guadagno co-spettrale tra Gold Future e gli altri indici finanziari.*



(segue)

Figura 5. *(Continuazione).*



Come si desume dai grafici che compongono la Fig. 5, le funzioni di guadagno co-spettrale tra *Gold future* e gli altri indici, evidenziano la funzione di guadagno che determinano le relazioni di antecedenza e conseguenza tra gli indici a confronto.

Nella prima immagine viene esplicata la funzione di guadagno co-spettrale considerando *Copper spot* "dipendente" da *Gold future* (andamento contrassegnato dalla linea blu) e, alternativamente, *Gold future* "dipendente" da *Copper spot* (andamento contrassegnato dalla linea verde). In questo caso il guadagno medio maggiore si ha quando *Copper spot* è antecedente a *Gold future*. Infatti quasi tutte le fasi osservate nel grafico sono più basse se si considera conseguente *Copper spot*, quindi il guadagno della serie congiunta è più basso in quest'ultimo caso.

Altre significanze in termini di guadagno, in questa parte di analisi esplorativa, sono inerenti alla funzione di guadagno co-spettrale considerando la relazione tra *Gold future* e *U.S. 10-years Treasure Note* e le due relazioni di *Gold future* con l'*indice S&P 500 volatility 30 gg* e con l'*indice S&P 500 volatility 60 e 90 gg*.

Invece risultano non rilevanti le funzioni di guadagno generate dalla relazione tra *Gold Future* e *S&P500* e tra *Gold Future* e *Thomson-Reuters-Jefferies CRB*.

A questo punto è opportuno prendere in considerazione la funzione di correlazione incrociata relative alle relazioni chiaramente identificate tramite la funzione di guadagno per definire i *lag* relazionali (ovviamente, nei casi di non rilevanza della funzione di guadagno, la funzione della *cross correlation* non identifica adeguatamente le relazioni intercorrenti tra le variabili oggetto di studio e dunque non è molto utile).

Nei grafici della Fig. 6, il significato di una relazione con lag *k-esimo*, laddove risulti statisticamente significativa, è quello di evidenziare un'influenza della serie ove essa appare sui valori del *Gold future* di *k* giorni dopo. Tale influenza può risultare di relazione diretta o inversa a seconda che il *lag* sia di segno positivo o negativo.

Ad esempio, la *cross-correlation* tra *Copper spot* e *Gold Future* evidenzia che le quotazioni di chiusura del primo hanno innanzitutto una cospicua influenza diretta il giorno stesso (al *lag* 0, relazione che potrebbe comunque essere spuria, ossia entrambi gli *asset* possono essere influenzati da altri), ma anche una lieve ma significativa influenza *negativa* con *lag* 8 e 17. Si rilevano anche influenze *borderline* (positive, ma quasi irrilevanti) con ritardo 6, 9 e 15.

L'influenza dell'asset dei Buoni del Tesoro U.S.A. a 10 anni sul *Gold Future* è fondamentalmente negativa, pur se non molto rilevante, sia al *lag* 0 (il giorno stesso) che, ancor meno fortemente, ai *lag* 5, 8 e 19. Vi è anche una influenza positiva di minima significanza al *lag* 2 (ossia 2 giorni dopo).

Figura 6. *Funzioni di cross-correlation tra Gold Future e gli indici finanziari da cui consegue.*

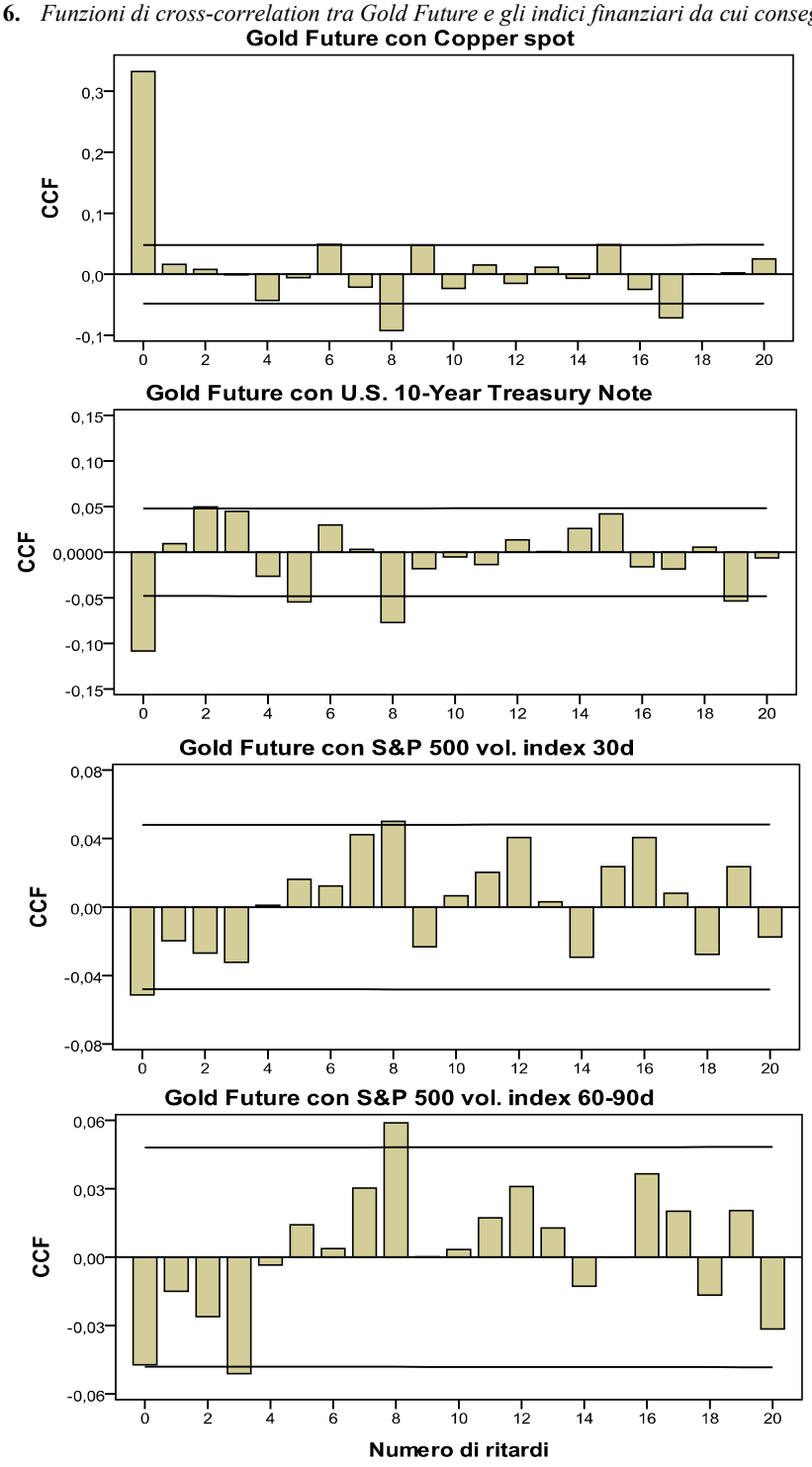
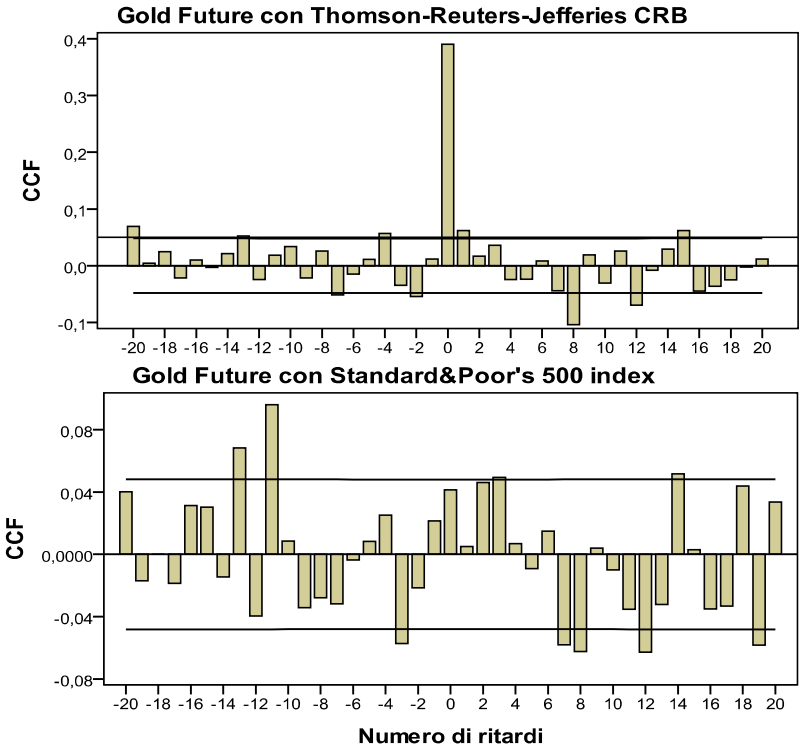


Figura 7. *Funzioni di cross-correlation tra Gold Future e gli indici finanziari da cui ha relazioni di antecedenza/conseguenza.*



La dipendenza del *Gold future* è minima, quasi insensibile, nell'incrocio con gli indici *S&P500 a 30 gg* e *a 60-90 gg*, ma si riconosce una influenza negativa borderline con *lag* 0 (ma nel secondo caso anche con *lag* 2), e una positiva con *lag* 8.

Come si è affermato, ove la funzione di guadagno non consente di stabilire se vi sia una serie antecedente e una conseguente, come nel caso degli indici *Thomson-Reuters-Jefferies CRB* e *S&P500*, sono altrettanto difficili da interpretare chiaramente eventuali relazioni significative: nel primo grafico della Fig. 7, ad esempio, constatiamo (senza considerare la correlazione al *lag* 0) che oltre ad alcune relazioni sul lato destro dei correlogrammi incrociati (ossia una sensibile influenza negativa del *TRJ-CRB* sul *Gold future* con *lag* 8 e con *lag* 12, e una lieve influenza positiva con *lag* 1 e *lag* 15), anche relazioni significative sul lato sinistro, ossia il TRJ-CRB viene influenzato, seppur in misura minimale, dal Gold future: negativamente con *lag* −2 e −7 (ossia a un incremento del secondo asset fa seguito abbastanza spesso un decremento del primo con 2 e 7 giorni di ritardo) e positivamente con *lag* −4, −13 e −20. Ancor più confusa si rivela la relazione tra *Gold Future* e *S&P500*, che non a caso è la coppia di asset più problematica nel test di Johansen.

4. Considerazioni conclusive

Il lavoro svolto ha permesso di esplorare le diverse relazioni che coesistono tra variabili finanziarie molto rilevanti nel mercato statunitense, identificando alcuni elementi di un certo interesse. Per prima cosa, il *Gold Future* risulta più spesso un asset influenzato (e non uno influenzante) dalle variazioni di altri indici del mercato finanziario considerati *secondari*, anche dopo aver reso stazionarie le serie.

In particolare, poi, una relazione che si presenta costantemente significativa, benché di lieve entità, è quella tra i valori di chiusura dei diversi indici e la quotazione di chiusura Gold Future 8 giorni (lavorativi) dopo.

Riconoscimenti

Gli Autori ringraziano la collega Caterina Marini per le sue osservazioni sulla stazionarietà delle serie temporali, nonché la società di consulenza finanziaria *Pragma 11 Sas* per la concessione dei dati esemplificativi qui utilizzati.

Riferimenti bibliografici

Battaglia, F. (2007). *Metodi di previsione statistica*, Springer-Verlag, Milano.

Bendat, J. S.; Piersol, A. G. (1966). *Measurement and Analysis of Random Data*, J. Wiley, New York.

Bendat, J. S.; Piersol, A. G. (2004) *Random data analysis and measurement procedures*, Wiley Series in Probability and Statistics (3rd Edition).

Box, G. E. P.; Cox, C. (1964). An Analysis of Transformations, *Journal of the American Statistical Association*, 65.

Cottrell, A.; Lucchetti, R. (2017). *Gretl User's Guide*, gretl documentation. URL http://sourceforge.net/projects/gretl/files/manual/gretl-guide-a4.pdf/download

Delvecchio, F. (1974). Modificazioni strutturali della curva dei matrimoni in Italia, *Giornale degli economisti e Annali di economia*. Anno XXXVI (nuova serie), n. 3-4 (marzo-aprile).

Dickey, D. A.; Fuller, W. A. (1979). Distribution of the Estimators for Autoregressive Time Series with a Unit Root. *Journal of the American Statistical Association*. 74 (366): 427–431. (doi:10.2307/2286348).

Engle, R. F.; Granger, C. W. J. (1987). Co-integration and error correction: Representation, estimation, and testing, *Econometrica*, 55: 251–276.

Granger, C. W. J.; Newbold, P. (1974). Spurious regressions in econometrics. *Journal of Econometrics*. **2** (2): 111–120. doi:10.1016/0304-4076(74)90034-7.

Hamilton, J. D. (1994). *Time Series Analysis*, Princeton University Press, Princeton, NJ.

Hannan, E. J. (1960). *Time Series Analysis*, Methuen, London.

Harris, R. (1995). *Using Cointegration Analysis in Econometric Modelling*. Prentice-Hall.

IBM Corporation (2012). *IBM SPSS Forecasting 21*. URL ftp://public.dhe.ibm.com/software/analytics/spss/documentation/statistics/21.0/en/client/Manuals/IBM_SPSS_Forecasting.pdf

Kendall, M. G. (1973). *Time-Series*, Charles Griffin & Co. Ltd., London & High Wycombe.

Kendall, M. G.; Stuart, A. (1976). *The Advanced Theory of Statistics - Volume 3 - Design and Analysis, and Time-Series*, Charles Griffin & Co. Ltd., London & High Wycombe.

Kwiatkowski, D.; Phillips, P. C. B.; Schmidt, P.; Shin, Y. (1992). Testing the null of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root? *Journal of Econometrics*, 54: 159–178.

Johansen, S. (1995). *Likelihood-Based Inference in Cointegrated Vector Autoregressive Models*, Oxford University Press, Oxford.

Johansen, S. (2000). Modelling of cointegration in the vector autoregressive model. *Economic Modelling*, 17: 359–373.

Malinvaud, E. (1971). *Metodi statistici dell'econometria*, UTET, Torino.

Oppenheim, A.; Verghese, G. C. (2010). *Signals, Systems, and Inference*, MIT OpenCourseWare. URL https://ocw.mit.edu/courses/electrical-engineering-and-computer-science/6-011-introduction-to-communication-control-and-signal-processing-spring-2010/readings/MIT6_011S10_notes.pdf

Said, S. E.; Dickey, D. A. (1984). Testing for Unit Roots in Autoregressive-Moving Average Models of Unknown Order. *Biometrika*. 71 (3): 599–607. (doi:10.1093/biomet/71.3.599).

Stuart, J. S. (1961). *Fourier Analysis*, Methuen & Co, London.

Trantner, C. T. (1951). *Integral Transforms in Mathematical Physics*. Methuen London.

Vajani, L. (1980). *Analisi statistica delle serie storiche*, vol. I, CLEUP, Padova.

Wold, H. (1954). *A study in the analysis of stationary time-series*, Almqvist & Wiksell, Stockolm.

Yaglom, A. M. (1958). Correlation Theory of processes with random stationary n^{th} increments, *Transactions of the American Mathematical Society*, Series 2, vol. 8: 87-141. Original: (1955), Корреляционная теория процессов со случайными стационарными *n*-ми приращениями, Математический сборник, 37(79), 1: 141-196).

https://sourceforge.net/projects/gretl/files/manual/gretl-guide-a4.pdf/download

UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO
Metodi e analisi statistiche 2017
DIPARTIMENTO DI ECONOMIA E FINANZA
ISBN: 978-88-6629-013-1
pp. 127-149

Database del mercato del lavoro a confronto: possibile integrazione per una analisi dinamica dell'occupazione

Caterina Marini, Vittorio Nicolardi*

Università degli Studi di Bari Aldo Moro

Riassunto: La crisi economica mondiale intervenuta nel 2007 ha causato una netta recessione nella quasi totalità delle economie occidentali e determinato un preoccupante aumento della disoccupazione in tutte le fasce di età, principalmente giovanili. Gli interventi di politica economica, accompagnati da altrettante riforme del mercato del lavoro messe in atto nella maggior parte dei Paesi dell'UE, hanno avuto conseguenze in alcuni casi positive, in altri discutibili, nei singoli Stati. Per un Paese come l'Italia l'analisi degli effetti sull'occupazione, però, non è facile affrontarla alla luce soprattutto di un mercato del lavoro che è stato totalmente stravolto sotto molti punti di vista. La necessità di integrare fonti di dati relative al mercato del lavoro provenienti da diversi database sembra rappresentare l'unica strada percorribile per riuscire a fare un po' di chiarezza sulla situazione attuale. Nel presente lavoro si analizzano separatamente i contenuti e la struttura dei database di ISTAT, Ministero del Lavoro e delle Politiche Sociali, e INPS: il primo riveniente dalla Rilevazione Continua delle Forze di Lavoro, il secondo costruito per scopi amministrativi attraverso le comunicazioni obbligatorie riguardanti tutti gli aspetti delle assunzioni (attivazioni, trasformazioni, proroghe e cessioni), il terzo definito dalle posizioni contributive e retributive di tutti i lavoratori. Dall'analisi sviluppata, si conclude che non solo l'integrazione è possibile, ma anche fortemente auspicabile per ottenere la completezza di un dato che, mai come ora, risulta essere altamente importante per l'evoluzione del Paese.

Keywords: Mercato del lavoro, RCFL, Comunicazioni Obbligatorie, Dati Longitudinali, Matrici di Transizione.

* Autore corrispondente: vittorio.nicolardi@uniba.it

1. Introduzione

La crisi economica mondiale intervenuta a partire dalla crisi dei subprime, esplosa negli Stati Uniti d'America nell'agosto del 2007, ha marcato una netta recessione in quasi tutti i Paesi dell'Unione Europea, oltre che del resto delle economie occidentali. Parallelamente, anche il mercato del lavoro ha conosciuto una crisi altrettanto drammatica, con un aumento della disoccupazione sia a livello generale, sia, in particolare, in riferimento alle classi giovanili di età.

I governi dei Paesi dell'Unione Europea hanno cercato di tamponare questa drammatica situazione occupazionale con interventi di varia natura, incentrati da un lato al sostegno dell'occupazione e dall'altro a favorire l'ingresso di nuovi lavoratori, in particolare giovani, nel mercato del lavoro.

L'analisi degli effetti di questi interventi di politica economica non è semplice, sia perché qualsiasi intervento di tipo legislativo mirato al mercato del lavoro non può prescindere da paralleli interventi finalizzati alla ripresa economica generale, sia perché tali interventi sono normalmente di durata pluriennale e spesso si sovrappongono nei loro effetti. Ciò è particolarmente vero per le fasce più giovani di età, a favore delle quali pesano sia gli interventi generali di lotta alla disoccupazione, sia interventi specificatamente orientati all'assunzione di persone nelle suddette classi.

L'analisi degli effetti degli interventi di politica economica sul mercato del lavoro necessita, quindi, di informazioni a livello disaggregato che permettano una analisi dinamica dell'andamento della disoccupazione al fine di verificare quali siano i flussi effettivi di passaggio tra le condizioni di disoccupazione o inattività e la condizione di occupato. Allo stesso tempo, altrettanto importante è l'analisi dinamica dei flussi di passaggio tra le modalità di occupazione a tempo determinato e a tempo indeterminato.

Allo stato attuale, però, la disponibilità di strutture di dati dinamiche relative al mercato del lavoro non sembra sufficiente per le finalità prepostesi. Nella maggior parte dei casi, tale indisponibilità è legata ai problemi di tutela della privacy che impedisce di identificare il singolo individuo per seguirne i cambi di stato relativamente alla sua situazione lavorativa nel tempo. A ciò si aggiunge anche il fatto che le diverse banche dati disponibili non contemplano tutte le variabili necessarie ad un'analisi strutturale del mercato del lavoro, essendo ognuna costruita in funzione degli scopi istituzionali costitutivi degli Enti che le gestiscono.

Pertanto, la necessità di integrare fonti di dati provenienti da diversi database sembra rappresentare l'unica strada, per quanto ponga una serie di problemi anche di tutela della privacy in materia statistica. Infatti, principalmente nell'ambito del mercato

del lavoro, l'integrazione dei database di diversa fonte non può che avvenire attraverso l'uso di dati personali e di dati identificativi, quale, in primis, il codice fiscale.

In senso generale, la tutela dei dati personali viene regolata dal D.Lgs. 196/2003 (Codice della privacy) modificato, in seguito, dal D.Lgs. 201/2011, per ciò che riguarda le persone fisiche, mentre per quanto riguarda i dati relativi alle persone giuridiche si fa riferimento alle norme in vigore riguardo il segreto statistico. Quest'ultimo è normato a livello comunitario dalla Raccomandazione COM(2005) 217 (Commission of the European Communities, 2005), la quale al Principio *5 - Riservatezza statistica* recita che "Deve essere assolutamente garantita la tutela della privacy dei fornitori di dati (famiglie, imprese, amministrazioni e altri rispondenti), così come la riservatezza delle informazioni da essi fornite e l'impiego di queste a fini esclusivamente statistici". Nel successivo Regolamento statistico (UE) 2015/759 (Unione Europea, 2005) viene sottolineato come "La riservatezza dei dati ottenuti a partire da dati amministrativi dovrebbe essere tutelata secondo i principi e gli orientamenti comuni applicabili a tutti i dati riservati utilizzati per la produzione di statistiche europee. È opportuno inoltre redigere e pubblicare quadri di valutazione della qualità applicabili a tali dati, nonché principi di trasparenza".

È interessante rilevare come le norme codificate nel D.Lgs. 196/2003 distinguono il caso in cui il titolare del trattamento dei dati sensibili sia un soggetto pubblico oppure un soggetto privato. Inoltre, in relazione al problema della privacy riguardante, in particolare, i dati in possesso dei soggetti pubblici da utilizzare per fini statistici, occorre distinguere tra diverse casistiche: dati amministrativi già disponibili presso l'Ente pubblico, dati amministrativi provenienti da altri soggetti e basi statistiche prodotti da soggetti SISTAN[1] e derivanti dall'integrazione di fonti amministrative diversificate. Nel caso in cui i dati siano già disponibili presso l'Ente, occorre ulteriormente distinguere tra dati di cui l'Ente stesso è titolare, dai dati di cui l'Ente dispone indirettamente in funzione delle proprie competenze istituzionali.

Nello specifico del trattamento dei dati amministrativi a fini statistici, il D.Lgs. 196/2003, all'art. 99, dispone che il soggetto pubblico che ha raccolto, acquisito e trattato dati personali per qualsiasi scopo, è sempre legittimato ad utilizzarli anche per fini statistici, storici o scientifici, in quanto tali finalità sono sempre considerate

[1] Il SISTAN (Sistema Statistico Nazionale) costituisce una rete di soggetti pubblici e privati che fornisce informazioni statistiche ufficiali al Paese e agli organismi internazionali. Il SISTAN è stato istituito con il D.Lgs. 322/1989 con l'intenzione di consentire una gestione più efficace dell'attività statistica nazionale attraverso il coordinamento tra diversi enti di produzione dell'informazione statistica sul territorio nazionale. In seguito all'emanazione del Regolamento Europeo 223/2009, il SISTAN è operativamente all'interno del Sistema Statistico Europeo (SSE).

compatibili con i diversi scopi per i quali i dati sono stati precedentemente raccolti o trattati. L'importanza dell'elaborazione per fini statistici anche dei dati personali viene evidenziata dalla norma che permette di utilizzare tali dati anche se l'Ente non ha preventivamente informato i singoli individui del possibile trattamento statistico delle informazioni da essi forniti, a condizione, però, che l'attività a cui viene finalizzato il trattamento dati sia inclusa nel Programma Statistico Nazionale[2]. Se quest'ultima eventualità non si verifica, il soggetto pubblico può elaborare i dati a fini statistici dando idonea pubblicità all'utilizzazione dei dati personali che si sta effettuando, sentito, comunque, preventivamente il Garante della privacy per verificare l'idoneità della procedura di pubblicizzazione. Resta, però, la necessità per i dati personali sensibili giudiziari di indicare l'eventuale normativa che stabilisca l'obbligo di fornitura dei tali tipologie di dati o, in alternativa, la normativa che permetta di non richiedere il consenso del soggetto interessato.

Di conseguenza, è evidente che le difficoltà ai fini di un'integrazione relativa al caso di specie trattato in questo lavoro è ancora lontana dalla risoluzione, per quanto l'ISTAT operi già da tempo in merito.

Il presente lavoro analizza da un lato le fonti statistiche ufficiali relative al mercato del lavoro dell'ISTAT, e dall'altro la struttura dei database amministrativi rivenienti da documenti obbligatori che ogni datore di lavoro sia pubblico, sia privato è tenuto a presentare all'Istituto Nazionale di Previdenza Sociale (INPS, d'ora innanzi) per la tracciabilità dei flussi contributivi e retributivi, e dal 2008 al Ministero del Lavoro e delle Politiche Sociali per il tramite dei centri per l'impiego. Potenzialmente, entrambe le banche dati amministrative potrebbe rappresentare una completa e precisa fonte di dati, anche statistica, non solo per l'analisi del lavoro alle dipendenze, ma anche delle altre forme di occupazione previste dalla normativa vigente.

Il lavoro è organizzato come segue: nella Sezione 2 si analizza la principale fonte di dati ISTAT in relazione al mercato del lavoro definita dalla Rilevazione Continua delle Forze di Lavoro e la recente pubblicazione di microdati in struttura longitudinale; nella Sezione 3 si analizza dal punto di vista normativo e statistico il database in possesso del Ministero del Lavoro e delle Politiche Sociali, ottenuto per mero scopo amministrativo attraverso le comunicazioni obbligatorie che riguardano tutti gli aspetti delle assunzioni (attivazioni, trasformazioni, proroghe e cessioni), e il database in possesso dell'INPS; nella Sezione 4 si descrivono le conclusioni raggiunte dal lavoro.

[2] Il Programma Statistico Nazionale, istituito in base all'art. 13 del D.Lgs. 322/1989 individua tutte le elevazioni statistiche ritenute di interesse pubblico rientranti nelle attività del SISTAN e ne stabilisce i relativi obiettivi formativi.

2. Le fonti ISTAT sul mercato del lavoro: la Rilevazione Continua delle Forze di Lavoro

All'interno del panorama delle fonti statistiche ufficiali, la Rilevazione Continua delle Forze di Lavoro (RCFL, d'ora innanzi) rappresenta una delle fonti statistiche più importanti sul mercato del lavoro visto dal lato dell'offerta, in quanto, oltre a misurare il livello della occupazione e della disoccupazione, fornisce informazioni preziose sulla modalità e sul grado di partecipazione al mercato del lavoro dei singoli individui costituenti la popolazione. In particolare, a partire dal 2004, la rilevazione ISTAT si è arricchita di una serie di nuove informazioni molto interessanti con riguardo proprio alle problematiche affrontate nel presente studio.

L'indagine sulle Forze di Lavoro è stata in senso assoluto la prima indagine campionaria effettuata dall'ISTAT. Una prima rilevazione pilota fu effettuata nel mese di settembre 1952, mentre a partire dall'aprile 1959 l'indagine assunse carattere di sistematicità assumendo la denominazione di Rilevazione Trimestrale delle Forze di Lavoro.

Molte delle caratteristiche della rilevazione, quali la cadenza trimestrale e la rotazione del campione di famiglie, si sono mantenute sostanzialmente invariate sino al 2003, mentre dal 2004 sono state apportate sostanziali innovazioni sia dal punto di vista della metodologia di rilevazione, sia dal punto di vista della definizione e della circoscrizione del mercato del lavoro.

La principale, ma non certamente unica, motivazione alla base della ristrutturazione intervenuta nell'indagine delle Forze di Lavoro nel 2004 è stata la necessità di rispondere a quanto stabilito dall'Unione Europea in tema di organizzazione delle indagini sulle Forze di Lavoro nel Regolamento n. 77/98 del 9 marzo 1998 pubblicato nella Gazzetta Ufficiale dell'Unione Europea.

I cambiamenti intervenuti a seguito della suddetta direttiva hanno riguardato sia il disegno di rilevazione, sia la definizione del questionario, sia, infine, la strategia di campionamento, fissando una serie di criteri comuni per la rilevazione delle Forze di Lavoro nei Paesi aderenti all'UE al fine di permettere la comparabilità delle statistiche a livello europeo.

In particolare, riguardo gli aspetti più interessanti per il presente studio, il regolamento UE ha stabilito che l'indagine sulle Forze di Lavoro dovesse assumere carattere continuativo, cioè le settimane di riferimento per la rilevazione devono essere ripartite su tutto l'arco dell'anno, e che fossero introdotte nuove variabili di interesse strutturale ed economico all'interno del questionario utilizzato per la rilevazione. Vincoli molto precisi sono stati, inoltre, imposti in merito alla precisione

delle stime, le quali non devono eccedere determinati livelli di errore per i dati annuali a livello di codice NUTS 2[3].

Dal punto di vista normativo, la popolazione di interesse della RCFL è costituita dagli individui componenti le famiglie residenti in Italia, anche se temporaneamente emigrati all'estero, al netto dei membri permanenti delle convivenze (ospizi, istituti religiosi, caserme, ecc.), mentre le unità di rilevazione sono costituite dalle cosiddette famiglie di fatto[4], intese come insiemi di persone legate da vincoli di matrimonio, parentela, affinità, adozione, tutela o da vincoli affettivi, coabitanti ed aventi dimora abituale nello stesso comune, anche se non residenti allo stesso domicilio. È appena il caso di specificare che nella classe delle famiglie di fatto così come definite in precedenza, rientrano anche quelle unipersonali, cioè costituite da un solo componente.

All'interno della popolazione di interesse, viene definita popolazione in età lavorativa l'insieme degli individui avente età superiore a 14 anni, la quale viene distinta, ai fini della RCFL, in occupati, disoccupati (o persone in cerca di occupazione) e inattivi. In particolare, occorre evidenziare come nella RCFL sia stato inserito un limite superiore dell'età lavorativa pari a 74 anni, ciò principalmente per evitare ambiguità nella definizione della componente delle persone in cerca di occupazione e, di conseguenza, nel calcolo dei tassi di disoccupazione e di attività ottenuti attraverso la precedente Rilevazione Trimestrale sulle Forze di Lavoro, in cui tale limite non era previsto.

In relazione al predetto grado di partecipazione al mercato del lavoro, si distinguono le Forze di Lavoro dalle non Forze di Lavoro, includendo in queste ultime tutti gli individui che, per motivi dipendenti o indipendenti dalla propria volontà, non partecipano o non possono partecipare al mercato del lavoro, mentre nella pri-

[3] La classificazione NUTS 2 (*Nomenclature of Territorial Units for Statistics*) è stata elaborata dall'EUROSTAT al fine di definire una divisione e definizione omogenea delle unità territoriali da utilizzare per la produzione di dati statistici all'interno dell'UE. In particolare, le unità a livello NUTS 1 comprendono aree con una popolazione tra i 3 e i 7 milioni di individui, a livello NUTS 2 tra gli 800 mila e i 3 milioni, a livello NUTS 3 tra i 150 mila e gli 800 mila. Per l'Italia le unità territoriali NUTS 2 coincidono con le Regioni.

[4] Non sono, quindi, considerati componenti della famiglia gli ospiti, i collaboratori domestici, gli affittuari di parte dell'abitazione e coloro che hanno lasciato definitivamente la famiglia, anche se non hanno ancora cambiato la residenza anagrafica. Inoltre, nel caso in cui più famiglie anagrafiche coabitino in una stessa casa, se queste costituiscono ai fini statistici famiglie di fatto vengono considerate come famiglie distinte nel processo di campionamento e viene intervistata soltanto quella eventualmente estratta. L'adozione della famiglia quale unità di rilevazione delle Forze di Lavoro segue la teoria economica secondo la quale questa è l'unità istituzionale che organizza e coordina sia la domanda per consumi sul mercato dei beni, sia l'offerta di lavoro dei propri componenti sul mercato dei fattori della produzione.

ma vengono ricompresi tutti coloro che effettivamente (occupati) o potenzialmente (disoccupati) partecipano al mercato del lavoro.

Un'altra importante innovazione metodologica introdotta nella RCFL e finalizzata a cogliere tutti quei fenomeni legati alla maggiore mobilità che caratterizza l'attuale mercato del lavoro è relativa alla rilevazione delle notizie riguardanti i componenti temporaneamente assenti del nucleo famigliare, per tramite dei familiari presenti. In particolare, sono considerati assenti temporanei coloro che sono lontani dal domicilio per svolgere un lavoro stagionale o temporaneo in altro Comune o all'estero, per svolgere servizio di leva o civile costitutivo, per svolgere il noviziato religioso, per ricovero in istituti di cura, per detenzione in attesa di giudizio, per viaggio di affari, turismo o breve cura, per necessità di servizio all'estero per conto dello Stato, per missioni di lavoro, per frequenza di corsi di qualificazione o aggiornamento professionale, per imbarco su navi della marina militare o mercantile.

Particolare importante ai fini della presente analisi costituisce la focalizzazione della RCFL sulla rilevazione di maggiori e più significative indicazioni circa l'evoluzione dinamica della situazione lavorativa degli intervistati, sia in termini di situazione occupazionale, sia di mobilità territoriale, e ciò, principalmente, in virtù, della sua caratterizzazione di continuità. Questo aspetto della RCFL viene sviluppato attraverso la richiesta, in apposite sezioni del questionario, di una serie di informazioni riguardanti sia la condizione lavorativa retrospettiva nei periodi precedenti la settimana di rilevazione, sia le eventuali variazioni di residenza intervenute nello stesso periodo. In particolare, nella Sezione I del questionario della RCFL, denominata Condizione autopercepita attuale e un anno prima e residenza, vengono richieste informazioni relativamente sia alla posizione lavorativa attuale, sia alla posizione lavorativa relativa allo stesso mese dell'anno precedente. Relativamente alla posizione lavorativa rivestita nell'anno precedente, vengono chieste una serie di informazioni sulla tipologia di contratto, se alle dipendenze o meno, se a tempo determinato o indeterminato, nonché in quale settore di attività economica fosse impiegato. Inoltre, viene chiesta anche la residenza dell'intervistato nello stesso mese dell'anno precedente, al fine di valutare anche gli spostamenti territoriali dello stesso.

Per quanto riguarda, però, la condizione professionale dichiarata dagli intervistati, occorre utilizzare una certa cautela nell'interpretare i dati retrospettivi, i quali possono scontare quello che tecnicamente viene indicato come effetto di autopercezione, in base al quale un lavoratore in una certa posizione lavorativa si ritiene in una posizione completamente diversa. Ciò può avvenire sia perché l'intervistato non è a conoscenza delle esatte definizioni relativamente alla situazione occupazionale, sia perché non conosce esattamente le condizioni contrattuali cui è sottoposto. Nel

primo caso, ad esempio, può verificarsi che un lavoratore assente per un tempo prolungato dal posto di lavoro si ritenga pienamente occupato, mentre ciò non è nelle statistiche ufficiali, mentre nel secondo caso alcuni lavoratori con contratti non standard possono sentirsi lavoratori dipendenti, principalmente perché operano a tempo pieno in un'azienda senza alcuna libertà di decisione circa gli orari e le modalità di espletamento del lavoro, nel mentre sono a tutti gli effetti lavoratori autonomi. Al fine di limitare le distorsioni dovute all'effetto di autopercezione nella Sezione I vengono utilizzate una serie di domande di controllo simili, ma in numero significativamente ridotto, a quelle della sezione che si occupa di rilevare le notizie sullo stato attuale di occupazione (Sezione C) dello stesso questionario.

Un'altra interessante sezione del questionario della RCFL è la Sezione E, denominata Precedenti esperienze di lavoro e destinata ai non occupati. In tale sezione si richiede se in passato l'intervistato svolgesse un'attività lavorativa e, in caso positivo, in quale anno avesse svolto tale attività lavorativa e in che anno avesse cessato di lavorare. Inoltre, viene richiesta l'età cui si è svolta l'ultima attività lavorativa, la tipologia di contratto, la condizione professionale, il settore di attività economica relativo al lavoro svolto, nonché le motivazioni che hanno determinato la cessazione dell'attività. Queste informazioni risultano di particolare interesse per valutare la condizione dei cosiddetti lavoratori scoraggiati, cioè coloro che pur non lavorando non cercano attivamente un'occupazione. In particolare, attraverso le informazioni di questa sezione si può analizzare in maniera strutturale la classe di individui denominata NEET (Not in Education, Employment, or Training), ovvero coloro che, nella maggior parte giovani, sono disoccupati e che al contempo non sono impegnati né in un corso di studi, né in attività di formazione[5].

2.1 L'analisi longitudinale attraverso i dati sulle Forze di Lavoro

Come già accennato in precedenza, la trasformazione dell'indagine trimestrale sulle Forze di Lavoro in indagine continua ha consentito l'implementazione di analisi di tipo dinamico sull'evoluzione del mercato del lavoro dal lato dell'offerta.

In tal senso, recentemente l'ISTAT ha intrapreso una politica di pubblicazione di micro-dati relativi alla RCFL, oltre che secondo la tradizionale struttura trasversale, anche seguendo la struttura longitudinale, ovvero la pubblicazione dei dati relativi alle stesse unità di rilevazione in periodi temporali successivi.

[5] Il riferimento è a qualsiasi tipo di educazione scolastica o universitaria e a qualsiasi genere di processo formativo, quali corsi professionali regionali o di altro tipo (tirocini, stage ecc.), attività educative (seminari, conferenze, lezioni private, corsi di lingua, informatica ecc.), con la sola esclusione delle attività formative informali quali l'autoapprendimento.

La possibilità di seguire dinamicamente la condizione professionale degli intervistati nel tempo deriva dalla struttura di campionamento utilizzata nella RCFL. Infatti, lo schema di campionamento utilizzato[6] è basato su un sistema di rotazione delle famiglie in base al quale la metà delle famiglie intervistate in un determinato trimestre T viene intervistata nuovamente a distanza di 3 e 12 mesi, mentre un quarto delle famiglie viene intervistato nuovamente a distanza di 15 mesi. In particolare, l'ISTAT fornisce trimestralmente due subset di micro-dati relativi alle unità di rilevazione che sono state intervistate in quel determinato trimestre (T) e nel trimestre precedente (T-1) o nello stesso trimestre dell'anno precedente (T-4), utili per un'analisi dinamica sia di tipo congiunturale, sia di tipo tendenziale.

L'ISTAT, a partire dall'aprile 2016, fornisce i file di micro-dati longitudinali per ciascuno dei quattro trimestri di ciascun anno. I dati contenuti nei database sono coerenti con la popolazione ricostruita in base alle risultanze censuarie e alla revisione post-censuaria delle anagrafi comunali, per cui i dati in essi contenuti non sono compatibili con quelli contenuti nei database precedenti a tale data. Al contempo, però, l'ISTAT sta effettuando un grosso lavoro di omogeneizzazione dei criteri di costruzione dei database longitudinali precedenti all'aprile 2016, sostituendo man mano che il lavoro di ricostruzione va avanti i vecchi database ricostruiti secondo i nuovi criteri.

Nell'analisi dei dati rivenienti nel database longitudinali è comunque necessario ricordare che tali dati non costituiscono un vero e proprio panel relativo a tutta la popolazione di interesse, poiché un individuo intervistato in uno dei comuni campione non verrà intervistato nuovamente se nel periodo di tempo intercorrente tra le interviste successive avrà cambiato residenza. In altre parole, la componente longitudinale della RCFL rappresenta solo la parte di popolazione residente in uno stesso comune sia all'inizio che alla fine del periodo considerato. Risulta evidente che il livello di precisione delle stime longitudinali risulti più basso rispetto alle stime trimestrali, per cui i risultati relativi ai sottogruppi di popolazione e/o ambiti territoriali ristretti potrebbero essere caratterizzati da un grado di incertezza molto elevato.

Premesso ciò, però, non si può non evidenziare che la disponibilità degli archivi longitudinali della RCFL permette di costruire uno strumento di analisi dinamica

[6] Il disegno campionario della RCFL si basa su uno schema di campionamento a due stadi con stratificazione delle unità di primo stadio, i comuni, e rotazione delle unità di secondo stadio, le famiglie. Esso prevede la sostituzione di una parte delle famiglie campione nei periodi successivi di rilevazione, per cui i campioni trimestrali risultano parzialmente sovrapposti in base ad uno schema di rotazione di tipo 2-2-2, in relazione al quale una famiglia viene inclusa nel campione per due rilevazioni successive, ne resta esclusa per i due successivi trimestri e, infine, viene reinserita nel campione per ulteriori due trimestri.

molto interessante ai fini dello studio dell'evoluzione del mercato del lavoro, le cosiddette *matrici di transizione*. Tali matrici permettono di valutare i flussi in entrata e in uscita relativamente ad una determinata condizione professionale fra due periodi di tempo, consentendo, quindi, di verificare le conseguenze che all'interno del mercato del lavoro possono aver avuto origine da crisi economiche, interventi di politica economica e finanziaria o interventi legislativi specifici dello stesso mercato del lavoro.

La costruzione di queste matrici utilizzando gli archivi trasversali della RCFL risulta, però, limitata dalla natura stessa del disegno campionario dell'indagine, che ha come finalità quella di fornire stime trimestrali trasversali dei principali indicatori del mercato del lavoro in base a quanto richiesto dai regolamenti comunitari 1575/2000 (Commission of the European Communities, 2000:1), e 1897/2000 (Commission of the European Communities, 2000:2).

Infatti, il campione trasversale della RCFL fornisce una stima della distribuzione per condizione professionale degli intervistati sia ad inizio che a fine periodo. Come già accennato, però, all'interno del periodo considerato parte della popolazione iniziale può uscire dal campione a causa di cambiamento di residenza, emigrazione o decesso, mentre, al contempo, possono entrare a far parte del campione finale gli individui che nel frattempo hanno compiuti 15 anni di età o che sono immigrati nel comune considerato. Nel primo caso, quindi, per alcuni individui costituenti il campione è nota soltanto la condizione professionale iniziale, mentre nel secondo caso è nota soltanto la condizione professionale finale.

3. I dati amministrativi per il mercato del lavoro

3.1 Il database del Ministero del Lavoro e delle Politiche Sociali

A partire dal 1997 numerosi sono gli interventi in materia di regolamentazione del mercato del lavoro italiano che il legislatore ha disposto attraverso l'approvazione di leggi e decreti legislativi non solo per riformare profondamente le modalità di accesso al mercato del lavoro, al fine di incrementare l'occupazione e prevenire situazioni critiche di disoccupazione giovanile e disoccupazione di lunga durata, ma anche per agevolare l'incontro tra domanda e offerta di lavoro a tutela altresì della parità di genere. In particolare, è con il D.Lgs. n. 469/1997 che, ai sensi dell'art. 1 della L. 59/1997 come modificata dalla L. 127/1997, si comincia a disciplinare il conferimento a Regioni ed Enti locali di funzioni e compiti per indirizzare il collocamento sul mercato del lavoro di disoccupati, lavoratori e persone in

cerca di prima occupazione, e promuovere politiche attive territoriali, fatte salve le materie di competenza del Ministero del Lavoro e delle Politiche Sociali come previsto dalla suddetta L. 59/1997.

E' proprio in relazione alla formalizzazione dei nuovi compiti e funzioni conferiti alle Regioni in materia di mercato del lavoro che vengono istituiti i centri per l'impiego, distribuiti a livello territoriale sulla base di un bacino di utenza provinciale non inferiore alle 100 mila unità (fatta eccezione per quelle situazioni socio geografiche opportunamente motivate) e costituiti per ricoprire appieno il ruolo detenuto sin dal 1949 da uffici periferici e strutture del Ministero del Lavoro e delle Politiche Sociali, che vengono soppressi proprio a decorrere dall'istituzione dei centri per l'impiego e comunque non oltre il 1 gennaio 1999 (D.lgs. 469/1997). L'importanza dei centri per l'impiego risulta essere considerevole non solo ai fini organizzativi delle tante iniziative definite dal legislatore per agevolare l'incontro tra domanda e offerta di lavoro sul territorio, ma anche ai fini del monitoraggio della quasi totalità degli aspetti legati al mercato del lavoro.

Quest'ultimo aspetto non è per nulla trascurabile, soprattutto in vista delle numerose riforme del lavoro che si sono susseguite e che definiscono le diverse forme contrattuali utilizzabili ancora oggi sia dai datori di lavoro privati, sia dagli enti pubblici economici. In relazione a quanto appena detto, al fine di rendere quanto più operativi i centri per l'impiego nelle funzioni e compiti assegnati in primis dal D.lgs. 469/1997, cui hanno fatto seguito ulteriori modificazioni e perfezionamenti legislativi, venne istituito sempre dallo stesso decreto legislativo di cui sopra il Sistema Informativo Lavoro, sostituito con la L. 30/2003 dalla omologa, in termini funzionali, Borsa Nazionale del Lavoro, che è un insieme di strutture organizzative e risorse informatiche hardware e software collegate in rete per consentire alle istituzioni amministrative (Ministero del Lavoro e delle Politiche Sociali, Regioni ed Enti locali) e ai soggetti abilitati alla mediazione tra domanda e offerta di lavoro[7] di rilevare, elaborare e diffondere dati in materia di collocamento e politiche del lavoro attive.

Il D.Lgs 181/2000, successivamente modificato e corretto dal D.Lgs 297/2002, definisce a seguire le disposizioni per agevolare l'incontro fra domanda ed offerta di lavoro in attuazione della L. 144/1999 in materia di investimenti, riordino degli incentivi all'occupazione e della normativa che disciplina l'INAIL, nonché riordino

[7] Per "soggetti abilitati" si intendono tutti i soggetti obbligati direttamente (datori di lavoro privati, esclusi i datori di lavoro domestico e gli armatori, pubbliche amministrazioni, enti pubblici economici e agenzie di somministrazione) e tutti gli organismi che ai sensi della normativa vigente e secondo le modalità e norme stabilite da ogni Regione e Provincia Autonoma possono effettuare le comunicazioni per proprio conto o terzi (ad esempio, consulenti del lavoro, dottori commercialisti, associazioni di categoria, periti agrari e agrotecnici, soggetti autorizzati all'attività di intermediazione).

degli enti previdenziali. In relazione a quanto previsto dal D.Lgs 469/1997 e successive modificazioni, e dalla L. 144/1999, il D.Lgs 181/2000 sopprime del tutto le liste di collocamento ordinarie e speciali, tranne le eccezioni previste dal decreto n. 2053/1963 del Presidente della Repubblica, dalla L. 223/1991 e dalla L. 68/1999, e definisce le modalità di assunzione e gli adempimenti successivi alla stessa (art. 4bis del D.Lgs. 181/2000, ma anche art. 9bis del D.lgs. 510/1996, già L.608/1996, che fa riferimento al lavoro autonomo in forma coordinata e continuativa) in relazione ai quali i datori di lavoro privati e gli enti pubblici procedono direttamente all'assunzione del lavoratore per qualsiasi tipologia di contratto, tranne i casi in cui sia obbligatorio il concorso pubblico, perché previsto dallo statuto degli enti, ed effettuano obbligatoriamente la comunicazione ai Servizi competenti (ovverosia i centri per l'impiego di cui al D.Lgs 469/1997 e gli altri organismi autorizzati e accreditati) nel cui ambito territoriale è ubicata la sede di lavoro entro e non oltre le ore 24 del giorno precedente l'inizio del rapporto di lavoro. Lo stesso D.Lgs 181/2000 dispone che qualunque variazione in essere del rapporto di lavoro, sia essa una cessazione, una trasformazione o una proroga, deve essere prontamente e obbligatoriamente comunicata entro il termine perentorio dei 5 giorni lavorativi sempre ai Servizi competenti presso i quali è ubicata la sede lavorativa.

Al fine di agevolare i datori di lavoro pubblici e privati, le comunicazioni obbligatorie suddette devono essere trasmesse obbligatoriamente per via telematica (D.I. 30 ottobre 2007, in attuazione della Legge Finanziaria 2007 n. 296/2006) per mezzo dei sistemi informatici resi disponibili presso le varie sedi dei Servizi Competenti secondo le modalità definite da ciascuna Regione o Provincia Autonoma.

Nello specifico, il Sistema Informatico per le Comunicazioni Obbligatorie (C.O., d'ora innanzi) si basa sulla interoperabilità dei sistemi locali realizzati dalle Regioni e dalle Province Autonome di Trento e Bolzano, e costituisce l'unico punto di accesso per l'invio telematico delle C.O. in un'unica soluzione contemporaneamente al Ministero del Lavoro e delle Politiche Sociali, all'INAIL, all'INPS, agli Uffici territoriali di Governo ed agli altri enti previdenziali sostitutivi o esclusivi, e mediante l'utilizzo di modelli elettronici standard utilizzati a seconda dei casi (instaurazione, trasformazione, proroga e cessazione di un rapporto di lavoro).

Per quanto descritto, si percepisce anche intuitivamente l'importanza da un lato politica, dall'altro statistica che il nuovo sistema di tracciabilità riveste per la conoscenza e gestione del mercato del lavoro, soprattutto per quel che riguarda il lavoro alle dipendenze. La mole di informazioni che viene raccolta attraverso le C.O., oltre che essere enorme, risulta essere anche completa, con l'eccezione dei casi di dichiarazioni mendaci che ci si augura sempre siano in numero ridotto, e puntuale, anche

per quel che riguarda datori di lavoro e lavoratori stranieri e agenzie per il lavoro estere (UE ed extra UE). Dal punto di vista statistico, quindi, la banca dati in possesso di ogni Regione e Provincia Autonoma per conto dei propri centri per l'impiego può rappresentare una fonte inestimabile di informazioni che da un lato completerebbe il dato rilevato con la RCFL dell'ISTAT, con le eccezioni del caso proprie della rilevazione stessa, dall'altro potrebbe consentire analisi specifiche "micro" territoriali e di settore considerata la dimensione individuale del dato. In quest'ottica, anche la validità e omogeneità del dato è garantita dal Sistema Informatico C.O. La validazione delle C.O. al momento dell'invio viene, infatti, assicurata dal fatto che un sistema di controllo sulla piattaforma, basato principalmente sulla correttezza della codifica di ogni variabile e dell'intersezione di variabili di controllo opportunamente definite, garantisce la coerenza dei dati in relazione anche alle definizioni standard già presenti nel sistema. Infatti, Il Sistema Informatico C.O. si basa su un insieme di informazioni e dizionari standard che garantisce l'uniformità e la facile e sicura condivisione delle informazioni a livello nazionale. Le "istruzioni" standard delle C.O. fanno proprio riferimento ai dati e alle informazioni contenuti nelle stesse, e ad alcune terminologie specifiche per la classificazione di informazioni già ritenute cruciali e fondamentali. Naturalmente, in relazione sia alle modalità di gestione delle C.O., sia all'evoluzione del mercato del lavoro, foss'anche la sola riforma del lavoro in continua trasformazione per sostenere l'occupazione, le istruzioni standard e di conseguenza le C.O. sono sottoposte a continui aggiornamenti.

3.1.1 Struttura e Contenuti delle Comunicazioni Obbligatorie

Le C.O., come dianzi già specificato, rappresentano un obbligo rispetto al quale nessun datore di lavoro pubblico e privato, tranne le eccezioni del caso come previsto dalla normativa vigente, può sottrarsi. Il Servizio Informatico C.O. abbraccia tutti i settori ATECO del sistema economico produttivo italiano attraverso l'implementazione di 6 moduli, ovverosia 6 modelli standard in base ai quali deve essere redatta la comunicazione obbligatoria in relazione alle varie circostanze. Ogni modello è a sua volta costituito da sezioni e, in alcuni casi specifici, sottosezioni. Inoltre, come si vedrà a breve, alcuni moduli sono interconnessi con altri moduli.

I due moduli chiave del Servizio Informatico C.O., maggiormente utilizzati perché abbracciano le casistiche più comuni e i cui contenuti verranno descritti dettagliatamente in seguito, sono l'Unificato LAV (UniLAV, d'ora innanzi) per i datori di lavoro pubblici e privati, e l'Unificato SOMM (UniSOMM, d'ora innanzi) per le Agenzie di Somministrazione. Gli altri moduli, a seguire, sono: l'Unificato URG (UniURG, d'ora innanzi) valido per qualsiasi settore per assunzioni d'urgenza dei

lavoratori, l'Unificato LAV-Cong (UniLAV-CONG, d'ora innanzi) utilizzato unicamente per l'assunzione congiunta di più lavoratori nel settore agricolo, l'Unificato VARDATORI (UniVAR, d'ora innanzi) unicamente utilizzato per comunicare le variazioni di diversa natura riguardanti la sola azienda (ragione sociale del datore di lavoro, incorporazione, fusione, cessione ramo d'azienda, ecc.), ed infine un modulo apposito per la "Comunicazione semplificata per l'assunzione d'urgenza nel settore del Turismo" (ComURG-TURISMO, d'ora innanzi). In particolare, sia l'UniLAV-CONG che l'UniVAR strutturalmente richiamano in parte l'UniLAV, anche per quel che riguarda la maggior parte delle informazioni raccolte. Per quel che riguarda, invece, l'UniURG e la ComURG-TURISMO, queste due tipologie di C.O. risulteranno completate e validate dal Servizio Informatico C.O. solo a fronte anche dell'invio dell'UniLAV entro e non oltre i 5 giorni per l'UniURG e 3 giorni per la ComURG-TURISMO successivi alla data di instaurazione del rapporto di lavoro.

Fra tutte le tipologie contrattuali, l'unica che non deve fare riferimento al Servizio Informatico C.O. per la registrazione del rapporto di lavoro è quello relativa all'assunzione di colf e badanti. Infatti, in base alla normativa vigente (L. 2/2009) il rapporto di lavoro domestico deve essere comunicato all'INPS mediante apposito modulo presente sulla pagina web dell'Ente. Successivamente, sarà l'INPS stessa a comunicare le posizioni lavorative domestiche attivate al Servizio Informatico C.O. per mezzo del nodo regionale di interscambio informativo tra enti, di cui sopra. Prima di procedere ad una breve descrizione dei due moduli principali, risulta importante evidenziare un aspetto comune a tutti i moduli.

Numerose variabili sono comuni ai 6 moduli, alcune di queste anche facoltative, ed indubbiamente il ruolo più importante ai fini di indagini statistiche mirate all'analisi del mercato del lavoro attraverso i suoi principali protagonisti, datori di lavoro e lavoratori, è il codice fiscale. Oltre ad altri campi ugualmente importanti, tutte le C.O. vedono la compilazione obbligatoria del campo "Codice Fiscale" prima ancora della denominazione del datore di lavoro e/o del lavoratore. Tralasciando del tutto quello che è il ruolo ormai strategico che ha assunto il codice fiscale individuale negli anni a fini amministrativi, emerge chiaramente come oramai lo stesso rivesta anche un ruolo strategico per analisi statistiche puntuali e dettagliate che vadano ben oltre l'utilizzo del macrodato di sintesi.

Nello specifico, può diventare realtà la possibilità di tracciare, sempre nel rispetto della normativa vigente in merito all'utilizzo di dati sensibili per fini statistici, il percorso individuale nel mercato del lavoro dei lavoratori assunti almeno a partire dall'anno 2008, anno in cui si inizia ad ottemperare la normativa vigente in

merito alle C.O., per evidenziarne carriere e criticità legate, ad esempio, a titoli di studio posseduti o domicilio o nazionalità, ma anche per valutare le politiche, anche fiscali, messe in atto dai governi centrali in materia di "lavoro". Allo stesso tempo, inoltre, si renderebbe possibile anche produrre analisi dal lato dei datori di lavoro per evidenziarne le decisioni in relazione, ad esempio, ai vari rapporti di lavoro possibili ed utilizzabili, la cui produzione si è susseguita negli anni sempre a partire dal 2008 e/o all'utilizzo di agevolazioni di tipo fiscale o normativo per l'instaurazione di un rapporto di lavoro piuttosto che di un altro, od anche per la trasformazione di un rapporto di lavoro già instaurato.

Di seguito, si procede alla breve descrizione dei due moduli fondamentali del Sistema Informatico C.O. e, soprattutto, delle principali variabili in essi contenuti.

Il modulo Unificato LAV

Tutti i datori di lavoro sia pubblici che privati appartenenti a qualunque settore di attività economica, ad eccezione delle agenzie per il lavoro in somministrazione per le quali, come dianzi visto, è previsto un modulo di C.O. specifico (UniSOMM), sono obbligatoriamente tenuti ad inviare direttamente o mediante soggetti abilitati al Ministero del Lavoro e delle Politiche Sociali la comunicazione di assunzione di nuovi lavoratori, o di trasformazione, proroga o cessazione di rapporti di lavoro già esistenti. Il modulo del Servizio Informatico C.O. preposto per tale obbligo è il modello UniLAV.

Può fare eccezione a questo obbligo l'assunzione dei lavoratori agricoli a tempo determinato, per i quali è previsto un modulo specifico laddove si tratti di contestuale assunzione di più soggetti (UniLAV-CONG). L'UniLAV è a sua volta composto di 9 Sezioni ognuna delle quali registra informazioni con finalità ben precise:

- Sezione *1 – Datore di Lavoro*: in questa sezione sono riportati i dati identificativi del datore di lavoro e del legale rappresentante dell'azienda (per le piccole/medie imprese le due figure possono anche coincidere), o dell'Ente pubblico. Inoltre, in questa stessa sezione dell'UniLAV vengono indicati sia il settore ATECO di appartenenza dell'azienda, se si tratta di privato, la sua sede legale e la sede in cui si svolgerà la prestazione lavorativa del lavoratore per il quale si è attivata la procedura di C.O. È sempre in questa prima sezione che si registra se, al contrario, si tratti di Pubblica Amministrazione (PA, d'ora innanzi) e tutte le corrispondenti informazioni relative alla localizzazione legale e operativa.
- Sezione *2 – Lavoratore*: in questa sezione sono riportati i dati identificativi del lavoratore, compresi cittadinanza, domicilio e livello di istruzione.

- Sezione *3 – Lavoratore co-obbligato*: questa sezione è identica nei contenuti alla sezione precedente, con l'unica differenza di riferirsi al lavoratore coobbligato nel caso si tratti di contratto di lavoro ripartito[8].
- Sezione *4 – Inizio*: in questa sezione sono registrati tutti i dati identificativi del rapporto di lavoro che si va ad instaurare e relativi a data di inizio e fine del rapporto (tranne che non si tratti di contratto di lavoro a tempo indeterminato), data di fine periodo formativo, se trattasi di rapporto di apprendistato, ente previdenziale e relativo codice al quale vengono versati i contributi, Posizione Assicurativa Territoriale INAIL del datore di lavoro, codice agevolazione INPS per il versamento dei contributi, tipologia contrattuale, tipologia di orario, qualifica professionale ISTAT, contratto collettivo nazionale applicato e corrispondente inquadramento, compenso annuo lordo, e se trattasi di lavoratore in mobilità o socio cooperativa o lavoratore appartenente a fascia protetta o lavoratore agricolo o lavoratore stagionale.

Le restanti 4 Sezioni del modulo UniLAV, ad eccezione della Sezione ULTIMA, perché obbligatoria, vengono compilate solo nei casi in cui si verifica la condizione che ne richiede una C.O.

- Sezione *5 – Proroga*: questo quadro di UniLAV è compilato nei casi in cui si tratti di contratto a termine o di durata temporanea e se ne voglia prolungare oltre il termine di scadenza la durata senza necessariamente trasformarlo in altro rapporto, oppure qualora si tratti di prosecuzione del contratto quando la scadenza non è definibile a priori (vedasi il caso di sostituzione per maternità). I dati che si rilevano in questa sezione sono gli stessi della Sezione 4, con l'aggiunta di informazioni relative alla data di inizio del rapporto di lavoro originario e la data del nuovo termine del rapporto come conseguenza della proroga.
- Sezione *6 – Trasformazione*: questa sezione è compilata dal datore di lavoro unicamente nei casi in cui si tratti di trasformazione del rapporto di lavoro, di trasferimento, o distacco/ comando del lavoratore. Anche in questo caso, come nel caso della Sezione 5, i dati che si rilevano in questa sezione sono gli stessi della Sezione 4, con l'aggiunta di informazioni relative alla data di trasformazione, al codice di trasformazione ed ad una serie di informazioni specifiche relative ai due casi distinti di trasferimento o distacco/comando (ad esempio, informazioni della precedente localizzazione del lavoro nel primo caso, o dati

[8] In base al D.lgs. 276/2003, il contratto di lavoro ripartito è un particolare contratto di lavoro mediante il quale due lavoratori assumono "in solido" l'adempimento di un'unica e identica obbligazione lavorativa. I lavoratori, pur restando ognuno personalmente e direttamente responsabile dell'adempimento in carica, possono gestire autonomamente attività e orari.

identificativi del datore di lavoro distaccatario / comandatario e della relativa azienda nel secondo).

- Sezione *7 – Cessazione*: questa sezione si applica per comunicare l'interruzione di un rapporto di lavoro qualunque esso sia (sia a tempo indeterminato, sia a tempo determinato), di una proroga, di un trasferimento, di un distaccamento. Anche in questo caso, le informazioni che si registrano in questa sezione sono le stesse della Sezione 4, con l'aggiunta di informazioni relative alla data di cessazione del rapporto ed alle relative cause.
- Sezione *8 – Tirocini*: la compilazione di questa sezione è obbligatoria per la comunicazione dei soli tirocini extracurriculari. In questa sezione sono indicate le informazioni relative al soggetto promotore, alla tipologia di tirocinio e alla categoria di tirocinante. In particolare, in relazione a quest'ultima variabile è definita la durata del tirocinio che varia, per legge, in relazione a che si tratti di un lavoratore in mobilità o persona presa in carico dai servizi sociali e/o sanitari o disabile o soggetto svantaggiato o disoccupato o ex studente oggi neoqualificato sia esso diplomato/laureato/dottorato/specializzato.
- Sezione *ULTIMA – Dati Invio*: questa sezione, non numerata in questa sede, costituisce in realtà la ricevuta di ricezione da parte del Servizio Informatico C.O. e rappresenta a tutti gli effetti una prova per il datore di lavoro dell'avvenuto adempimento dell'obbligo previsto per legge. Essa registra oltre ai dati identificativi del datore di lavoro o del soggetto abilitato che ne fa le veci che effettua la C.O., il tipo di C.O., se trattasi di assunzione obbligatoria e, nel caso, le ragioni della stessa. Inoltre, in questa sezione è registrato il codice identificativo della C.O. in essere come assegnato dal Sistema Informativo quando la C.O. è presa in carico e, eventualmente occorra, anche il codice identificativo della C.O. precedente (vedasi quanto detto dinanzi in relazione all'UniURG).

Il modulo Unificato SOMM

Attraverso il modello UniSOMM, le agenzie per il lavoro, similmente a quanto previsto per i datori di lavoro privati e pubblici, hanno l'obbligo di inviare per via telematica al Ministero del Lavoro e delle Politiche Sociali la comunicazione relativa a tutti i rapporti lavorativi di somministrazione instaurati ed eventuali variazioni degli stessi (proroga, trasformazione e cessazione).

La struttura del modulo UniSOMM si presenta leggermente più semplificata rispetto a quella dell'UniLAV, tant'è che il numero di sezioni in esso contenute ammonta a 7 anziché 9, e ciò è dovuto al fatto che tutte le variazioni del rapporto sono accorpate in un'unica sezione attraverso sottosezioni che, peraltro, contengono solo

le informazioni fondamentali della variazione comunicata. Inoltre, alcune sezioni risultano simili alle corrispondenti sezioni dell'UniLAV, tranne chiaramente per il fatto che si faccia riferimento non già a datori di lavoro, ma ad agenzie per il lavoro. Questo è il caso della Sezione *1 – Agenzia di Somministrazione*, della Sezione *2 – Lavoratore*, e della Sezione *ULTIMA – Dati Invio*, che difatti presentano gli stessi dati e le stesse informazioni. Le ultime quattro sezioni del modulo UniSOMM sono, invece, specifiche per i rapporti di lavoro in somministrazione, e la procedura di assunzione e reclutamento del lavoratore interessato. Di seguito si descrivono brevemente i contenuti e la struttura delle sezioni del modello UniSOMM non comuni con il modello UNILAV.

- Sezione *3 – Rapporto Agenzia/Lavoratore*: questa Sezione riporta i dati relativi al rapporto di somministrazione che si instaura tra Agenzia di Somministrazione, che a tutti gli effetti farà da intermediario tra individuo e mercato del lavoro unicamente per questa tipologia di contratto, e il Lavoratore. I principali dati raccolti in questa Sezione della C.O. riguardano la data di inizio e fine del rapporto di somministrazione (tranne che non si tratti di rapporto a tempo indeterminato), data di fine periodo formativo se trattasi di rapporto di apprendistato, ente previdenziale e relativo codice al quale vengono versati i contributi, tipologia contrattuale e se trattasi di lavoratore in mobilità. Per molti versi, questa sezione appare molto simile alla Sezione 4 di UniLAV.
- Sezione *4 – Ditta Utilizzatrice*: in questa sezione verranno indicati tutti i dati identificativi della ditta che farà ricorso ad un lavoratore in somministrazione per il tramite dell'Agenzia del lavoro. In questa stessa sezione dell'UniSOMM viene indicata la data di inizio e di fine del contratto di somministrazione che si instaura tra Agenzia di Somministrazione e Ditta Utilizzatrice, ed inoltre altre informazioni relative al settore ATECO di appartenenza della ditta se si tratta di privato, la sua sede legale e la sede in cui si svolgerà la "missione" del lavoratore per il quale si è attivata la procedura di C.O. Ed è sempre in questa prima sezione che si registra se, al contrario, si tratti di PA e tutte le corrispondenti informazioni relative alla localizzazione legale e operativa. Anche in questo caso, la strutturazione di questa sezione richiama quella della Sezione 1 di UniLAV.
- Sezione *5 – Rapporto Ditta Utilizzatrice/Lavoratore*: questa sezione registra tut–ti i dati identificativi della missione, ovverosia la prestazione di lavoro, del lavoratore presso la ditta utilizzatrice. I dati in questione si riferiscono a data di inizio e fine della missione, tipologia di orario, qualifica professionale ISTAT, contratto collettivo nazionale applicato e corrispondente inquadramento, descrizione attività, voce di tariffa INAIL della lavorazione prestata, se trattasi di la-

voratore agricolo, se trattasi di attività rischiosa per esposizioni al biossido di silicio (silicosi) o inalazione di fibre di asbesto (asbestosi).

- Sezione *6 – Variazione*: la seguente sezione registra tutte le possibili variazioni che possono intercorrere nel rapporto di somministrazione o nella missione, e riguardano la proroga, la trasformazione o la cessazione degli stessi. Ognuna di queste variazioni viene registrata nella corrispondente sottosezione con i dati relativi alla data di accadimento della variazione nel caso di trasformazione e cessazione, e relativi codici di identificazione degli stessi nella casistica tabulata, o alla data di termine della variazione nel caso di proroga.

3.2 Il database dell'INPS

La banca dati in possesso dell'INPS completa il quadro di riferimento relativo alle strutture di dati disponibili in merito al fenomeno dell'occupazione, sebbene le ragioni che ne determinano la compilazione siano di natura molto diversa da quelle precedentemente descritte relativamente ai database ISTAT e Ministero del Lavoro e delle Politiche Sociali. Dal punto di vista statistico, però, tale aspetto non ne sminuisce l'importanza e, anzi, al contrario definisce un elemento ulteriore di completezza e possibile integrazione dell'informazione relativa al fenomeno oggetto di studio.

La banca dati dell'INPS registra la consistenza dei rapporti di lavoro per il tramite delle denunce telematiche attraverso la piattaforma dell'Ente che, in base alla L. 326/2003, tutti i datori di lavoro sono obbligatoriamente tenuti a presentare ai fini della trasmissione dei dati contributivi e retributivi e relativi compensi riferiti al mese precedente per ogni singolo lavoratore assunto, ed entro l'ultimo giorno del mese successivo. Per i lavoratori dipendenti il mese di competenza è quello cui si riferisce la busta paga, mentre per i lavoratori parasubordinati il mese di competenza è quello in cui è stato erogato il compenso. La banca dati INPS contiene anche altri dati, sempre di natura finanziaria, relativi alla posizione assicurativa del lavoratore, al sostegno assistenziale cui è sottoposto, quale ad esempio maternità, malattia, congedo parentale, assegni familiari, per citarne alcuni, all'accantonamento TFR e corrispondenti informazioni correlate. Numerosi sono i dati relativi anche all'azienda datrice di lavoro al fine di inquadrarla nella banca dati INPS, quali codice fiscale, ragione sociale e posizione contributiva. In base al messaggio INPS n.ro 011903 del 2009, la denuncia mensile relativa ai lavoratori dipendenti e parasubordinati da parte delle aziende datrici di lavoro avviene per il tramite di un sistema di inoltro denominato UNIEMENS che sostituisce appieno, mediante l'accorpamento in un unico documento, i due modelli precedentemente utilizzati: DM10/2 per la comunicazione dei dati contributivi aggregati dei lavoratori presenti

in azienda, e EMENS per la comunicazione dei dati retributivi individuali e nominali di ogni singolo lavoratore.

Da quanto esposto, appare chiaro che il database INPS si presta a numerose applicazioni di analisi, non solo dal lato della domanda, ma anche dell'offerta, e il dato sulla consistenza dell'occupazione indirettamente determinato completa quello già desumibile dal database del Ministero Lavoro e delle Politiche Sociali nella sua dimensione di flusso e quello del database ISTAT nella sua dimensione di stock.

Come già evidenziato dianzi, il Sistema Informatico per le C.O. garantisce l'invio telematico delle informazioni relative ai lavoratori dipendenti e collaboratori in unica soluzione tra i vari Enti e Istituzioni. Pertanto, la dichiarazione sulla posizione contributiva e retributiva UNIEMENS completa il quadro delle informazioni relative ai suddetti lavoratori. Ma non si può non evidenziare come all'interno delle banche dati INPS trovino collocazione le registrazioni contributive e retributive di tutte le altre tipologie di occupato diverse da quelle relative ai lavoratori dipendenti e parasubordinati: lavoratori delle Pubbliche Amministrazioni, lavoratori autonomi con Partita IVA non parasubordinati e non iscritti alle Casse Professionali, lavoratori assunti mediante il sistema dei buoni lavoro (i cosiddetti voucher), lavoratori degli Enti Pubblici Economici (molti dei quali già Società Pubbliche).

L'unico aspetto un po' incerto relativo al database INPS è legato alla natura essenzialmente amministrativo-contabile-finanziaria del dato. Nonostante le procedure amministrative di controllo e perfezionamento delle dichiarazioni UNIEMENS finalizzate alla normalizzazione del dato dichiarato, l'aggiornamento continuo dei dati, anche pregressi, a causa del miglioramento della piattaforma informatica, dei ritardi nella trasmissione oltre il tempo limite, della correzione delle denunce già inviate, rende utilizzabile l'informazione fruita con un margine di errore dell'ana–lisi eseguita che in alcuni casi rischia di essere elevato.

4. Conclusioni

Dall'analisi condotta, risulta evidente come i tre principali database riguardanti il mercato del lavoro italiano forniscano informazioni preziose ma parziali sulla dinamica dello stesso, principalmente perché utilizzano due approcci complementari al fenomeno, uno dal lato della domanda (Ministero del Lavoro e delle Politiche Sociali, e INPS) e l'altro dal lato dell'offerta (ISTAT) di lavoro.

La possibile integrazione tra gli archivi ISTAT e gli archivi amministrativi del Ministero e dell'INPS, non rientrando tale attività tra quelle previste nel Piano Statistico Nazionale, richiede la stesura di una specifica norma di legge, allo stato at-

tuale non disponibile, che consenta la condivisione da parte del Ministero e/o dell'INPS dei database in loro possesso con l'ISTAT comprensivi di quei dati personali che permettano l'accoppiamento dei record dei tre database relativi ad uno stesso individuo.

Nel presente lavoro si sono separatamente analizzati i tre database mettendone in evidenza la struttura e le potenzialità di ognuno ai fini di un approccio dinamico all'analisi del mercato del lavoro. I dati desumibili dalla RCFL risultano a tale scopo molto utili per l'analisi dinamica, in particolare per ciò che attiene l'analisi delle ripercussioni che le politiche di intervento legislativo e fiscale hanno sul mercato del lavoro. A tale scopo, in particolare, con le limitazioni evidenziate nel testo, risultano di grande utilità i dati longitudinali e le matrici di transizione da essi ricavabili.

Occorre, però, evidenziare come i dati rilevati dalla RCFL, relativamente alle posizioni lavorative degli intervistati, si basino sulla valutazione autopercepita della propria condizione professionale, il che può introdurre distorsioni in relazione a quello che è il reale inquadramento lavorativo. Inoltre, ogni individuo viene intervistato una sola volta all'interno del trimestre di riferimento, per cui possono sorgere problemi relativamente alla valutazione della variazione della condizione professionale per coloro che possono ricoprire più posizioni all'interno dello stesso trimetre o, ancor più, per coloro che cambiano condizione lavorativa nei due trimestri in cui non figurano nel campione della RCFL.

Di contro, però, la RCFL presenta il vantaggio di rilevare anche i lavoratori autonomi, permettendo un'analisi delle matrici di transizione più esaustiva, mentre ciò non accade per i dati contenuti nel database in possesso del Ministero, il quale non rileva affatto le posizioni relative ai lavoratori autonomi se non nella forma coordinata e continuativa.

Al contrario della RCFL, il database a disposizione del Ministero rileva oggettivamente la condizione di occupato alle dipendenze per il tramite di una comunicazione obbligatoria da parte dei datori di lavoro, il che permetterebbe di seguire l'evoluzione professionale di ogni singolo individuo nell'arco della sua vita. I dati sono registrati per posizione lavorativa, ma fino a quando non sarà varata un'apposita legge o sarà disposta una struttura di codici identificativi unici che permetta di criptare i dati personali e/o sensibili ai fini dell'utilizzo statistico dei dati stessi, non sarà possibile sviluppare alcuna analisi specifica sui microdati amministrativi.

Inoltre, come già evidenziato poc'anzi, il database del Ministero non registra i lavoratori autonomi tranne che nell'eccezione dianzi riportata, in quanto questi non hanno obbligo di comunicazione, per cui diviene difficile valutare un cambiamento

di posizione professionale, come ad esempio la stabilizzazione di un lavoratore a partita IVA verso un lavoro dipendente a tempo sia determinato che indeterminato.

Il database INPS, al contrario, supera in parte il limite dianzi evidenziato del database del Ministero per quanto esso stesso presenti alcune limitazioni importanti di altra natura. Il database INPS ha natura essenzialmente amministrativo-contabile e consente di ricavare la consistenza dell'occupazione in modo indiretto. Esso ha il grande vantaggio di registrare tutti i lavoratori per i quali si effettui un versamento di contributi e/o si apra una posizione assicurativa e, pertanto, completa di fatto l'informazione ricavabile dalle banche dati ISTAT e Ministero.

Di contro, però, oltre che presentare la stessa limitazione della banca dati del Ministero causata dall'assenza di una struttura di codici identificativi unica che tuteli la privacy del lavoratore e al tempo stesso permetta l'utilizzo del dato a fini statistici, risulta anche carente di tante altre informazioni utili relative al lavoratore quali, ad esempio, titolo di studio e residenza, per citarne alcuni.

Inoltre, un altro aspetto non trascurabile che caratterizza negativamente il data base INPS è legato all'instabilità più o meno elevata, a seconda dei casi, del dato in esso contenuto. Infatti, per il database INPS si assiste ad un continuo aggiornamento del dato che, nonostante il continuo monitoraggio della piattaforma telematica, può essere oggetto di correzione o registrazione tardiva oltre la scadenza prevista.

L'integrazione, pertanto, delle tre banche dati relative al mercato del lavoro attualmente disponibili (ISTAT) o potenzialmente tali (Ministero del Lavoro e delle Politiche Sociali, e INPS) ai fini di un'analisi dinamica del mercato del lavoro è possibile oltre che fortemente auspicata, per quanto non poche possono essere le difficoltà per omogeneizzare e stabilizzare un dato importante dal punto di vista economico e politico, quale appunto risulta quello relativo ad una grande fetta di occupati che non siano appartenenti alle categorie dei lavoratori dipendenti e parasubordinati.

Riferimenti bibliografici

Commission of the European Communities (1998). *Council Regulation (EC) No 577/98 of 9 March 1998 on the organisation of a labour force sample survey in the Community*. Brussels.

Commission of the European Communities (2000, 1). *Commission Regulation (EC) No 1575/2000 of 19 July 2000 implementing Council Regulation (EC) No 577/98 on the organisation of a labour force sample survey in the Community concerning the codification to be used for data transmission from 2001 onwards*. Official Journal of the European Communities. Brussels.

Commission of the European Communities (2000, 2). *Commission Regulation (EC) No 1897/2000 of 7 September 2000 implementing Council Regulation (EC) No 577/98 on the organisation of a labour force sample survey in the Community concerning the operational definition of unemployment. Official Journal of the European Communities*. Brussels.

Commission of the European Communities (2005). *Recommendation of the Commission on the independence, integrity and accountability of the national and Community statistical authorities, COM (2005)*. Brussels.

Decreto Legislativo 19 dicembre 2002, n. 297 "Disposizioni modificative e correttive del decreto legislativo 21 aprile 2000, n. 181, recante norme per agevolare l'incontro tra domanda e offerta di lavoro, in attuazione dell'articolo 45, comma 1, lettera a), della L. 17 maggio 1999, n. 144". Gazzetta Ufficiale Serie Generale n. 11 del 15.01.2003. Roma.

Decreto Legislativo 21 aprile 2000, n. 181 "*Disposizioni per agevolare l'incontro fra domanda e offerta di lavoro, in attuazione dell'articolo 45, comma 1, lettera a), della L. 17 maggio 1999, n. 144"*. Gazzetta Ufficiale Serie Generale n. 154 del 04.07.2000. Roma.

Decreto Legislativo 23 dicembre 1997, n. 469 "*Conferimento alle Regioni e agli Enti locali di funzioni e compiti in materia del mercato del lavoro, a norma dell'articolo della legge 15 marzo 1997, n. 59"*. Gazzetta Ufficiale Serie Generale n. 5 del 08.01.1998. Roma.

Deville J. C., Särndal C. E. (1992). *Calibration Estimator in Survey Sampling*. Journal of the American Statistical Association, vol. 87, pp. 376-382. London.

European Union (2015). *Regulation (EU) 2015/759 of the European Parliament and of the Council of 29 April 2015 amending Regulation (EC) No 223/2009 on European statistics*. Brussels.

ISTAT (2006). *La rilevazione sulle forze di lavoro: contenuti, metodologie, organizzazione*. Collana Metodi e norme n. 32 – 2006. Roma.

ISTAT (2014). *Rilevazione sulle Forze di Lavoro. Aspetti metodologici dell'in–dagine*. Roma.

ISTAT (2017). *Rilevazione sulle Forze di Lavoro. Dati longitudinali. Aspetti metodologici dell'indagine*. Roma.

Ministero del Lavoro e delle Politiche Sociali (2016). *Comunicazioni Obbligatorie. Modelli e Regole. Settembre 2016*. Roma.

Ministero del Lavoro e delle Politiche Sociali (2016). *Sistema CO. Controlli di coerenza dei dati. Settembre 2016*. Roma.

UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO *Metodi e analisi statistiche 2017* ISBN: 978-88-6629-013-1
DIPARTIMENTO DI ECONOMIA E FINANZA pp. 151-178

Il Cruscotto Regionale dell'Innovazione: una nuova metodologia di misurazione della performance innovativa delle regioni italiane

Rossana Mancarella*, Stefano Marastoni*

Agenzia Regionale per la Tecnologia e l'Innovazione - ARTI - della Puglia

Riassunto: Obiettivo principale del presente lavoro è quello di sperimentare un nuovo strumento di misurazione teso a quantificare la performance innovativa delle regioni italiane. Questo progetto è stato condotto con il proposito di distin–guersi dallo strumento già noto del *Regional Innovation Scoreboard* (RIS): a. accogliendo le indicazioni della letteratura più autorevole, secondo cui la quantificazione della *performance* innovativa di una qualsiasi unità territoriale debba ricorrere necessariamente ad un approccio metodologico che contempli uno spettro molto ampio ed articolato dei fenomeni e delle attività innovative; b. utilizzando, pertanto, un sistema di indicatori più significativo ed esauriente rispetto a quello definito e implementato nel RIS.

Keywords: Innovazione, *performance* innovativa, misurazione della *perfomance*, indice sintetico ponderato, sistemi innovativi regionali, RIS.

1. Cosa è il Cruscotto Regionale dell'Innovazione (CRI)

L'oggetto del presente lavoro consiste nell'illustrazione di un sistema di misurazione della performance innovativa delle regioni italiane – il Cruscotto Regionale dell'Innovazione (CRI) -, potenzialmente estendibile a tutte le regioni europee, che intende distinguersi da quello implementato nel *Regional Innovation Scoreboard*

* Autori corrispondenti: s.marastoni@arti.puglia.it ; r.mancarella@arti.puglia.it .
Il presente lavoro è frutto del lavoro comune degli Autori, ma è attribuibile a S. Marastoni la redazione dei capitoli 1, 2, 5 e del paragrafo 4.1 e a R. Mancarella quella del capitolo 3 e dei paragrafi 4.2 e 4.3.

(RIS) commissionato e diffuso dall'*European Commission, Directorate-General for Internal Market, Industry, Entrepreneurship and SMEs.*

Tale RIS è l'estensione, a livello di analisi regionale, dell'*European Innovation Scoreboard* (EIS) che, a partire dal 2001 è stato riproposto regolarmente nell'ambito dell'Unione Europea (EU), assumendo una crescente importanza per l'analisi e per l'adozione delle politiche dell'UE dedicate alla Ricerca & Innovazione (R&I).

Infatti, in questa sede si sostiene che, qualora si misurino le *performance* innovative delle regioni europee[1] si debba ricorrere necessariamente ad un approccio metodologico che contempli uno spettro molto ampio ed articolato dei fenomeni e delle attività innovative, utilizzando un sistema di indicatori più significativo ed esauriente rispetto a quello definito e implementato nel RIS.

Del resto, una parte rilevante della letteratura suggerisce di estendere lo studio dei fenomeni e dei processi che contribuiscono a determinare la cosiddetta "performance innovativa" di un Paese, di una regione o di un territorio, poiché essi sono sempre più legati alla dotazione e alla qualità del capitale intellettuale, nonché all'intensità e alla velocità della diffusione della conoscenza nelle interazioni sociali.

Nel corso degli anni, man mano che l'economia mondiale transitava nell'era della Conoscenza e del Digitale, è diventato sempre più importante misurare le performance innovative dei sistemi nazionali e regionali dell'Innovazione, allo scopo di fornire ai *policy maker* precise indicazioni sulle strategie e sugli interventi da adottare e finanziare. Oggi, la misurazione della "performance innovativa" rappresenta uno dei temi di maggiore interesse a livello internazionale e, in particolare, nell'ambito dell'Unione Europea che, attraverso la *Smart Specialization Strategy* (S3), spinge molto sulla capacità degli Stati Membri di potenziare e diversificare le misure a sostegno della R&I per contribuire alla loro "crescita intelligente, sostenibile ed inclusiva".

La specializzazione intelligente implica l'identificazione degli *asset* principali e dei vantaggi competitivi di ogni Paese e di ogni Regione, evidenziando le loro più significative convergenze di competenze (scientifiche, produttive e socio-culturali) al fine di concentrare gli investimenti pubblici e privati negli ambiti di "eccellenza", massimizzando i relativi impatti sul territorio. Senza un quadro conoscitivo dei territori ampio e ricco di dati, capace di monitorare continuativamente l'evoluzione del contesto socio-economico, risulterà più difficile descrivere i processi innovativi

[1]La nomenclatura delle *performance* innovative delle regioni europee, nel linguaggio statistico-normativo della EU, si esprime in NUTS (*Nomenclature des Unités Territoriales Statistiques*), livelli 1, 2 e 3.

e costruire *policy* più efficaci, a partire da quelle azioni mirate ad implementare la S3 nell'EU e nei relativi sistemi regionali.

Infatti, i dati contenuti nel presente lavoro incrociano le esigenze manifestate dai *policy maker* regionali nel dotarsi di uno strumento di controllo che, considerando un ampio spettro di indicatori rappresentativi di attività, processi e fenomeni innovativi, sia adatto a misurare la *performance* innovativa delle regioni italiane e a compararla tra loro, nell'ottica di supportare con maggiore efficacia l'analisi e la definizione delle S3 regionali.

Nella costruzione del CRI, quindi, è stata utilizza una metodologia prettamente quantitativa, il cui sviluppo risulta agevolato dal pieno accesso alle fonti istituzionali sempre aggiornate (e.g.: EUROSTAT; ISTAT), nonché disponibili, consultabili e scaricabili on-line.

Si ritiene che un'approfondita e tempestiva conoscenza della realtà regionale sia di grande utilità per istituzioni, studiosi e operatori che, nell'ambito delle proprie competenze, si occupano di Innovazione.

Il presente lavoro, pertanto, descrive un'attività di elaborazione e diffusione di conoscenze utili sia agli studi sui sistemi innovativi territoriali, sia a supportare l'indirizzo delle politiche regionali per l'Innovazione.

2. L'impianto del CRI

2.1 Il quadro di riferimento

L'impianto del CRI è stato costruito coerentemente con il quadro teorico di riferimento fornito dalla letteratura relativa alle proprietà dei sistemi regionali di innovazione e a quella più specifica che concerne la valutazione delle performance di tali sistemi, ossia dai modelli dell'*European Innovation Scoreboard* (EIS) e dal già citato RIS.

A tal fine, nello specifico, sono state esaminate e utilizzate le seguenti edizioni degli *Scoreboard* diffusi dalla Commissione Europea:

- *Pro Inno Europe - Regional Innovation Scoreboard*, 2009,
- *Innovation Union Scoreboard*, 2014,
- *Regional Innovation Scoreboard* 2014,
- *Innovation Union Scoreboard,* 2015,
- *European Innovation Scoreboard* 2016,
- *Regional Innovation Scoreboard* 2016,

nonché la seguente fondamentale pubblicazione ad esse strettamente collegata:

- H. Hollanders, "*Measuring innovation: the European Innovation Scoreboard*", 2009.

Inoltre, sono stati consultati i seguenti modelli di *Scoreboard* delineati da alcune regioni italiane, nonché un importante documento di analisi del sistema innovativo pugliese:

- *Scoreboard* Regionale dell'Innovazione per la comparazione delle performance del sistema innovativo lombardo – settembre 2006;
- *Scoreboard* Regionale dell'Innovazione per la comparazione delle performance del sistema innovativo piemontese – ottobre 2007;
- Analisi del Sistema Innovativo Regionale Puglia CERPI, 2008;
- *Innovation Scoreboard* Regione Campania, 2012.

Tale quadro teorico di riferimento ha consentito la definizione dell'impianto del CRI. In primo luogo, così come indicato dagli approcci metodologici presenti in letteratura (soprattutto dallo schema di riferimento dell'EIS - che misura le *performances* dei Paesi – relativo all'edizione del 2010 nella quale è stata superata l'impostazione precedente a 18 indicatori. Nel nuovo EIS gli indicatori sono aumentati fino a 25, suddivisi in 3 macro-categorie: elementi abilitanti, attività delle imprese, risultati. I 25 indicatori semplici si suddividono nelle predette tre categorie a seconda della loro natura e funzione interpretativa), nel CRI sono stati individuati i macrofattori e le dimensioni di *input*, di *output* e di processo caratterizzanti la *performance* innovativa oggetto di analisi.

La struttura finale del CRI prevede, quindi, n. 3 "ambiti" che sono relativi a:

1. "*driver* dell'Innovazione";
2. "attività innovativa delle imprese";
3. "risultati dell'Innovazione".

A ciascuno di tali ambiti sono collegate le "dimensioni" caratterizzanti il comportamento del sistema regionale di innovazione. Le dimensioni critiche individuate preliminarmente sono dieci in totale.

Quelle collegate all'ambito "driver dell'Innovazione" sono tre e fanno riferimento a "Risorse Umane", "Finanza per l'innovazione" e "Supporto istituzionale".

Le dimensioni connesse all'ambito "Attività innovativa delle imprese" sono due: "Relazioni di rete"; "Innovatività".

Le dimensioni relative all'ambito "Risultati" sono cinque: "Valorizzazione delle attività innovative"; "Imprese innovatrici"; "Effetti sulla specializzazione produtti-

va"; "Effetti sulla specializzazione dell'Export"; "Sopravvivenza delle Imprese Innovative".

Successivamente sono stati selezionati gli indicatori considerati più idonei per l'analisi. Sulla base della migliore letteratura esistente, il CRI ha preso in considerazione n. 28 indicatori.

Tuttavia, per poter essere selezionati, tali indicatori devono garantire preliminarmente i seguenti criteri:

- l'accessibilità dei dati presso le fonti statistiche istituzionali (ISTAT, EUROSTAT, DPS – Dipartimento per le Politiche di Sviluppo e Coesione, UNIONCAMERE);
- la rilevazione degli stessi dati in tutte le regioni italiane;
- la garanzia di aggiornamento periodico dei dati;
- la rilevanza del fenomeno innovativo misurato da ogni indicatore;
- la rappresentatività dei fenomeni innovativi rispetto alle loro dimensioni.

L'impianto del CRI (Tabelle 1.a, 1.b e 1.c), pertanto, è organizzato sotto forma di matrice ed è articolato per ambiti, dimensioni, indicatori, fonti statistiche, unità di misura e anno di riferimento del dato più aggiornato:

Ovviamente, è bene precisare che ogni indicatore rappresenta solo una parte di una realtà poliedrica, complessa e in continua evoluzione. Infatti, se vengono considerati singolarmente, gli indicatori non sono in grado di rappresentare la performance innovativa dei sistemi territoriali. Pertanto, il CRI promuove un'analisi congiunta di numerosi indicatori che osservano lo stesso fenomeno da più angolature, con l'obiettivo di aumentare la capacità di interpretarlo nella sua complessità (Sirilli, 2000). Infatti, i livelli di *input* e di *output* che quantificano le attività innovative regionali devono essere rilevati facendo ricorso ad un ampio set di indicatori semplici.

L'analisi della *performance* innovativa delle regioni parte dalla misurazione delle variabili indipendenti, ovvero degli indicatori di input. A tale scopo, gli indicatori da cui si preferisce partire sono quelli relativi alla quantificazione del capitale umano e dei livelli di istruzione e culturali dell'unità di analisi considerata, ovvero le regioni. La quantità e la qualità delle risorse umane costituiscono fattori di importanza determinante sia per la creazione di nuova conoscenza, sia per la sua diffusione nel contesto socio-economico, contribuendo a creare meccanismi dinamici e innovativi.

Gli indicatori del primo blocco (descritti nella sezione iniziale della Tabella 1.a) sono suddivisi in due gruppi: nove indicatori dedicati ai fenomeni di istruzione/formazione/cultura e due indicatori (10 e 11) relativi all'occupazione.

Tabella 1.a *Impianto del CRI Ambito: Driver dell'Innovazione.*

Dimensione	*Cod. Ind.*[2]	*Indicatore*	*Fonte*[3]	*Unità di misura*	*Ultimo aggiornam. disponibile*
Risorse Umane	1	Studenti in educazione secondaria e post-secondaria (livelli 3 e 4 ISCED UNESCO 2011), NON terziaria / Popolazione 15-24 anni	EUROSTAT – Database Regional statistics	%	2012
	2	Studenti in educazione terziaria (livelli 5 e 6 ISCED UNESCO 2011) / Popolazione in età 20-24 anni	EUROSTAT – Database Regional statistics	%	2012
	3	Ricercatori in tutti i settori (in ULA) / Popolazione attiva	EUROSTAT – Database Regional statistics	%	2013
	4	Popolazione 20-24 anni con diploma scuola secondaria superiore / Popolazione in età 20-24 anni	ISTAT	%	2015
	5	Popolazione 25-64 anni con titolo di studio secondario e post secondario (livelli 3 e 4 ISCED UNESCO 2011), NON terziario / Popolazione in età 25-64 anni	EUROSTAT – Database Regional statistics	%	2015
	6	Popolazione 25-64 anni con titolo di studio terziario (livelli da 5 a 8 ISCED UNESCO 2011) / Popolazione in età 25-64 anni	EUROSTAT – Database Regional statistics	%	2015
	7	Popolazione 30-34 anni con titolo di studio terziario (livelli da 5 a 8 ISCED UNESCO 2011) / Popolazione in età 30-34 anni	EUROSTAT – Database Regional statistics	%	2015
	8	Laureati in discipline scientifiche e tecnologiche / Popolazione in età 20-29 anni	ISTAT	‰ abitanti	2012
	9	Adulti 25-64 anni che frequentano un corso di studio o di formazione professionale / Popolazione in età 25-64 anni	ISTAT	%	2015
	10	Addetti in attività scientifiche e tecnologiche / Popolazione attiva	EUROSTAT – Database Regional statistics	%	2015
	11	Persone con titolo di studio terziario (livelli da 5 a 8 ISCED UNESCO 2011) e occupate in attività scientifiche e tecnologiche / Popolazione attiva	EUROSTAT – Database Regional statistics	%	2015
Finanza per l'Innovazione	12	Investimenti in capitale di rischio early stage / PIL regionale (a prezzi correnti)	ISTAT	‰	2014
	13	Investimenti in capitale di rischio expansion e replacement / PIL regionale (a prezzi correnti)	ISTAT	‰	2014
	14	Investimenti fissi lordi / PIL regionale (ai prezzi dell'anno precedente)	ISTAT	%	2013
Supporto istituzionale	15	Incidenza (Spesa pubblica in R&S - Spesa per R&S intramuros di Università, EPR e PA regionali)/PIL regionale (a prezzi corr.)	ISTAT	%	2012
	16	Incidenza della spesa pubblica in R&I - Spesa Fondi Strutturali in R&I / Spesa Fondi Strutturali totale	DPS - Banca dati di Open Coesione	%	30.04.2016

Fonte: Data retrieval a cura di S. Marastoni su dati Istat e Eurostat.

[2] Per "Cod. Ind." si intende "Codice Identificativo dell'Indicatore".

[3] Gli indicatori di Fonte ISTAT, riportati nelle tabelle 1.a, 1.b e 1.c, sono estratti dalla Banca dati di indicatori territoriali per le politiche di sviluppo.

Il secondo blocco è composto da tre indicatori (12, 13 e 14 Tab. 1.a) diretti a quantificare gli *input* collegati al finanziamento privato dell'innovazione in termini sia di apporto del capitale di rischio al sistema innovativo, sia del livello degli investimenti fissi in quanto fenomeno che condiziona significativamente la capacità innovativa dei territori osservati:

Un altro mini-blocco, formato da due indicatori (15 e 16 Tab. 1.a) è quello relativo alla quantificazione degli *input* che rilevano il supporto istituzionale all'Innovazione, attribuendo un peso adeguato alle attività di R&I stimolate ed avviate grazie all'intervento pubblico:

Il quarto blocco, in realtà, coincide con il solo indicatore 17 della Tabella 1.b ed è volto a quantificare l'*input* che misura le relazioni di rete fra imprese, poiché in quei contesti in cui si registra una maggiore intensificazione e concentrazione di relazioni tra attori economici è più probabile che le dinamiche di apprendimento collettivo risultino più efficaci, incidendo positivamente sulla propensione dei sistemi territoriali osservati ad introdurre innovazioni:

Il quinto blocco è composto dai quattro ultimi indicatori della Tab. 1.b, relativi alla quantificazione degli *input* che rilevano la capacità innovativa delle imprese, soprattutto la capacità specifica di creare nuova conoscenza anche mediante l'utilizzo di infrastrutture digitali:

Tabella 1.b *Impianto del CRI Ambito: Attività Innovativa delle Imprese.*

Dimensione	*Cod Ind.*	*Indicatore*	*Fonte*	*Unità di misura*	*Ultimo aggiornam. disp.*
Relazioni di rete	17	Numero di Imprese (con almeno 10 addetti) dei settori industria e servizi che hanno definito accordi di cooperazione per l'Innovazione / Numero totale di imprese attive (con almeno 10 addetti) nei settori industria e servizi	EUROSTAT – CIS	%	2012
Innovatività	18	Indice di diffusione della banda larga nelle imprese - Numero di imprese (con più di 10 addetti) dei settori industria e servizi che dispongono di un collegamento a banda larga / Numero di imprese attive (con più di 10 addetti) dei settori industria e servizi	ISTAT	%	2015
	19	Numero di Imprese (con almeno 10 addetti) dei settori industria e servizi con attività innovative / Numero totale di imprese attive (con almeno 10 addetti) nei settori industria e servizi	EUROSTAT – CIS	%	2012
	20	Incidenza della spesa delle Imprese in R&S 1- Spesa delle imprese pubbliche e private per R&S / PIL regionale (a prezzi correnti)	ISTAT	%	2013
	21	Incidenza della spesa delle Imprese in R&S 2 - Spesa per Innovazione delle Imprese attive (con almeno 10 addetti) dei settori industria e servizi / Totale addetti delle Imprese attive (con almeno 10 addetti) dei settori industria e servizi.	EUROSTAT – CIS	Spesa media per addetto in migliaia di euro correnti	2012

Fonte: Data retrieval a cura di S. Marastoni su dati Istat e Eurostat.

Tabella 1.c *Impianto del CRI Ambito: Risultati dell'Innovazione.*

Dimensione	***Cod. Ind.***	***Indicatore***	***Fonte***	***Unità di misura***	***Ultimo aggiornamento disponibile***
Valorizzazione delle attività innovative	22	Numero di brevetti depositati presso l'EPO / Totale popolazione attiva	EUROSTAT – Database Regional statistics	N. di brevetti per milioni di unità di popolazione attiva	2012
	23	Numero di marchi registrati presso l'EUTM / Totale popolazione attiva	EUROSTAT – Database Regional statistics	N. di marchi per milioni di unità di popolazione attiva	2015
Imprese Innovatrici	24	Numero di Imprese (con almeno 10 addetti) dei settori industria e servizi con attività innovative di prodotto-processo / Numero totale di imprese attive (con almeno 10 addetti) nei settori industria e servizi	EUROSTAT – CIS	%	2012
Specializzazione produttiva	25	Numero degli addetti nei settori manifatturieri a medio-alta e alta tecnologia / Totale occupati	EUROSTAT – Database Regional statistics	%	2015
	26	Numero degli addetti nei settori manifatturieri ad alta tecnologia e nei servizi ad elevata intensità di conoscenza / Totale occupati	EUROSTAT – Database Regional statistics	%	2015
Specializzazione dell'export	27	Esportazioni di prodotti specializzati e high-tech secondo la tassonomia di Pavitt / Totale esportazioni	Unioncamere - Banca dati STARNET	%	2014
Sopravvivenza delle Imprese innovative	28	Numero di imprese ad alta intensità di conoscenza sopravissute a tre anni / Totale imprese ad alta intensità di conoscenza	ISTAT	%	2013

Fonte: Data retrieval a cura di S. Marastoni su dati Istat e Eurostat.

Se in letteratura si riscontra una generale conformità nell'individuazione degli indicatori di *input*, la situazione appare molto più controversa per l'esplicitazione degli indicatori di *output*. Inoltre, si rileva che gli studiosi preferiscono non utilizzare un'ampia varietà di indicatori finalizzati a quantificare i livelli di *output* delle attività innovative. Ciò che emerge con evidenza è il grande utilizzo dei dati sulla capacità brevettuale, poiché essa è reputata come la variabile dipendente per eccellenza, ovvero come l'*output* delle attività innovative più considerato e più apprezzato in letteratura.

Nel CRI, invece, oltre all'indicatore centrato sui brevetti, si propone un mix di altri sei indicatori di *output* volti ad illustrare una componente essenziale della *performance* innovativa delle regioni.

Pertanto, il blocco finale è dedicato ai sette indicatori diretti a quantificare altrettanti livelli di *output* delle attività innovative, partendo proprio dall'andamento del fenomeno brevettuale (indicatore 22 della Tabella 1.c) che può essere conside-

rato come il più importante *output* codificabile, perché in grado di misurare la valorizzazione economica dell'attività inventiva attraverso la protezione legale della proprietà intellettuale:

In questa sede si propone anche l'utilizzo dell'indicatore relativo all'andamento di un'altra tipologia di proprietà intellettuale: i marchi (indicatore 23 della Tabella 1.c). Questo perché i marchi categorizzano e fissano l'innovazione immateriale e spesso rappresentano un fattore chiave nella strategia innovativa, specie quella legata all'innovazione commerciale e di mercato:

Tra gli *output* codificabili della capacità innovativa delle imprese si devono senz'altro considerare quelli relativi alla quantità di nuovi prodotti immessi sul mercato. Pertanto, in questa sede si propone di considerare un indicatore relativo alla consistenza numerica delle imprese innovatrici in termini di prodotto/processo (indicatore 24 della Tabella 1.c):

Per quanto riguarda gli indicatori relativi al fenomeno della specializzazione produttiva delle imprese (indicatori 25 e 26 della Tabella 1.c), questi possono senz'altro essere inquadrati come un *output* specifico (anche se non codificabile) dei processi di apprendimento continuo:

Tra gli output della capacità innovativa di un sistema produttivo si devono anche considerare quelli relativi alle esportazioni dei prodotti innovativi. Pertanto, in questa sede si propone di trattare un indicatore relativo alle esportazioni di prodotti specializzati e high-tech, secondo la tassonomia di Pavitt[4], rapportate al totale delle esportazioni (indicatore 27):

Infine, in questa sede si propone di utilizzare il tasso di sopravvivenza delle imprese innovative nei primi tre anni di vita (indicatore 28 della Tabella 1.c) come un *output* (non codificabile) diretto a misurare l'efficacia del fenomeno degli *spillover*.

L'analisi statistica, di cui alcuni dettagli sono riportati in appendice, ha poi consentito di confermare le scelte effettuate, di verificare l'esistenza di relazioni tra indicatori semplici e le dimensioni cui questi sono associati, di minimizzare la ridondanza di informazioni.

Al termine della fase di estrazione, raccolta ed elaborazione dei dati, gli stessi sono stati inseriti nella matrice del CRI (Cfr. Tabella A.1 posta in Appendice).

Si segnala che nel processo di costruzione del CRI sono stati elaborati e trattati sensibilmente i dati relativi a due indicatori la cui misurazione è necessariamente "derivata":

[4] Per avere indicazioni sul contenuto tecnologico dei beni commercializzati, Pavitt ha classificato i prodotti esportati, raggruppandoli in tre classi distinte (1. agricoltura e materie prime; 2. prodotti tradizionali e standard; 3. prodotti specializzati e high tech).

- Indicatore n. 16; Spesa Fondi Strutturali in R&I / Spesa Fondi Strutturali totale;
- Indicatore n. 27: Esportazioni di prodotti specializzati e high-tech secondo la tassonomia di Pavitt / Totale esportazioni.

Infatti, per quanto riguarda l'indicatore 16, i dati relativi alla Spesa (sia in R&I, sia totale) dei Fondi Strutturali sono stati ricostruiti per tutte le regioni italiane, calcolando la somma della spesa certificata di ciascun intervento finanziato con i predetti Fondi e censito nella Banca Dati di Open Coesione,

Inoltre, con riferimento all'indicatore 27, i valori delle esportazioni sia totali, sia riferite ai soli prodotti specializzati e high tech (secondo la tassonomia di Pavitt) sono stati calcolati sommando i dati provinciali disponibili presso la Banca Dati di STARNET (Unioncamere) per tutte le regioni italiane.

2.2 Analisi di benchmarking

Il CRI consente di valutare i diversi elementi alla base della performance innovativa delle regioni italiane e di stimare un indice complessivo di essa, anche nell'ottica della comparazione tra le regioni stesse.

I dati sono stati raccolti ed elaborati non solo per le regioni italiane, ma anche per le ripartizioni geografiche di "Nord", "Centro", "Sud", "Isole" e "Italia", al fine di effettuare le opportune analisi di *benchmarking* (poiché tali ripartizioni risultano fondamentali per il confronto dei loro valori medi con quelli delle singole regioni).

Di seguito si elencano le unità territoriali per le quali sono stati raccolti i dati:

Tabella 2. *Unità territoriali considerate nella costruzione del CRI*

Unità territoriale	***Unità territoriale***
Piemonte	Abruzzo
Valle D'Aosta / Vallée D'Aoste	Molise
Liguria	Campania
Lombardia	Puglia
Provincia Autonoma di Bolzano / Bozen	Basilicata
Provincia Autonoma di Trento	Calabria
Veneto	Sicilia
Friuli-Venezia Giulia	Sardegna
Emilia-Romagna	*Nord*
Toscana	*Centro*
Umbria	*Sud*
Marche	*Isole*
Lazio	*Italia*

2.3 Punti di debolezza e di forza del CRI

Per quanto riguarda i punti di debolezza:

- i dati presenti nel CRI non sono omogenei rispetto al momento temporale della rilevazione e, pertanto, l'analisi della performance innovativa deve riferirsi al quadro più aggiornato disponibile presso le fonti statistiche istituzionali, relativamente all'intervallo temporale 2012-2016.

  Nello specifico, gli indicatori sono aggiornati:
  - n. 9 al 2012;
  - n. 4 al 2013;
  - n. 3 al 2014;
  - n. 11 al 2015;
  - n. 1 al 2016.

  Comunque, se si considera che nel RIS 2016 i dati più recenti fanno riferimento al 2014 per due indicatori, al 2013 per tre indicatori, al 2012 per sei indicatori e al 2011 per un indicatore, il CRI risulta significativamente più aggiornato rispetto al RIS stesso. È importante evidenziare che l'aggiornamento degli *Scoreboard* risente dei tempi tecnici con cui le fonti statistiche ufficiali rilevano e pubblicano i dati di riferimento. Pertanto, sia il CRI, sia il RIS e, in generale, tutti gli altri strumenti analoghi, non riescono a cogliere le dinamiche più recenti dei fenomeni innovativi;
- sono stati gestiti i valori mancanti[5] di alcuni indicatori elementari per alcune regioni e/o ripartizioni territoriali, sostituendoli con la media calcolata sull'indicatore stesso.

  Tuttavia, si deve evidenziare che anche nel RIS 2016, con riferimento ai 12 indicatori considerati, ben il 24,5% dei dati non è disponibile per le 214 regioni europee nel periodo di osservazione e di comparazione diacronica considerato nella sua metodologia, ovvero negli ultimi cinque anni. Inoltre, molti dati regionali rilevati attraverso la CIS (la rilevazione degli istituti statistici

[5] Risultano mancanti i seguenti valori degli indicatori elementari: 12. *Investimenti in capitale di rischio early stage / PIL regionale (a prezzi correnti)* per la Valle d'Aosta, il Molise e le due Province Autonome di Bolzano e di Trento; 13. *Investimenti in capitale di rischio expansion e replacement / PIL regionale (a prezzi correnti)* per la Valle d'Aosta e le due Province Autonome di Bolzano e di Trento; 15. *Incidenza della spesa pubblica in R&S - Spesa per R&S intramuros di Università, EPR e PA regionali / PIL regionale (a prezzi correnti)* e 17. *Numero di Imprese (con almeno 10 addetti) dei settori industria e servizi che hanno definito accordi di cooperazione per l'Innovazione / Numero totale di imprese attive (con almeno 10 addetti) nei settori industria e servizi* per il Molise; 25. *Numero degli addetti nei settori manifatturieri a medio-alta e alta tecnologia / Totale occupati* per la Valle d'Aosta; 26. *Numero degli addetti nei settori manifatturieri ad alta tecnologia e nei servizi ad elevata intensità di conoscenza / Totale occupati* per la Valle d'Aosta e il Molise.

nazionali sull'Innovazione nelle Imprese è coordinata a livello europeo con la *Community Innovation Survey* – CIS. Essa raccoglie informazioni sulle attività innovative svolte dalle imprese che operano nei settori economici dell'industria e dei servizi) risultano scarsi o addirittura non disponibili per diversi anni o addirittura per l'intero periodo considerato.

È anche importante sottolineare che, al fine di completare la griglia dei dati da elaborare per la misurazione delle performance, il RIS 2016 (ma anche quelli delle edizioni precedenti) utilizza una tecnica statistica di stima e imputazione dei dati mancanti chiamata "regionalizzazione. In altre parole, ogni volta che i dati afferenti una rilevazione della CIS risultano non disponibili a livello regionale, ma disponibili in forma aggregata a livello nazionale, viene implementata una tecnica che, in merito agli indicatori sulla forza lavoro e sulle imprese, presuppone la stessa intensità settoriale tra i Paesi e le regioni ad essi appartenenti. Si tratta di una tecnica statistica senz'altro consentita, ma che rende l'impianto complessivo del RIS abbastanza critico.

In merito ai punti di forza, invece, il CRI:

- mostra e, quindi, rende trasparenti i dati rilevati per ciascuno dei 28 indicatori, mentre nel RIS e negli altri *Scoreboard* regionali vengono illustrati solo i valori indice;
- presenta e rende visibile nell'analisi di *benchmarking*, diversamente da quanto si può riscontare nel RIS e negli altri *Scoreboard* regionali, il confronto con i valori medi delle aggregazioni territoriali di riferimento, quali "Nord", "Centro", "Sud", "Isole" e "Italia", che risulta fondamentale per l'analisi del posizionamento dei vari sistemi regionali rispetto ai contesti territoriali su cui vengono attuate delle specifiche misure e *policy* per l'innovazione, originate dalle stesse fonti di finanziamento (Fondi Strutturali);
- prende in considerazione 28 indicatori contro i 12 del RIS 2016. Tale scelta consente di analizzare in maniera più ampia e articolata i fenomeni innovativi, anche in considerazione della loro rilevanza economico-sociale e degli impatti che essi determinano sui processi di sviluppo del territorio regionale;
- costruisce e calcola un Indice Sintetico Ponderato Generale di Innovazione delle regioni italiane (ISPGI) e delle relative ripartizioni territoriali, mentre nel RIS e in tutti gli altri *Scoreboard* esistenti in letteratura, l'indice di Innovazione Generale delle regioni viene calcolato come semplice media aritmetica tra i valori indice.

3. La metodologia statistica di misurazione della performance innovativa delle regioni

Completata la fase del *data retrieval* e definito il *dataset*, si è proceduto a svincolare gli indicatori elementari dall'unità di misura con cui essi sono stati espressi, attraverso un processo di standardizzazione (o normalizzazione) degli stessi. In altri termini, per i dati di ciascuno dei 28 indicatori è stato calcolato un valore normalizzato che misura il fenomeno sotteso. Tale valore è stato determinato su base 10, cioè, assegnando alla regione e/o alla ripartizione territoriale più performante il valore 10. Successivamente, sul dato statistico relativo alle altre regioni/ripartizioni territoriali, è stata eseguita una attribuzione proporzionale dei valori indice, appunto, su base 10: ogni indice, dunque, varia tra 0 e 10, qualunque sia il suo originale campo di variazione (cfr. Tabella A.2 posta in appendice)[6].

Dopo aver concluso l'operazione di standardizzazione dei dati, è stato avviato il procedimento di ponderazione degli indicatori elementari. Innanzi tutto è stata individuata la variabile da correlare con i valori indice di ciascun indicatore elementare. La scelta della variabile è caduta su un fondamentale dell'economia regionale (che è senz'altro condizionato dai processi di innovazione), ossia la produttività del lavoro, determinata sulla base del rapporto tra PIL regionale (a prezzi correnti) e numero di occupati. Inoltre, tale variabile costituisce una misura fondamentale della capacità competitiva dei sistemi economici.

Conseguentemente, è stata costruita, per ogni regione/ripartizione territoriale, la predetta variabile. Questa è stata standardizzata rispetto al valore massimo per liberarsi dell'unità di misura e rendere il fenomeno confrontabile con i valori indice disponibili, anch'essi standardizzati rispetto al valore massimo.

Quindi, è stato quantificato il livello di correlazione dei valori indice di ciascun indicatore semplice con la produttività pro-capite, utilizzando il coefficiente di correlazione non parametrica di *Spearman* (cfr. Tabella A.3 di Appendice).

La scelta è ricaduta su tale strumento non parametrico in quanto, a parità di risultati, esso risulta più potente del coefficiente di *Pearson* quando viene violato il requisito della normalità delle variabili: nel caso presente, infatti, è stato verificato

[6]Anche nel RIS 2016 viene effettuata un'operazione di normalizzazione dei dati trasformandoli opportunamente al fine di ridurre l'asimmetria positiva registrata sui dati grezzi
Per tutti gli indicatori i dati grezzi sono stati trasformati calcolandone la radice di indice *k* allorquando, per la distribuzione degli stessi, è stato registrato un indice di *Skewness* maggiore di 1, in modo da ottenere un indice minore di 1 dopo tale procedimento. L'indice di *Skewness*, infatti, misura l'asimmetria di una distribuzione: quando è pari a zero, la distribuzione è simmetrica (non sempre); invece, per valori positivi/ negativi la distribuzione è, rispettivamente, asimmetrica positiva/negativa.

che, tramite il test di *Shapiro-Wilk,* nessuno degli indicatori elementari si distribuisce normalmente (cfr. Tabella A.3bis posta in appendice).

Inoltre, mentre la correlazione parametrica di *Pearson* valuta una relazione biunivoca che deve essere forzatamente di tipo lineare, quella non parametrica stima la monotonicità che, essendo una condizione meno stringente, si realizza sempre in presenza di una relazione lineare e molto spesso anche in presenza di relazioni non lineari (ad eccezione delle correlazioni paraboliche e simili); infatti, se due variabili sono legate da una relazione monotòna, i loro ranghi avranno una relazione di tipo lineare.

Nel passaggio successivo è stato pesato ogni valore indice relativo a ciascun indicatore elementare con la misura della predetta correlazione (cfr. la Tabella A.4 in appendice).

3.1 L'Indice Sintetico Ponderato di Innovazione Generale delle regioni (ISPGI)

Infine, è stato costruito per ogni regione / ripartizione territoriale l'indice sintetico ponderato, inteso come rapporto tra la somma di tutti gli indicatori elementari pesati con i rispettivi coefficienti di cograduazione di Spearman e la somma di tutti i coefficienti di cograduazione. Formalizzando si ha:

$$QI = \frac{\sum w_i \mathrm{I}_i}{\sum w_i}$$

dove w_i rappresenta gli indici di cograduazione di Spearman e I_i i valori di ogni singolo indice elementare standardizzato. Tale indice di sintesi assume valori compresi tra zero e dieci[7].

4. Risultati della misurazione e comparazione tra CRI e RIS

4.1 Il raffronto tra il ranking CRI VS RIS

Sia il CRI che il RIS si propongono come strumenti di misurazione delle performance innovative delle regioni italiane. Applicando la metodologia statistica di misurazione messa a punto nel CRI, si è definito il ranking relativo all'indice Sintetico Ponderato di Innovazione Generale delle regioni italiane e lo si è raffrontato con quello determinato dal RIS 2016.

[7] Il campo di variazione dell'indice è il risultato del processo di trattamento dei dati grezzi e della costruzione dell'indice stesso (paragrafo 3).

Come ampiamente descritto nel seguito, i due strumenti mostrano delle analogie e alcune differenze rilevanti, sia dal punto di vista metodologico così come nei risultati.

Innanzi tutto, nella Tab. 3 e, rispettivamente, nella Tab. 4 sono riportati i due ranking citati.

Tabella 3. *Ranking CRI, ordinato per indice di Innovazione Generale*

Ripartizione territoriale	***ISPGI***[8]	***Rango***
Emilia Romagna	7,58	1
Friuli Venezia Giulia	7,42	2
Lombardia	7,26	3
Piemonte	7,11	4
Provincia autonoma di Trento	6,80	5
Lazio	6,78	6
Toscana	6,49	7
Veneto	6,41	8
Liguria	6,39	9
Provincia autonoma di Bolzano	6,30	10
Marche	5,81	11
Abruzzo	5,66	12
Umbria	5,62	13
Valle d'Aosta	5,29	14
Molise	5,03	15
Campania	4,99	16
Basilicata	4,76	17
Puglia	4,74	18
Sardegna	4,51	19
Calabria	4,40	20
Sicilia	4,32	21
Nord	6,62	6,88
Centro	6,46	7,60
Sud	4,93	16,33
Isole	4,42	20,00
Italia	5,89	

Fonte: Elaborazioni a cura di R. Mancarella su dati Istat e Eurostat.

Preliminarmente, si è ritenuto opportuno verificare e quantificare l'eventuale correlazione tra i due indicatori. I risultati di tale analisi inducono a confermare le prime considerazioni descrittive, ovvero i due indicatori mostrano delle logiche analogie ma, di fatto, esistono delle innegabili differenze.

[8] Con l'acronimo ISPGI si intende l'Indice Sintetico Ponderato di Innovazione Generale del CRI.

Tabella 4. *Ranking RIS, ordinato per indice di Innovazione Generale*

Ripartizione territoriale	***ISGI***[9]	***Rango***
Friuli Venezia Giulia	0,46	1
Piemonte	0,45	2
Veneto	0,43	3
Lazio	0,42	4
Lombardia	0,42	5
Emilia Romagna	0,42	6
Provincia autonoma di Trento	0,40	7
Toscana	0,38	8
Liguria	0,37	9
Umbria	0,37	10
Abruzzo	0,36	11
Provincia autonoma di Bolzano	0,35	12
Molise	0,34	13
Marche	0,32	14
Basilicata	0,32	15
Campania	0,32	16
Puglia	0,32	17
Sicilia	0,30	18
Valle d'Aosta	0,30	19
Calabria	0,26	20
Sardegna	0,24	21
Nord[10]	0,40	7,25
Centro	0,38	8,40
Sud	0,32	15,33
Isole	0,27	19,50
Italia	0,36	

Fonte: Data retrieval a cura di S. Marastoni.

Infatti, utilizzando entrambi i coefficienti *r* di *Pearson* e *ρ* di *Spearman*[11], si è registrata una correlazione tra l'ISPGI e l'ISGI molto alta in entrambi i casi ma non prossimi al valore massimo[12], consentendo dunque di escludere la perfetta equivalenza dei due indicatori, confermando però la loro similarità di comportamento.

[9] Con l'acronimo ISGI si intende l'Indice Sintetico di Innovazione Generale del RIS.

[10] I valori dell'ISGI relativi alle ripartizioni territoriali Nord, Centro, Sud e Isole sono stati calcolati dagli autori.

[11] Il primo è il più noto coefficiente di correlazione lineare, parametrico, che non tiene conto della scala di misura degli indicatori (*Scale free*). Il secondo è un coefficiente di correlazione non parametrico, con assunzioni meno stringenti rispetto al primo*:* non tiene conto, infatti, né della scala di misura degli indicatori né della loro distribuzione, né della forma dell'eventuale relazione (ed è quindi adatto anche per lo studio di relazioni non lineari): esso è dunque sia *Scale free* che *Distribution free.*

[12] Ricordando che il campo di variazione per entrambi i coefficienti è [-1; +1], si è registrato un valore pari a 0.908 per il coefficiente di *Pearson* e pari a 0.911 per il coefficiente di *Spearman.*

L'approfondimento analitico relativo al raffronto tra i due strumenti di misurazione, consente di rappresentare le seguenti evidenze.

4.2 Confronto analitico per ranghi

Dal confronto tra ranghi, per i due indicatori si possono osservare analogie, sensibili differenze e importanti scostamenti nel posizionamento di alcune regioni italiane.

Le differenze più consistenti si osservano (Tabelle 3 e 4) per:

- l'Emilia Romagna: primo posto ISPGI *vs* sesto ISGI;
- Il Piemonte: quarto posto ISPGI *vs* secondo ISGI;
- Lazio: sesto posto ISPGI *vs* quarto ISGI;
- Veneto: ottavo posto ISPGI *vs* terzo ISGI;
- Provincia Autonoma di Bolzano: decimo posto ISPGI *vs* dodicesimo ISGI;
- Marche: undicesimo posto ISPGI *vs* quattordicesimo ISGI;
- Valle d'Aosta: quattordicesimo posto ISPGI *vs* diciannovesimo ISGI.

4.3 Capacità discriminante degli indicatori

L'analisi prosegue con la valutazione della capacità discriminante dei due diversi indici di Innovazione Generale, utilizzando il criterio della coerenza interna tramite un test proposto da R. *Likert* nel contesto della teoria della misura[13].

Il test della differenza interquartile di *Likert*[14,] di per sé, è indipendente dalla scala degli indicatori, ma la loro descrizione risulta più chiara se gli indicatori sono espressi nelle stesse unità di misura. Dunque è opportuno procedere alla normalizzazione dei due indicatori.

Sono stati testati tre differenti tipi di normalizzazione (cfr. Tabella A.5 in appendice):

- Campo di variazione teorico: i valori vengono relativizzati rispetto al massimo teorico (per esempio 10 per il ISPGI);
- Campo di variazione empirico: i valori vengono relativizzati rispetto al valore massimo registrato;
- La classica normalizzazione Z tramite scarti standardizzati[15].

[13]In realtà tale criterio viene qui utilizzato per analogia: pur non essendo giudizi valutativi, gli indicatori ISPGI e ISGI esprimono di per sé una valutazione, e presentano adeguate caratteristiche: rendono giustizia delle differenze tra regioni (sensibilità), possono classificare anche quelle regioni che non presentano tutte le caratteristiche studiate (specificità), etc.

[14]Cfr., ad es., Delvecchio (1995), pp 44-45.

[15]Si tenga conto che con la standardizzazione classica $\mathbf{z=(x-\mu)/\sigma}$ si perde l'informazione relativa alla differenza di variabilità complessiva dei due indicatori: infatti la variabile risultante ha $\mu=0$; $\sigma=1$.

Dai risultati di una prima analisi descrittiva degli indicatori standardizzati, l'ISPGI sembrerebbe esibire una maggiore capacità discriminante rispetto all' ISGI.

In effetti per il ISPGI sono stati registrati dei valori caratteristici maggiori rispetto all' ISGI (si rimanda Tabella A.6 in appendice):

- Campo di variazione (0,33 ISPGI vs 0,21 ISGI);
- Coefficiente di variazione (1,82; 20,91 ISPGI vs 0,98; 13,72 ISGI)
- Differenza interquartilica (0,19; 0,59 ISPGI vs 0,10; 0,46 ISGI)

Il criterio della coerenza interna di *Likert*, invece, considera la media del quartile con valori più alti (primo quartile) e la media del quartile con valori più bassi (terzo quartile): quanto più è elevata la differenza tra queste due medie (scarto interquartile) tanto maggiore è il potere discriminante dell'indicatore.

Come si evince dall'analisi della Tabella A.6 (posta in appendice), l'ISPGI esibisce una maggiore capacità discriminante rispetto all' ISGI (scarto interquartile: ISPGI 0,27; 0,82 vs 0,15; 0,72 ISGI)

Per valutare la significatività dello scarto interquartile *Likert* propose il test:

$$Z = \frac{(\mu_4 - \mu_1)}{\sqrt{\left(\frac{\hat{\sigma}_4^2}{n/4}\right) + \left(\frac{\hat{\sigma}_1^2}{n/4}\right)}}$$

in cui μ_4 e μ_1 sono, rispettivamente, la media del quartile alto e basso, mentre $\hat{\sigma}_4^2$ e $\hat{\sigma}_1^2$ ne rappresentano le rispettive varianze.

Il test di *Likert* si distribuisce all'incirca come una normale standardizzata (valore soglia z=1,96) e assicura un buon livello di coerenza e quindi di capacità discriminante per valori superiori a 2.

Come illustrato in Tabella A6 (posta in appendice) entrambi gli indicatori risultano significativi (ISPGI 16,74 vs ISGI 9,91), ma al ridursi del valore empirico del test si incrementa la probabilità di errore (p-value): dunque l'ISPGI presenta una probabilità di errore minore dell'ISGI.

A questo risultato si potrebbero muovere due critiche:

- non sarebbe corretto, a rigor di logica matematica, usare le varianze considerate per dati non metrici; tuttavia, data la composizione dei due indicatori (determinata come sintesi di indici in massima parte metrici) l'approssima–zione sembra soddisfacente;
- presumibilmente l'ISPGI riesce ad essere più discriminante rispetto al ISGI perché costruito come sintesi di una maggior quantità di indicatori elemen-

tari; invero, gli indicatori elementari devono essere rappresentativi[16] ovvero devono essere in grado di rappresentare il fenomeno d'interesse in tutte le sue componenti. Naturalmente, quanto più vasto e complesso è il fenomeno da fotografare, tanto più numerosi dovrebbero essere gli indicatori. Considerata la complessità del fenomeno indagato e l'immediata disponibilità dei dati di fonte amministrativa, per costruire l'ISPGI è sembrato opportuno non focalizzarsi sulla parsimonia dell'indicatore, pur prestando comunque molta attenzione a non inserire indicatori elementari ridondanti.

5. Conclusioni

È opportuno sottolineare che l'impianto del CRI, pur essendo frutto di un intenso lavoro di ricerca e di elaborazione, deve considerarsi come un punto di partenza nell'analisi del sistema di innovazione regionale che deve arricchirsi di nuovi indicatori e nuove elaborazioni derivanti dall'evoluzione degli ecosistemi territoriali. Ulteriori studi e approfondimenti potranno in futuro condurre ad una revisione di alcuni degli elementi dell'impianto che in questa sede vengono presentati.

Perché preferire il CRI al RIS?

1. L'Italia è una realtà molto eterogenea a livello economico e sociale, immagine che entrambi gli strumenti di misurazione riescono a fotografare con un certo dettaglio. Tuttavia, se si ipotizza che in futuro il divario socio-economico diminuisca, è lecito presumere che l'eterogeneità osservata tra essi si riduca in misura all'incirca proporzionale per l'ISPGI e per l'ISGI. Il test di *Likert* indica, infatti, che l'ISGI, meno sensibile, cesserà di essere significativamente discriminante molto prima dell'ISPGI, ovvero al ridursi dell'eterogeneità delle regioni, il test interquartile condotto sull'ISPGI continuerà ad essere statisticamente significativo anche quando l'ISGI abbia perso di significatività. In caso di incremento del divario, invece, la stessa eterogeneità tra strumenti di misurazione potrebbe acutizzarsi.
2. La metodologia di misurazione del CRI è l'unica, tra quelle attualmente esistenti, ad aver introdotto ed implementato un sistema di ponderazione degli indicatori elementari, finalizzato alla costruzione dell'ISPGI. Il calcolo di tale indice rappresenta, secondo gli autori, il vero e più pregiato valore aggiunto del CRI. Si consideri che nel RIS 2016 (e in tutte le edizioni precedenti), l'indice di innovazione regionale viene calcolato come media non ponderata

[16] Delvecchio (1995), pp 87-88.

dei punteggi normalizzati dei 12 indicatori (adottando, quindi, una semplice media aritmetica tra i valori indice dei 12 indicatori considerati).

3. Il CRI è uno strumento estendibile a tutte le regioni europee, nonché a tutti gli *Innovation Regional Systems* (IRS) del mondo, purché siano disponibili gli stessi dati in maniera omogenea per i n. 28 indicatori considerati.
4. La struttura del CRI consente facilmente di studiare i sistemi innovativi regionali sia per "ambito", sia per "dimensione", contribuendo alla disamina analitica delle diverse componenti che determinano la performance innovativa dei territori.
5. È anche possibile disaggregare l'analisi per gli anni di rilevazione, considerando solo quelli in cui è presente la maggior parte dei dati afferenti gli indicatori considerati nel CRI.
6. L'impianto del CRI permette di valutare analiticamente e comparativamente sia i fenomeni innovativi più performanti, sia quelli meno performanti degli IRS.
7. Con le future rilevazioni periodiche annuali, il CRI potrà arricchirsi di quelle serie storiche che consentiranno un'analisi evolutiva degli IRS, ovvero lo studio anche comparato delle loro traiettorie nel tempo. Tale analisi offrirà un concreto e prezioso aiuto ai *policy maker* che quotidianamente si trovano a gestire il ciclo della programmazione regionale.

Riconoscimenti

Il presente lavoro descrive una delle molteplici attività che l'Agenzia Regionale per la Tecnologia e l'Innovazione – ARTI – della Puglia sta realizzando al fine di supportare i decisori pubblici regionali nella definizione e nella manutenzione evolutiva della cosiddetta *Smart Specialization Strategy* (S3) della Regione Puglia.

Bibliografia di riferimento

Acs, Z. J.; Anselin, L.; Varga, A. (2002). Patents and Innovation Counts as Measures of Regional Production of New Knowledge, *Research Policy*, 31: 1069-1085.

Asheim, B. T.; Gertler, M. S. (2005). The Geography of Innovation: Regional Innovation Systems. In: Fagerberg J., Mowery D. C. and Nelson R. R.(eds.) *The Oxford Handbook of Innovation*. Oxford: Oxford University Press, pp. 291-317.

Baù, M.; Cagnina, M.R.; Chiarvesio, M.; Compagno, C.; Fornasier, E.; Lauto, G.; Mazzurana, P.A.M.; Pittino, D.; Visintin, F. (2011). *Misurare le performance innovative di un sistema regionale*, Franco Angeli ed., collana Economia - Ricerche.

Brenner, T.; Greif, S. (2006). The dependence of innovativeness on the local firm population —an empirical study of German patents, *Industry and Innovation*, Zitiert von: 63 - Ähnliche Artikel - Alle 13 Versionen.

Browne, M. W. (1982). Covariance structures. In: D.M. Hawkins (ed.), *Topics in applied multivariate analysis*. Cambridge University Press: 72-141.

Buesa, M.; Heijs, J.; Martinez Pellittero, M.; Baumert, T. (2006). Regional Systems of Innovation and the Knowledge Production Function: the Spanish Case, *Technovation*, 26: 463-472.

Calderero Gutiérrez, A.; Fernández Macho, J.; Kuittinen, J, H. (2009). *European regional innovation. A methodological and updated alternative for RIS* - Dyna (Spain), Vol. 84 n. 6, pp. 501-516.

Camagni, R.; Capello, R. (2002). Apprendimento collettivo, innovazione e contesto locale. In: Camagni R. and Capello R. (a cura di), *Apprendimento collettivo e competitività territoriale*, Franco Angeli, Milano.

Capriati, M. (2003). Spesa pubblica e capacità innovative delle regioni italiane, *Atti della XXIV Conferenza dell'Associazione Italiana di Scienze Regionali*, Perugia, 8-10 ottobre 2003.

Carlsson, B.; Jacobbson, S.; Holmen, M.; Rickne, A.(2002). Innovation Systems: Analitical and Methodological Issues, *Research Policy*, 31: 233-245.

Chung, S. (2002). Building a National Innovation Systems Through Regional Innovation Systems, *Technovation*, 22: 485-491.

Cicchitelli, G. (2012). *Statistica. Principi e metodi*, II edizione. Pearson Italia, Milano-Torino.

Ciciotti, E. (2000). *Competitività e territorio. L'economia regionale nei Paesi industrializzati*, Carocci, Roma.

Cooke, P. (2001). Regional innovation systems, clusters and the knowledge economy. *Industrial and Corporate Change*, 4 (10): 945-974.

Cooke, P.; Gomez Uranga, M.; Etxebarria, G. (1997). Regional systems of Innovation: Institutional and Organisational Dimensions, *Research Policy*, 26: 475-491.

Delvecchio, F. (2015). *Statistica per l'analisi dei fenomeni sociali*. Cleup, Padova.

Delvecchio, F. (1995). *Scale di misura e indicatori sociali*, Cacucci Editore, Bari, pp 87-88.

Doloreaux, D. (2002). What We Should Know About Regional Innovation Systems of Innovation, *Technology in Society*, 24: 243-263.

d'Ovidio, F. D. (2007). *Valutazione di fenomeni sociali e dei servizi di pubblica utilità*. Cacucci Editore Bari.

Edquist, C. (2005). Systems of Innovation: Perspectives and Challenges. In: Fagerberg J., Mowery D. and Nelson R.R. (eds.), *The Oxford Handbook of Innovation*, Norfolk, Oxford University Press.

Edquist, C. (2001). *Systems of Innovation for Development (SID)*, Background paper for the UNIDO World Industrial Development Report (WIDR), written for Investment Promotion and Institutional Capacity-building division, Industrial Policies and Research Branch, United Nations Industrial Development Organisation (UNIDO), January.

Edquist, C. (1997). *Systems of innovation: Technologies, institutions and organizations*, London, Pinter Publishers.

Edquist, C.; Hommen, L. (1999). Systems of Innovation: theory and policy from the demand side, *Technology in Society*, 21: 63–79.

Edquist, C.; McKelvey, M.D. (2000). Systems of Innovation. *Growth, Competiti–veness and Employment*, Cheltenham, Northampton: Edward Elgar.

Freeman, C.; Soete, L. (1997). *The Economics of Industrial Innovation*, Pinter, London.

Furman, J. L.; Porter, M. E.; Stern, S. (2002). The Determinants of National Innovative Capacity, *Research Policy*, 31: 899-933.

Hollanders, H. (2014). Do innovation scoreboards adequately measure regional innovation?, *Conference "Week of Innovative Regions in Europe (WIRE 2014)"*, Session n. 6, Athens, 12 June 2014.

Hollanders, H. (2009). Measuring innovation: the European Innovation Scoreboard. In: Villalba E. (ed.), *"Measuring Creativity", Proceedings from the conference, 'Can creativity be measured?'*, Luxembourg: Publications Office of the European Union, pp. 27-40.

Kline, S. J.; Rosenberg, N. (1986). An overview of innovation. In: Landau R. and Rosenberg N. (eds.); *The positive sum game*, Washington D.C., National Academy Press.

Legrenzi, P. (2005). *Creatività e innovazione*; Il Mulino, collana "Farsi un'idea".

Lundvall, B.-Å. (1992). *National Innovation Systems: Towards a Theory of Innovation and Interactive Learning*, London, Pinter Publishers.

Lundvall, B.-Å.; Johnson, B.; Lorenz, E. (2002). Why all this fuss about codified and tacit Knowledge? *Industrial and Corporate Change*, No 2: 245-62.

Malerba, F. (2000). La teoria evolutiva: i recenti sviluppi. In: Malerba F. (a cura di), *Economia dell'Innovazione*, Carocci, Roma.

Manjón, J. V. C. (2010). A Proposal of Indicators and Policy Framework for Innovation Benchmark in Europe, *Journal of Technology Management and Innovation*, Volume 5, Issue 2.

Montobbio, F. (2000). *National System of Innovation. A Critical Survey*. ESSY Working Paper, n. 2.

Nauwelaers, C.; Reid, A. (1995). Methodologies for the Evaluation of Regional Innovation Potential, *Scientometrics*, 34: 497-511.

Nelson, R.R. (1993). *National Innovation Systems: A Comparative Analysis*, Oxford, Oxford University Press.

Niosi, J. (2000). Regional systems of innovation: market pull and government push. In: Holbrook J.A and Wolfe D. (eds*) Knowledge, clusters and regional innovation*. Montreal: McGill-Queen's University Press.

Nonaka, I.; Takeuchi, H. (1995). *The Knowledge Creating Company*, Oxford/New York: Oxford University Press.

Padmore, T.; Gibson, H. (1998). Modelling Systems of Innovation: II. A Framework for Industrial Cluster Analysis in Regions, *Research Policy*, 26: 625-641.

Park, S.O. (2001). Regional Innovation Strategies in the Knowledge-Based Economy, *GeoJournal*, 53: 29-38.

Perry, B.; May, T. (2007). Governance, Science Policy and Regions: an introduction, *Journal of Regional Studies*, Volume 41, Issue 8, pp. 1039-1050.

Sirilli, G. (2000). La misurazione della ricerca: metodi e indicatori. In: Garonna P. e Iammarino S. (a cura di), *L'economia della ricerca*, Il Mulino, Bologna.

Trigilia, C. (2007). *La costruzione sociale dell'innovazione: economia, società e territorio*. Firenze University Press.

Pubblicazioni e Documenti istituzionali:

- Scoreboard Regionale dell'Innovazione per la comparazione delle performance del sistema innovativo lombardo – settembre 2006.
- Scoreboard Regionale dell'Innovazione per la comparazione delle performance del sistema innovativo piemontese – ottobre 2007.
- Analisi del Sistema Innovativo Regionale Puglia CERPI, 2008.
- Pro Inno Europe - Regional Innovation Scoreboard, 2009.
- Innovation Scoreboard Regione Campania, 2012.
- Innovation Union Scoreboard, 2014.
- Regional Innovation Scoreboard 2014.
- Innovation Union Scoreboard, 2015.
- European Innovation Scoreboard 2016.
- Regional Innovation Scoreboard 2016.

APPENDICE

Tabella A.1[17] *(Fonte: Data retrieval a cura di S. Marastoni su dati Istat e Eurostat).*

Codice identificativo indicatore	*Piemonte*	*Valle d'Aosta*	*Liguria*	*Lombardia*	*Provincia autonoma di Bolzano*	*Provincia autonoma di Trento*	*Veneto*	*Friuli Venezia Giulia*	*Emilia Romagna*	*Toscana*	*Umbria*	*Marche*	*Lazio*	*Abruzzo*	*Molise*	*Campania*	*Puglia*	*Basilicata*	*Calabria*	*Sicilia*	*Sardegna*
1	47,8	47,6	47,2	45,4	45,9	50,2	48,0	48,6	48,9	47,7	45,7	49,3	46,0	44,9	46,1	45,1	46,5	47,7	44,2	43,9	45,6
2	54,8	20,5	56,0	61,7	9,3	64,5	48,8	67,2	80,5	78,0	71,1	70,3	100,4	97,3	48,8	54,7	43,4	24,6	42,5	47,3	51,7
3	0,6	0,3	0,5	0,5	0,3	0,9	0,4	0,6	0,6	0,6	0,4	0,3	0,6	0,3	0,2	0,4	0,3	0,3	0,2	0,3	0,3
4	81,2	76,4	82,0	79,0	75,4	85,5	88,7	85,1	81,7	79,7	88,7	86,1	82,7	83,9	85,2	76,1	77,7	86,0	79,3	71,8	68,1
5	44,8	41,2	44,3	43,8	50,1	51,0	45,6	47,8	45,1	43,1	48,1	45,5	46,3	46,4	41,4	36,3	35,1	43,6	38,8	36,6	34,9
6	16,6	15,5	19,7	19,3	16,3	18,7	15,9	17,6	20,3	19,3	19,8	18,5	23,3	17,0	18,0	14,9	13,3	15,5	15,7	13,2	14,7
7	24,0	25,9	26,2	29,5	25,3	31,7	26,4	26,9	28,8	29,8	31,8	28,7	31,6	24,9	32,4	18,5	18,6	22,8	24,2	18,2	18,6
8	17,6	0,8	15,1	16,3	2,3	17,4	12,0	18,7	18,7	16,6	12,2	16,3	17,9	9,8	3,5	11,2	6,7	4,7	10,3	8,0	7,9
9	7,3	7,6	7,2	8,1	13,4	10,0	7,1	10,3	8,7	9,0	8,5	7,4	8,2	7,1	7,7	5,4	5,6	6,0	5,9	4,7	7,8
10	29,4	24,6	30,1	33,3	27,6	30,5	27,9	30,1	31,3	27,2	25,9	26,7	30,2	25,1	27,6	23,9	21,3	25,7	22,0	22,4	23,6
11	12,7	11,4	15,7	15,8	12,0	14,5	12,2	13,9	14,8	13,9	13,4	12,7	17,2	12,2	14,0	13,8	11,2	12,8	12,5	11,7	12,9
12	0,1	n.d.	1,1	2,4	n.d.	n.d.	1,2	1,3	3,7	3,9	3,0	1,3	1,7	11,6	n.d.	2,3	19,6	0,1	11,8	1,1	21,0
13	31,7	n.d.	200,7	140,1	n.d.	n.d.	5,7	28,3	264,2	67,5	31,1	19,4	11,2	9,1	62,7	46,8	3,3	2,3	3,1	2,9	18,2
14	21,8	24,1	15,7	16,7	26,3	24,6	17,3	18,6	17,4	15,5	17,3	16,8	16,1	24,9	21,5	14,3	15,6	20,1	19,7	14,1	16,5
15	0,4	0,1	0,6	0,3	0,2	0,9	0,4	0,6	0,5	0,6	0,6	0,4	1,1	0,5	n.d.	0,7	0,6	0,5	0,4	0,6	0,7
16	24,4	20,4	20,5	12,2	16,4	0,5	16,9	18,8	15,9	32,6	30,9	21,3	12,5	3,7	26,8	13,6	8,2	11,2	7,7	6,8	18,6
17	8,7	2,1	3,4	4,5	4,4	4,4	4,7	6,9	3,8	4,1	3,5	3,2	5,8	2,8	n.d.	3,1	3,6	2,2	3,2	1,3	1,8
18	93,3	95,2	99,0	95,6	96,1	96,6	95,4	97,2	95,0	94,3	96,5	92,5	91,0	93,3	90,8	92,9	89,5	94,7	94,2	93,9	95,0
19	53,1	33,0	42,5	54,2	47,7	50,4	58,0	58,5	53,8	55,9	47,1	44,4	50,7	47,7	38,1	45,2	46,9	52,2	40,6	43,0	44,6
20	1,6	0,2	0,7	0,9	0,4	0,8	0,8	0,8	1,1	0,6	0,2	0,4	0,5	0,3	0,3	0,5	0,2	0,1	0,0	0,3	0,0
21	7,9	5,4	7,7	7,0	6,3	4,9	5,6	6,5	7,2	5,5	4,8	4,9	4,8	7,8	3,0	4,4	7,5	4,4	3,5	7,4	3,1
22	201,3	109,4	133,5	201,5	249,0	118,6	220,2	492,6	279,4	143,3	75,6	129,1	52,4	45,6	7,7	28,6	26,3	27,9	26,2	12,7	13,5
23	225,6	448,5	151,4	394,9	546,6	168,3	394,4	287,7	309,3	270,0	134,5	233,5	195,5	115,0	67,2	124,6	96,0	73,1	59,8	51,7	60,0
24	39,2	22,3	28,0	39,1	30,6	33,7	43,5	41,1	36,0	37,7	31,6	27,2	32,6	32,4	20,1	25,3	30,4	28,3	23,8	26,3	26,0
25	11,2	n.d.	4,1	9,4	2,7	3,9	7,7	7,0	9,9	4,5	4,1	7,3	2,7	7,1	8,0	3,1	2,4	7,9	0,5	1,4	0,6
26	3,7	n.d.	2,9	5,0	2,0	2,7	2,7	2,4	3,1	2,8	2,2	2,7	6,8	3,2	n.d.	2,0	1,4	2,1	1,4	1,7	1,6
27	48,5	18,4	46,0	45,9	28,6	39,3	35,0	50,5	49,4	29,9	32,6	52,8	65,8	62,7	12,3	40,5	40,5	66,7	19,6	10,5	3,2
28	52,2	50,7	53,2	56,1	58,3	60,5	56,1	58,4	56,2	55,3	52,3	55,7	51,7	59,1	49,8	49,1	53,5	49,9	47,4	48,8	50,4

[17] Per economia editoriale, non sono rappresentati i dati delle ripartizioni territoriali e quello nazionale.

Tabella A.2[18] *(Fonte: Elaborazioni a cura di R. Mancarella su dati Istat ed Eurostat).*

Codice identificativo indicatore	*Piemonte*	*Valle d'Aosta*	*Liguria*	*Lombardia*	*Provincia autonoma di Bolzano*	*Provincia autonoma di Trento*	*Veneto*	*Friuli Venezia Giulia*	*Emilia Romagna*	*Toscana*	*Umbria*	*Marche*	*Lazio*	*Abruzzo*	*Molise*	*Campania*	*Puglia*	*Basilicata*	*Calabria*	*Sicilia*	*Sardegna*
1	9,52	9,48	9,40	9,04	9,14	10,00	9,56	9,68	9,74	9,50	9,10	9,82	9,16	8,94	9,18	8,98	9,26	9,50	8,80	8,75	9,08
2	5,46	2,04	5,58	6,15	0,93	6,42	4,86	6,69	8,02	7,77	7,08	7,00	10,00	9,69	4,86	5,45	4,32	2,45	4,23	4,71	5,15
3	6,71	3,76	5,88	5,76	3,29	10,00	4,71	7,53	7,29	6,47	4,12	3,41	7,41	3,29	2,71	4,59	2,94	3,06	2,71	3,53	3,65
4	9,15	8,61	9,24	8,91	8,50	9,64	10,00	9,59	9,21	8,99	10,00	9,71	9,32	9,46	9,61	8,58	8,76	9,70	8,94	8,09	7,68
5	8,78	8,08	8,69	8,59	9,82	10,00	8,94	9,37	8,84	8,45	9,43	8,92	9,08	9,10	8,12	7,12	6,88	8,55	7,61	7,18	6,84
6	7,12	6,65	8,45	8,28	7,00	8,03	6,82	7,55	8,71	8,28	8,50	7,94	10,00	7,30	7,73	6,39	5,71	6,65	6,74	5,67	6,31
7	7,41	7,99	8,09	9,10	7,81	9,78	8,15	8,30	8,89	9,20	9,81	8,86	9,75	7,69	10,00	5,71	5,74	7,04	7,47	5,62	5,74
8	9,41	0,43	8,07	8,72	1,23	9,30	6,42	10,00	10,00	8,88	6,52	8,72	9,57	5,24	1,87	5,99	3,58	2,51	5,51	4,28	4,22
9	5,45	5,67	5,37	6,04	10,00	7,46	5,30	7,69	6,49	6,72	6,34	5,52	6,12	5,30	5,75	4,03	4,18	4,48	4,40	3,51	5,82
10	8,83	7,39	9,04	10,00	8,29	9,16	8,38	9,04	9,40	8,17	7,78	8,02	9,07	7,54	8,29	7,18	6,40	7,72	6,61	6,73	7,09
11	7,38	6,63	9,13	9,19	6,98	8,43	7,09	8,08	8,60	8,08	7,79	7,38	10,00	7,09	8,14	8,02	6,51	7,44	7,27	6,80	7,50
12	0,04	2,31	0,54	1,12	2,31	2,31	0,59	0,61	1,77	1,84	1,42	0,64	0,82	5,54	2,31	1,08	9,34	0,04	5,65	0,52	10,00
13	1,20	1,99	7,60	5,30	1,99	1,99	0,22	1,07	10,00	2,56	1,18	0,73	0,42	0,35	2,37	1,77	0,13	0,09	0,12	0,11	0,69
14	8,28	9,16	5,97	6,35	10,00	9,33	6,57	7,08	6,61	5,90	6,56	6,36	6,13	9,45	8,15	5,42	5,94	7,64	7,46	5,35	6,28
15	3,64	0,91	5,45	2,73	1,82	8,18	3,64	5,45	4,55	5,45	5,45	3,64	10,00	4,55	4,88	6,36	5,45	4,55	3,64	5,45	6,36
16	7,49	6,24	6,28	3,73	5,04	0,15	5,17	5,75	4,88	10,00	9,46	6,52	3,82	1,13	8,20	4,15	2,50	3,44	2,37	2,08	5,69
17	10,00	2,41	3,91	5,17	5,06	5,06	5,40	7,93	4,37	4,71	4,02	3,68	6,67	3,22	4,44	3,56	4,14	2,53	3,68	1,49	2,07
18	9,42	9,62	10,00	9,66	9,71	9,76	9,64	9,82	9,60	9,53	9,75	9,34	9,19	9,42	9,17	9,38	9,04	9,57	9,52	9,48	9,60
19	9,08	5,64	7,26	9,26	8,15	8,62	9,91	10,00	9,20	9,56	8,05	7,59	8,67	8,15	6,51	7,73	8,02	8,92	6,94	7,35	7,62
20	10,00	1,26	4,47	5,66	2,39	5,22	4,72	5,28	6,86	3,90	1,51	2,77	3,14	1,95	1,95	3,33	1,45	0,31	0,19	1,57	0,25
21	10,00	6,84	9,75	8,86	7,97	6,20	7,09	8,23	9,11	6,96	6,08	6,20	6,08	9,87	3,80	5,57	9,49	5,57	4,43	9,37	3,92
22	4,09	2,22	2,71	4,09	5,06	2,41	4,47	10,00	5,67	2,91	1,53	2,62	1,06	0,93	0,16	0,58	0,53	0,57	0,53	0,26	0,27
23	4,13	8,21	2,77	7,22	10,00	3,08	7,22	5,26	5,66	4,94	2,46	4,27	3,58	2,10	1,23	2,28	1,76	1,34	1,09	0,95	1,10
24	9,01	5,13	6,44	8,99	7,03	7,75	10,00	9,45	8,28	8,67	7,26	6,25	7,49	7,45	4,62	5,82	6,99	6,51	5,47	6,05	5,98
25	10,00	4,67	3,66	8,39	2,41	3,48	6,88	6,25	8,84	4,02	3,66	6,52	2,41	6,34	7,14	2,77	2,14	7,05	0,45	1,25	0,54
26	5,44	4,04	4,26	7,35	2,94	3,97	3,97	3,53	4,56	4,12	3,24	3,97	10,00	4,71	4,04	2,94	2,06	3,09	2,06	2,50	2,35
27	7,27	2,76	6,90	6,88	4,29	5,89	5,25	7,57	7,41	4,48	4,89	7,92	9,87	9,40	1,84	6,07	6,07	10,00	2,94	1,57	0,48
28	8,63	8,38	8,79	9,27	9,64	10,00	9,27	9,65	9,29	9,14	8,64	9,21	8,55	9,77	8,23	8,12	8,84	8,25	7,83	8,07	8,33

[18] Per economia editoriale, non vengono rappresentati i dati relativi alle ripartizioni territoriali e quello nazionale.

Tabella A.3

Indici + Anno di Rilevazione	*Coefficienti di Correlazione Rho di Spearman*			
	PIL_PROCAPITE			
	2012	*2013*	*2014*	*2015*
1_2012	**0,266**	0,244	0,241	0,270
2_2012	**0,208**	0,175	0,175	0,197
3_2013	0,613	**0,572**	0,572	0,567
4_2015	0,029	0,032	0,062	**0,038**
5_2015	0,502	0,485	0,447	**0,479**
6_2015	0,475	0,484	0,469	**0,473**
7_2015	0,401	0,448	0,436	**0,435**
8_2012	**0,338**	0,301	0,299	0,335
9_2015	0,498	0,526	0,499	**0,508**
10_2015	0,693	0,706	0,683	**0,677**
11_2015	0,329	0,339	0,312	**0,284**
12_2014	**0,167**	0,128	0,094	0,071
13_2014	0,598	0,607	**0,615**	0,570
14_2013	0,192	**0,230**	0,206	0,225
15_2012	**0,239**	0,258	0,256	0,297
16_2015	0,063	0,080	0,084	**0,054**
17_2012	**0,475**	0,481	0,484	0,482
18_2015	0,465	0,454	0,413	**0,425**
19_2012	**0,319**	0,270	0,240	0,267
20_2013	0,609	**0,582**	0,582	0,580
21_2012	**0,333**	0,314	0,347	0,339
22_2012	**0,699**	0,649	0,633	0,671
23_2015	0,815	0,769	0,766	**0,797**
24_2012	**0,425**	0,388	0,379	0,400
25_2015	0,035	0,128	0,175	**0,168**
26_2015	0,616	0,631	0,637	**0,618**
27_2014	0,125	0,035	**0,008**	0,048
28_2013	0,505	**0,470**	0,463	0,511

Fonte: Elaborazioni a cura di R. Mancarella su dati Istat ed Eurostat

Tabella A.3 bis

Indici + Anno di Rilevazione	*Test di normalità di Shapiro-Wilk*
	p-value
1_2012	0,039
2_2012	0,005
3_2013	0,041
4_2015	0,009
5_2015	0,026
6_2015	0,006
7_2015	0,017
8_2012	0,029
9_2015	0,044
10_2015	0,033
11_2015	0,017
12_2014	0,000
13_2014	0,000
14_2013	0,020
15_2012	0,004
16_2015	0,039
17_2012	0,052
18_2015	0,001
19_2012	0,043
20_2013	0,002
21_2012	0,002
22_2012	0,001
23_2015	0,001
24_2012	0,003
25_2015	0,000
26_2015	0,004
27_2014	0,000
28_2013	0,004
Pil procapite 2012	0,013
Pil procapite 2013	0,039
Pil procapite 2014	0,080
Pil procapite 2015	0,202

Fonte: Elaborazioni a cura di R. Mancarella su dati Istat ed Eurostat

Tabella A.4[19] *(Fonte: Elaborazioni a cura di R. Mancarella su dati Istat ed Eurostat).*

Codice identificativo indicatore	*Piemonte*	*Valle d'Aosta*	*Liguria*	*Lombardia*	*Provincia autonoma di Bolzano*	*Provincia autonoma di Trento*	*Veneto*	*Friuli Venezia Giulia*	*Emilia Romagna*	*Toscana*	*Umbria*	*Marche*	*Lazio*	*Abruzzo*	*Molise*	*Campania*	*Puglia*	*Basilicata*	*Calabria*	*Sicilia*	*Sardegna*
1	2,53	2,52	2,50	2,40	2,43	2,66	2,54	2,57	2,59	2,53	2,42	2,61	2,44	2,38	2,44	2,39	2,46	2,53	2,34	2,33	2,42
2	1,13	0,42	1,16	1,28	0,19	1,33	1,01	1,39	1,66	1,61	1,47	1,45	2,08	2,01	1,01	1,13	0,90	0,51	0,88	0,98	1,07
3	3,84	2,15	3,36	3,30	1,88	5,72	2,69	4,31	4,17	3,70	2,36	1,95	4,24	1,88	1,55	2,62	1,68	1,75	1,55	2,02	2,09
4	0,35	0,33	0,35	0,34	0,32	0,37	0,38	0,37	0,35	0,34	0,38	0,37	0,35	0,36	0,37	0,33	0,33	0,37	0,34	0,31	0,29
5	4,21	3,87	4,16	4,11	4,71	4,79	4,28	4,49	4,24	4,05	4,52	4,27	4,35	4,36	3,89	3,41	3,30	4,09	3,64	3,44	3,28
6	3,37	3,15	4,00	3,92	3,31	3,80	3,23	3,57	4,12	3,92	4,02	3,76	4,73	3,45	3,65	3,02	2,70	3,15	3,19	2,68	2,98
7	3,22	3,48	3,52	3,96	3,40	4,26	3,54	3,61	3,87	4,00	4,27	3,85	4,24	3,34	4,35	2,48	2,50	3,06	3,25	2,44	2,50
8	3,18	0,14	2,73	2,94	0,42	3,14	2,17	3,38	3,38	3,00	2,20	2,94	3,23	1,77	0,63	2,02	1,21	0,85	1,86	1,44	1,43
9	2,77	2,88	2,73	3,07	5,08	3,79	2,69	3,90	3,30	3,41	3,22	2,81	3,11	2,69	2,92	2,05	2,12	2,27	2,24	1,78	2,96
10	5,98	5,00	6,12	6,77	5,61	6,20	5,67	6,12	6,36	5,53	5,27	5,43	6,14	5,10	5,61	4,86	4,33	5,22	4,47	4,55	4,80
11	2,10	1,88	2,60	2,61	1,98	2,40	2,02	2,30	2,45	2,30	2,22	2,10	2,84	2,02	2,31	2,28	1,85	2,12	2,07	1,93	2,13
12	0,01	0,39	0,09	0,19	0,39	0,39	0,10	0,10	0,29	0,31	0,24	0,11	0,14	0,92	0,39	0,18	1,56	0,01	0,94	0,09	1,67
13	0,74	1,22	4,67	3,26	1,22	1,22	0,13	0,66	6,15	1,57	0,72	0,45	0,26	0,21	1,46	1,09	0,08	0,05	0,07	0,07	0,42
14	1,91	2,11	1,37	1,46	2,30	2,15	1,51	1,63	1,52	1,36	1,51	1,46	1,41	2,18	1,88	1,25	1,37	1,76	1,72	1,23	1,44
15	0,87	0,22	1,31	0,65	0,44	1,96	0,87	1,31	1,09	1,31	1,31	0,87	2,39	1,09	1,17	1,52	1,31	1,09	0,87	1,31	1,52
16	0,41	0,34	0,34	0,20	0,27	0,01	0,28	0,31	0,27	0,54	0,51	0,35	0,21	0,06	0,45	0,23	0,14	0,19	0,13	0,11	0,31
17	4,75	1,15	1,86	2,46	2,40	2,40	2,57	3,77	2,07	2,24	1,91	1,75	3,17	1,53	2,11	1,69	1,97	1,20	1,75	0,71	0,98
18	4,01	4,09	4,25	4,10	4,13	4,15	4,10	4,17	4,08	4,05	4,14	3,97	3,91	4,01	3,90	3,99	3,84	4,07	4,04	4,03	4,08
19	2,89	1,80	2,32	2,95	2,60	2,75	3,16	3,19	2,93	3,05	2,57	2,42	2,76	2,60	2,08	2,46	2,56	2,84	2,21	2,34	2,43
20	5,82	0,73	2,60	3,29	1,39	3,04	2,75	3,07	3,99	2,27	0,88	1,61	1,83	1,13	1,13	1,94	0,84	0,18	0,11	0,92	0,15
21	3,33	2,28	3,24	2,95	2,65	2,06	2,36	2,74	3,03	2,32	2,02	2,06	2,02	3,29	1,26	1,85	3,16	1,85	1,47	3,12	1,31
22	2,86	1,55	1,89	2,86	3,53	1,68	3,13	6,99	3,97	2,03	1,07	1,83	0,74	0,65	0,11	0,41	0,37	0,40	0,37	0,18	0,19
23	3,29	6,54	2,21	5,76	7,97	2,45	5,75	4,20	4,51	3,94	1,96	3,40	2,85	1,68	0,98	1,82	1,40	1,07	0,87	0,75	0,87
24	3,83	2,18	2,74	3,82	2,99	3,29	4,25	4,02	3,52	3,68	3,09	2,66	3,19	3,17	1,96	2,47	2,97	2,76	2,33	2,57	2,54
25	1,68	0,78	0,61	1,41	0,40	0,58	1,15	1,05	1,48	0,67	0,61	1,09	0,40	1,06	1,20	0,46	0,36	1,18	0,07	0,21	0,09
26	3,36	2,50	2,64	4,54	1,82	2,45	2,45	2,18	2,82	2,54	2,00	2,45	6,18	2,91	2,50	1,82	1,27	1,91	1,27	1,55	1,45
27	0,91	0,34	0,86	0,86	0,54	0,74	0,66	0,95	0,93	0,56	0,61	0,99	1,23	1,18	0,23	0,76	0,76	1,25	0,37	0,20	0,06
28	4,36	4,23	4,44	4,68	4,87	5,05	4,68	4,87	4,69	4,62	4,37	4,65	4,32	4,93	4,16	4,10	4,47	4,17	3,96	4,07	4,21

[19] Per economia editoriale, non vengono rappresentati i dati relativi alle ripartizioni territoriali e quello nazionale.

Tabella A.5

CRI					***RIS***				
Regioni	*indice*	*rango*	*campo 0-1*	*campo variaz.*	*Regioni*	*indice*	*rango*	*campo 0-1*	*campo variaz*
Emilia Romagna	7,58	1,0	0,76	1,00	*Friuli Venezia Giulia*	0,46	1	0,46	1,00
Friuli Venezia Giulia	7,42	2,0	0,74	0,95	*Piemonte*	0,45	2	0,45	0,999
Lombardia	7,26	3,0	0,73	0,90	*Veneto*	0,43	3	0,43	0,89
Piemonte	7,11	4,0	0,71	0,86	*Lazio*	0,42	4	0,42	0,86
Prov. aut. di Trento	6,80	5,0	0,68	0,76	*Lombardia*	0,42	5	0,42	0,85
Lazio	6,78	6,0	0,68	0,75	*Emilia Romagna*	0,42	6	0,42	0,82
Toscana	6,49	7,0	0,65	0,67	*Prov. aut. di Trento*	0,40	7	0,40	0,75
Veneto	6,41	8,0	0,64	0,64	*Toscana*	0,38	8	0,38	0,67
Liguria	6,39	9,0	0,64	0,63	*Liguria*	0,37	9	0,37	0,60
Prov. aut. di Bolzano	6,30	10,0	0,63	0,61	*Umbria*	0,37	10	0,37	0,58
Marche	5,81	11,0	0,58	0,46	*Abruzzo*	0,36	11	0,36	0,56
Abruzzo	5,66	12,0	0,57	0,41	*Prov. aut. di Bolzano*	0,35	12	0,35	0,53
Umbria	5,62	13,0	0,56	0,40	*Molise*	0,34	13	0,34	0,48
Valle d'Aosta	5,29	14,0	0,53	0,30	*Marche*	0,32	14	0,32	0,39
Molise	5,03	15,0	0,50	0,22	*Basilicata*	0,32	15	0,32	0,38
Campania	4,99	16,0	0,50	0,21	*Campania*	0,32	16	0,32	0,38
Basilicata	4,76	17,0	0,48	0,13	*Puglia*	0,32	17	0,32	0,36
Puglia	4,74	18,0	0,47	0,13	*Sicilia*	0,30	18	0,30	0,27
Sardegna	4,51	19,0	0,45	0,06	*Valle d'Aosta*	0,30	19	0,30	0,26
Calabria	4,40	20,0	0,44	0,02	*Calabria*	0,26	20	0,26	0,11
Sicilia	4,32	21,0	0,43	0,00	*Sardegna*	0,24	21	0,24	0,00
Ripartizioni territoriali	*indice medio*	*rango medio*	*campo 0-1*	*campo variaz.*	*Ripartizioni territoriali*	*indice medio*	*rango medio*	*campo 0-1*	*campo variaz.*
Nord	6,62	6,88	0,66	0,71	*Nord*	0,40	7,25	0,40	0,73
Centro	6,46	7,60	0,65	0,66	*Centro*	0,38	8,40	0,38	0,66
Sud	4,93	16,33	0,49	0,19	*Sud*	0,32	15,33	0,32	0,38
Isole	4,42	20,00	0,44	0,03	*Isole*	0,27	19,50	0,27	0,14
Italia	5,89		0,59	0,48	*Italia*	0,36		0,36	0,56

Fonte: Elaborazioni a cura di R. Mancarella su dati Istat e Eurostat

Tabella A.6

	CRI				***RIS***			
Valori caratteristici	***valori originali***	***campo 0-1***	***campo variaz***	*z*	***valori originali***	***campo 0-1***	***campo variaz***	*z*
Varianza	1,07	0,01	0,10	1,00	0,00	0,00	0,08	1,00
C.V.	18,16	1,82	20,91	-	0,98	0,98	13,72	-
Mediana	5,81	0,58	0,46	-0,08	0,36	0,36	0,56	-0,01
Min	4,32	0,43	0,00	-1,52	0,24	0,24	0,00	-2,02
Max	7,58	0,76	1,00	1,64	0,46	0,46	1,00	1,59
1° Quartile	4,88	0,49	0,17	-0,98	0,32	0,32	0,37	-0,68
3° Quartile	6,79	0,68	0,76	0,87	0,42	0,42	0,83	0,99
Campo Variazione	3,26	0,33	1,00	3,15	0,21	0,21	1,00	3,61
Differenza Interquartilica	1,92	0,19	0,59	1,85	0,10	0,10	0,46	1,67
Media Quartile Alto	7,23	0,72	0,89	1,30	0,44	0,44	0,92	1,30
Media Quartile. Basso	4,55	0,45	0,07	-1,30	0,28	0,28	0,20	-1,29
Varianza Quarile Alto	0,07	0,00	0,01	0,07	0,00	0,00	0,00	0,06
Varianza Quartile Basso	0,03	0,00	0,00	0,03	0,00	0,00	0,02	0,21
Varianza Quartili Intermedi	0,35	0,00	0,03	0,33	0,00	0,00	0,02	0,26
Scarto Interquartile	2,69	0,27	0,82	2,60	0,15	0,15	0,72	2,59
Test Likert	16,74	16,74	16,74	16,74	9,91	9,91	9,91	9,91

Fonte: Elaborazioni a cura di R. Mancarella su dati Istat e Eurostat

UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO
Metodi e analisi statistiche 2017
DIPARTIMENTO DI ECONOMIA E FINANZA
ISBN: 978-88-6629-013-1
pp. 179-184

Un indice composito dei fattori di rischio della salute derivante dagli stili di vita

Monica Carbonara*

ISTAT, Ufficio territoriale per la Puglia

Riassunto: Molti studi dimostrano che un'alimentazione scorretta, il fumo, l'abuso di alcol ed un'insufficiente attività fisica sono fattori di rischio per numerose patologie cronico-degenerative che possono causare disuguaglianze di salute. A partire dai dati dell'Indagine multiscopo "Aspetti della vita quotidiana" 2015 dell'Istat è stato elaborato un indicatore composito costruito attraverso la combinazione di cinque indicatori elementari che descrivono gli stili di vita della popolazione italiana. I valori ottenuti forniscono elementi utili per l'individuazione di potenziali aree di criticità.

Keywords: salute, stili di vita, indici sintetici.

1. Introduzione

Secondo le stime dell'Organizzazione Mondiale della Sanità, in Europa l'86% dei decessi ed il 76% della perdita di anni di vita in buona salute sono provocati da patologie croniche causate da fattori di rischio modificabili: ipertensione, fumo di tabacco, sedentarietà, elevato consumo di alcol, ipercolesterolemia, obesità e scarso consumo di frutta e verdura. Diversi studi, infatti, hanno dimostrato che abitudini e comportamenti malsani sono fattori di rischio che possono causare disuguaglianze di salute in quanto contribuiscono notevolmente ad aumentare il rischio di insorgenza di diverse patologie e a peggiorare il loro decorso.

L'obiettivo di questo lavoro è costruire un indice sintetico, mediante l'applicazione di una opportuna combinazione di indicatori relativi ai fattori di rischio o di

* email: mocarbon@istat.it.

protezione derivanti dagli stili di vita e tratti dall'Indagine multiscopo "Aspetti della vita quotidiana" 2015. Tale indicatore è utile per valutare la sostenibilità degli attuali livelli di salute della popolazione italiana e individuare potenziali aree di criticità.

2. Metodologia

Dall'Indagine Istat sono stati selezionati gli indicatori che costituiscono i principali fattori prevedibili, quali il consumo di tabacco, il sovrappeso e l'obesità, i comportamenti a rischio nel consumo di alcol, la mancanza di attività fisica e la cattiva alimentazione.

Per ciascuno di essi è stata definita la polarità, ossia il segno della relazione tra ciascun indicatore e il fenomeno che si intente misurare (Tabella 1).

Tabella 1. *Indicatori e polarità*

Indicatori	Polarità
Proporzione standardizzata di persone di 18 anni o più in sovrappeso o obese *(valori percentuali)*	–
Proporzione standardizzata di persone di 14 anni o più che dichiarano di fumare attualmente *(valori percentuali)*	–
Proporzione standardizzata di persone di 14 anni e più che presentano almeno un comportamento a rischio nel consumo di alcol *(valori percentuali)*	–
Proporzione standardizzata di persone di 14 anni e più che non praticano alcuna attività fisica *(valori percentuali)*	–
Proporzione standardizzata di persone di 3 anni e più che consumano quotidianamente almeno 4 porzioni di frutta e/o verdura *(valori percentuali)*	+

Per il calcolo dell'indicatore sintetico è stato utilizzato Adjusted Mazziotta-Pareto Index ($AMPI^{+/-}$), che rispetta tutti i requisiti di un indice composito:
- comparabilità spaziale e temporale;
- non sostituibilità[1] degli indicatori elementari;
- semplicità e trasparenza di calcolo;
- immediata fruizione e interpretazione dei risultati di output;
- robustezza dei risultati ottenuti.

Gli indicatori elementari sono stati svincolati dalla loro unità di misura e depurati dalla loro variabilità mediante una standardizzazione min-max[2] degli indicatori elementari.

[1] Le componenti di un indice sintetico sono dette non sostituibili se non è ammessa compensazione tra di esse.

Data la matrice originaria degli indicatori $X=\{x_{ij}\}$, si costruisce la matrice standardizzata $R=\{r_{ij}\}$ in cui:

$$r_{ij} = \frac{(x_{ij} - Min\, x_j)}{(Max\, x_j - Min\, x_j)} \cdot 60 + 70$$

dove:

- x_{ij} è il valore dell'indicatore j nell'unità i,
- $Min\, x_j = Rif\ x_j - \Delta x_j$ e $Max\, x_j = Rif\ x_j + \Delta x_j$ sono i *goalposts* dell'indicatore j.

Nei *goalposts,* definendo $Inf\ x_j$ e $Sup\ x_j$, rispettivamente, il minimo e il massimo dell'indicatore *j-esimo* per tutto il periodo considerato:

- $Rif\ x_j$ è il valore di riferimento dell'indicatore *j-esimo*,
- $\Delta = (Sup\ x_j - Inf\ x_j)/2$.

I valori r_{ij} così ottenuti saranno compresi all'incirca nell'intervallo 70-130, dove 100 rappresenta il dato di riferimento (media nazionale)

Gli indicatori normalizzati sono aggregati con peso uguale mediante media aritmetica semplice, una funzione di sintesi additiva che in quanto tale presuppone un effetto compensativo fra gli indicatori elementari.

Nell'ipotesi di non sostituibilità o sostituibilità parziale degli indicatori elementari, in questa applicazione l'effetto compensativo della media aritmetica (effetto medio) è corretto aggiungendo alla media un fattore (coefficiente di penalità) che dipende dalla variabilità dei valori normalizzati di ciascuna unità (denominata variabilità orizzontale), ossia dalla variabilità degli indicatori rispetto ai valori di riferimento utilizzati per la normalizzazione.

Indicando con Mr_i e Sr_i la media aritmetica e, rispettivamente, lo scostamento quadratico medio dei valori normalizzati degli indicatori dell'unità i, l'indice è dato da:

$$AMPI_i^{+/-} = Mr_i \pm Sr_i cv_i$$

dove $cv_i = Sr_i / Mr_i$ è il coefficiente di variazione dei valori normalizzati degli indicatori dell'unità i.

[2] Questa procedura consiste nel riproporzionare il valore assunto da ciascuna unità in modo che oscilli tra il valore più basso assunto dall'indicatore, posto uguale a 0, e quello più elevato, posto uguale a 1.

Il fattore correttivo è funzione diretta del coefficiente di variazione dei valori normalizzati degli indicatori per ogni unità e, a parità di media aritmetica, consente di penalizzare le unità che presentano un maggiore squilibrio fra gli indicatori, spingendo verso l'alto il valore dell'indice (più è basso il valore dell'indice, peggiori sono gli stili di vita).

La robustezza del metodo individuato è stata valutata attraverso un'analisi di influenza che ha permesso di verificare se e con quale intensità cambia la graduatoria dell'indice composito a seguito dell'eliminazione dall'insieme di partenza di un indicatore elementare.

L'analisi è stata condotta utilizzando il software COMIC (Composite Indices Creator), sviluppato dall'Istat, che consente di calcolare ed eventualmente confrontare gli indici compositi prodotti in modo agevole, senza dover ricorrere a programmi informatici scritti *ad hoc*.

3. Risultati

L'analisi descrittiva mostra la presenza di correlazioni significative tra sedentarietà e peso (r = 0,877), alimentazione e alcol (r= 0.403), alcol e peso (r = -0.707), sedentarietà e alcol (r = -0.823), alimentazione e peso (r = -0.798) e alimentazione e sedentarietà (r = -0,655) (Tabella 2).

Tabella 2. *Matrice di correlazione*

Indicatore base	Peso	Fumo	Alcol	Sedentarietà	Alimentazione
Peso	1.000	0.123	-0.707	0.877	-0.798
Fumo	0.123	1.000	-0.256	0.192	0.197
Alcol	**-0.707**	-0.256	1.000	-0.823	0.403
Sedentarietà	**0.877**	0.192	**-0.823**	1.000	-0.655
Alimentazione	**-0.798**	0.197	**0.403**	**-0.655**	1.000

Dall'analisi di influenza si evince che la variabile che più influisce nella graduatoria dell'indice composito è il fumo (σ=2,57) (Figura 1).

I valori più elevati dell'indice composito si registrano in Trentino Alto Adige (116,7), Marche (110,7) e Piemonte (109,5). Il Molise (78,1) occupa l'ultima posizione della graduatoria, anche se è tutto il Mezzogiorno ad essere maggiormente svantaggiato (Figura 2).

Figura 1. *Analisi di influenza - Scarto quadratico medio degli shifts delle graduatorie.*

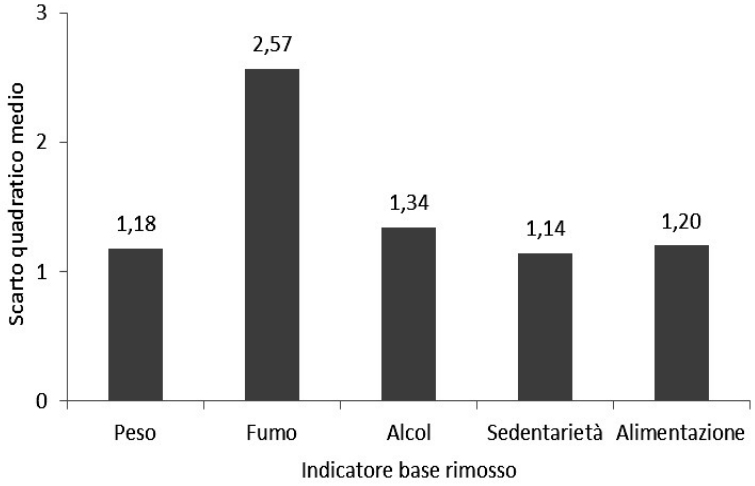
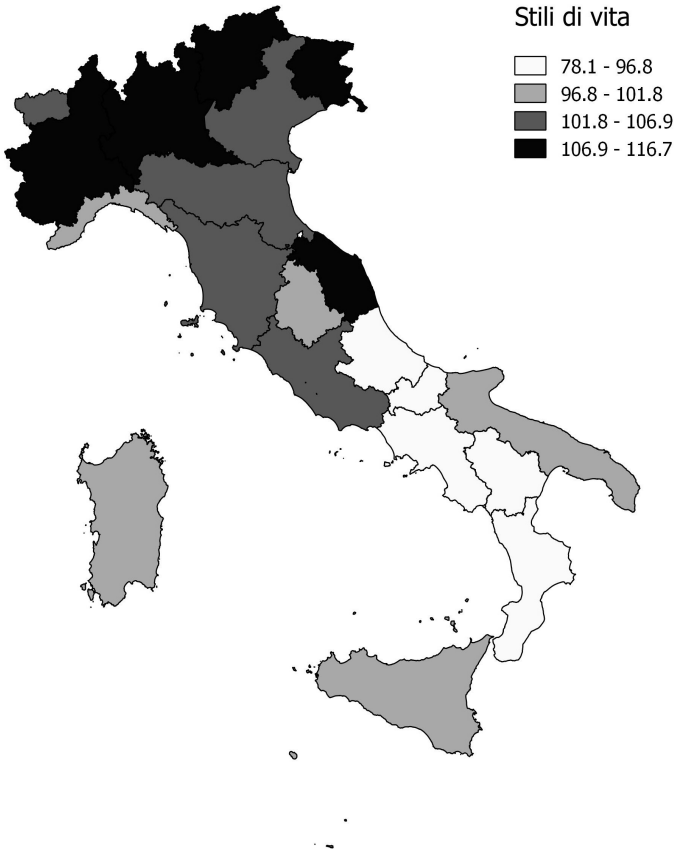


Figura 2. *Indice composito degli stili di vita delle regioni italiane (Indice Italia: 100). Anno 2015.*



4. Conclusioni

L'indicatore composito ha consentito di evidenziare le disuguaglianze geografiche nelle esposizioni regionali ai principali fattori di rischio analizzati: fumo, alcol, eccesso ponderale, inattività fisica e alimentazione scorretta.

In particolare, l'indice ha confermato il divario tra le regioni del Mezzogiorno e quelle del Centro-Nord legato all'adozione stili di vita scorretti. Il Mezzogiorno, infatti, registra la quota più alta di persone in sovrappeso e obese, stili di vita sedentari e un inadeguato consumo di frutta e verdura, soprattutto tra le donne. L'abitudine al fumo e il consumo a rischio di alcol, invece, sono più diffusi al Centro-Nord.

Le differenze territoriali negli stili di vita insalubri inducono a riflettere sulla necessità di adottare politiche mirate che contrastino i comportamenti a rischio e favoriscano la salute e il benessere di tutti i cittadini.

Riferimenti bibliografici

Mazziotta, M.; Pareto, A. (2011). Nuove misure del benessere: dal quadro teorico alla sintesi degli indicatori. *SISmagazine-Rivista on-line della SIS*. http://old.sis-statistica.org/magazine/spip.php?article194

Mazziotta, M.; Pareto, A. (2013). Methods for Constructing Composite Indices: One for All or All for One?, *Rivista Italiana di Economia Demografia e Statistica*, vol. LXVII, n. 2: pp 67-80. http://www.sieds.it/listing/RePEc/journl/2013LXVII_N2_10_Mazziotta_Pareto.pdf

Mazziotta, M.; Pareto, A. (2014). A Composite Index for measuring italian regions' development over time, *Rivista Italiana di Economia Demografia e Statistica*, Volume LXVIII, n. 3/4: pp. 127-134. http://www.sieds.it/listing/RePEc/journl/2014LXVIII_3-4_RIEDS_127-134_Mazziotta_Pareto.pdf

Massoli, P.; Mazziotta, M.; Pareto, A.; Rinaldelli, C.: *COMposite Index Creator – COMIC - Programma per il calcolo di indicatori compositi e relativa analisi di influenza sviluppato nell'ambito delle attività della commissione scientifica del progetto sul Benessere Equo Sostenibile*. ISTAT.

UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO *Metodi e analisi statistiche 2017*
DIPARTIMENTO DI ECONOMIA E FINANZA
ISBN: 978-88-6629-013-1
pp. 185-210

La gestione del rischio di credito attraverso metodi statistici: una verifica empirica

Domenico Summo*

Università degli studi di Bari Aldo Moro (Italy)

Riassunto: L'analisi di *credit scoring* è un sistema automatizzato adottato dalle banche e dagli intermediari finanziari per valutare le richieste di finanziamento della clientela. Si tratta di un insieme di tecniche statistico-economiche che valutano la probabilità di insolvenza del richiedente (soggetto privato o impresa) mediante l'utilizzo di un punteggio (*il credit score)* che permette di classificare il richiedente il finanziamento come solvente o insolvente e quindi meritevole o meno di una concessione creditizia. In ambito aziendale, i dati utilizzati in queste procedure sono il risultato dell'analisi e della classificazione del bilancio in base a criteri che evidenziano e permettono di valutare le performance economico-finanziarie del richiedente. L'output del modello (*lo score*) viene confrontato con un *benchmark* fissato dalla banca o dall'intermediario (*il valore di cut off*) e al di sotto di tale valore eventuali richieste di finanziamento verranno respinte o sottoposte a revisione. È opportuno chiarire la sottile distinzione tra *credit scoring* e *credit rating* che utilizzano metodologie e indicatori di rischio differenti, nonostante siano spesso impropriamente utilizzati come sinonimi; mentre il *credit scoring* calcola la probabilità che il debitore risulti insolvente, il *credit rating* è un'opinione espressa da apposite agenzie specializzate sul merito creditizio. Nonostante questa sostanziale differenza i due sistemi si sono fortemente legati con l'accordo di Basilea 2, e pertanto i sistemi di *rating* si servono dell'implementazione di modelli di *scoring* per classificare la clientela.

Keywords: *credit scoring;* analisi discriminante; regressione logistica.

* Autore corrispondente: domenico.summo@uniba.it

1. Introduzione

Oggi molte aziende richiedono strumenti in grado di valutare con attenzione e continuità l'area del credito al fine di esporsi nei confronti dei propri clienti in modo mediato tra l'opportunità di fare profitto e il rischio di perdere i propri crediti. Le aziende, sia di piccole che di grandi dimensioni, sentono sempre più l'esigenza di usufruire di informazioni utili in tempo reale, relative sia al mercato di sbocco che ai rispettivi partner commerciali dislocati ovunque. L'informazione è diventata un elemento fondamentale, sia per orientare i processi produttivi e le strategie aziendali, sia per ridurre i diversi rischi cui l'azienda è soggetta.

In questo nuovo sistema si inseriscono molte società di servizi che forniscono ai propri clienti le informazioni e i dati necessari a ridurre i rischi, e ad un tempo per accrescere proprio quell'esperienza e quella capacità di cui l'azienda stessa ha bisogno per rimanere competitiva in un mercato che diventa sempre più globale. Tali nuove società di servizi fungono da legame tra le aziende della *old-economy* e le nuove tecnologie prodotte dalla *new-economy*; esse hanno permesso di disporre di un'incredibile quantità di dati e di informazioni, in brevissimo tempo, provenienti da qualsiasi parte del mondo industrializzato, aiutano a realizzare in breve tempo studi di fattibilità, informano in merito alle procedure necessarie per ottenere finanziamenti. In questo ambito, tante sono le società specializzate nella valutazione degli affidamenti commerciali, ma poche sono quelle che adottano metodologie oggettive e prettamente statistiche; a tal proposito, è doveroso affermare che in questi ultimi anni le stesse hanno manifestato l'impellente bisogno di munirsi di nuovi modelli di *scoring*, sicuri e soprattutto oggettivi.

Gli indici di bilancio vengono correntemente utilizzati nella determinazione dei rischi connessi agli impieghi bancari. La loro diffusione operativa appare, tuttavia, ancora contenuta, sebbene sia in forte crescita nelle fasi di selezione, di revisione e di monitoraggio dell'andamento dei rapporti creditizi. La maggior parte delle applicazioni automatizzate utilizzate in Italia per la valutazione dei fidi alle imprese sono state, fino a pochissimi anni fa, di tipo soggettivo e univariato ed erano finalizzate al monitoraggio degli andamenti dei rapporti creditizi sui dati interni alla banca e sui dati di ritorno dalla Centrale dei rischi (Savona e Sironi, 2000).

Molte società di software offrono modelli di monitoraggio automatico che consentono di scegliere gli indici ritenuti rilevanti per la composizione dell'indicatore di sintesi. Il vero successo di un sistema di scoring consiste però proprio nell'accurata individuazione statistica degli indici e dei loro pesi.

2. – Fonte dei dati e scopo del lavoro

La società Eurocredit, grazie al proprio sistema informativo, all'armonizzazione tra esperienza acquisita, investimenti tecnologici e risorse umane, è capace di soddisfare tutte le esigenze di un'azienda, dall'identificazione del potenziale cliente e del suo grado di solvibilità fino alla gestione delle transazioni economiche.

L'Eurocredit è specializzata nella realizzazione di indagini di natura economico-valutative, si impegna anche a monitorare continuamente il pro–prio pacchetto clienti e a fornire loro il supporto informatico necessario a migliorarne l'assetto organizzativo. Essa raccoglie, incrocia, valuta dati e notizie sulle imprese allo scopo di delineare il livello di rischio e il relativo grado di affidabilità, evitando, in tal modo, di concedere crediti per forniture a soggetti economici insolventi. I servizi Eurocredit rendono possibile, in tal modo, l'individuazione di nuovi potenziali clienti e la scelta in maniera più accurata dei propri fornitori, attraverso un aggiornamento costante della loro affidabilità e delle potenzialità commerciali.

Dato che il rischio di credito rappresenta un problema, con il quale le banche e le aziende si devono continuamente confrontare e dato che ormai i vecchi criteri di misurazione e verifica dell'affidabilità di un partner commerciale non possono essere più applicabili, Eurocredit ha elaborato un proprio modello di scoring, con lo scopo di assegnare un punteggio di rischio agli affidamenti commerciali e di determinare l'esposizione massima consigliata per una fornitura di merce.

Tale modello, in particolare, si articola in sei indicatori principali che hanno il compito di delineare il profilo comportamentale dell'azienda in esame. Essi si riferiscono alla forma giuridica, all'esperienza nei pagamenti, all'anzianità aziendale, al settore d'attività, al territorio nel quale l'azienda opera e al risultato d'esercizio. Ciascuna di queste sei componenti risulta, a sua volta, scorporata in diversi segmenti, al fine di determinare in modo più dettagliato gli aspetti del soggetto in esame e di analizzare più attentamente le varie caratteristiche, combinandole con i valori medi del settore di attività e con gli indicatori generali del mercato. Per quanto riguarda le società di capitale, l'Eurocredit combina i sei indicatori con gli indici di redditività e gli indici finanziari.

Secondo la società Eurocredit, i soli indici di redditività non sono più sufficienti a determinare il rischio di credito; oggi, per assegnare un pun–teggio di rischio, soprattutto nel settore commerciale, non si può più

prescindere dal settore merceologico, dall'andamento del mercato e dal ciclo di vita dei beni prodotti. Il punteggio di rischio aumenta quando un'azienda adempie ai propri pagamenti con tempi più lunghi rispetto a quelli medi del settore di appartenenza; questa variabile rientra nella determinazione del rischio di credito perché essa potrebbe mascherare una situazione di cattiva gestione o, nei casi più gravi, una situazione di crisi che potrebbe irreversibilmente determinare una situazione d'insolvenza. L'anzianità e l'esperienza sono, invece, fattori favorevoli nella determinazione del punteggio; infatti, un'azienda che, data la sua esperienza, si rinnova continuamente, segue i flussi del mercato e cerca di conquistare maggiori quote di mercato, è da ritenersi affidabile e abbastanza sicura da intraprendere rapporti commerciali.

E' indubbiamente da ritenersi meno affidabile un'azienda che, sia pur senza rilevanti esposizioni creditizie, stenta a rinnovarsi e rimane statica ed ancorata nella sua nicchia di mercato. Secondo Eurocredit, per determinare tale punteggio è molto importante considerare il settore d'attività dell'azienda; infatti, è indubbiamente più rischioso esporsi con imprese che operano in settori poco dinamici, con una domanda in fase recessiva o addirittura in crisi. Tra i sei indicatori Eurocredit ha inserito anche il luogo dove l'azienda svolge la propria attività, in quanto si ritiene che la determinazione di tale punteggio possa essere influenzata anche dal livello di sviluppo dell'area industriale di riferimento, dallo sviluppo delle infrastrutture e dal rischio connesso alla criminalità.

Scopo del presente lavoro è la valutazione e l'affidabilità di un modello empirico di *credit scoring* nella determinazione dei rischi connessi agli affidamenti commerciali attraverso l'analisi discriminante e la regressione logistica. Per tale studio è stato utilizzato un database, composto da 1392 aziende pugliesi appartenenti al commercio all'ingrosso e al commercio al dettaglio fornito da Eurocredit, società leader nei sistemi di prevenzione e controllo del rischio di credito.

3. - Metodologie statistiche per l'analisi delle insolvenze aziendali

Le metodiche più utilizzate per l'analisi delle insolvenze aziendali possono essere classificate in tre principali categorie: l'analisi discriminante, la regressione logistica (*logit analysis*) e le reti neurali.

Nel presente lavoro, le elaborazioni sono state effettuate con l'impiego delle prime due tecniche statistiche.

3.1 – L'analisi discriminante

L'analisi discriminante è una tecnica multivariata di classificazione con–dotta, in genere, per definire un'adeguata modalità di assegnazione dei casi ai differenti gruppi in funzione di una serie di variabili[1]. I vari gruppi sono stati già definiti al momento dell'analisi, pertanto l'interesse è rivolto a definire un modello che consenta di assegnare un nuovo caso ad un gruppo predefinito, in funzione di un certo numero di variabili. Per realizzare un modello corretto bisogna affrontare due tipi di problemi: il primo che si può definire "di carattere classificatorio", riguardante la formazione dei gruppi (stima della funzione discriminante generata dai valori delle varia–bili osservate sulle unità che costituiscono i diversi gruppi); il secondo, di natura predittiva, riguardante l'assegnazione di una nuova unità ad uno dei gruppi.

In campo economico-finanziario si ricorre all'analisi discriminante per il bisogno di segmentare la clientela e per prevedere se una determinata azien–da può essere a rischio di insolvenza.

Le fasi principali per sviluppare in maniera adeguata una funzione discriminante, che sia capace di classificare le osservazioni in modo soddisfacente, sono essenzialmente sei: definizione degli obiettivi, progettazione dell'indagine, verifica delle ipotesi di base per la costruzione della regola di classificazione, stima dei coefficienti delle funzioni discriminanti, interpretazione dei risultati e validazione dei risultati (Brasini e Tassinari, 2002).

Nella prima fase si prevede che vengano indicati con precisione gli obiettivi, invece, nella seconda si mette a punto il progetto e l'obiettivo dell'indagine stessa; perciò, per utilizzare al meglio un modello di analisi discriminante, è necessario tener conto dei criteri di selezione della variabile dipendente, di quelle indipendenti e della numerosità del campione, necessaria per la stima delle funzioni discriminanti delle variabili. E' importante fissare il numero delle modalità della variabile dipendente, in modo da rendere i vari gruppi mutuamente esclusivi ed esaustivi tra loro, al fine di non creare dubbi e di facilitare l'assegnazione di ciascun caso ad uno solo dei gruppi. Per procedere alla selezione delle variabili indipendenti, si deve tener conto del modello teorico che costituisce il fondamento dell'indagine, e delle conoscenze empiriche del caso in esame e ovviamente di un qualche legame con la variabile dipendente.

[1] L'analisi discriminante (De Helguero, 1909; Fisher, 1936; Rao, 1952) interviene nell'attribuzione, a uno fra più gruppi multivariati e transvarianti, di ogni nuovo elemento individuale che si aggiunga all'insieme osservato. Cfr. Scardovi, 1998.

Dopo aver determinato la variabile dipendente e quelle esplicative si pone l'obiettivo di individuare in forma di equazione la combinazione esistente tra le variabili in esame. Tale equazione deve essere quella che meglio discrimina tra i gruppi, definiti a priori, tale che vi sia la massima variabilità tra i gruppi e la minima variabilità all'interno (Delvecchio, 2010)**.**

La funzione discriminante impiegata ha equazione:

$$Z = w_0 + w_1X_1 + w_2X_2 + \ldots\ldots\ldots + w_pX_p$$

dove:

Z = variabile dipendente;

X_k= generica variabile indipendente, con k = 1,2,3….p;

w_k= è il coefficiente di ponderazione associato alla variabile X_k.

La variabile dipendente Z, come si è già detto, assume il carattere di punteggio discriminante, ottenuto sommando i prodotti delle p variabili indipendenti per i rispettivi coefficienti di ponderazione, i quali vengono determinati mediante un conveniente metodo statistico di adattamento.

Una volta determinata l'equazione di discriminazione, è importante effettuare un controllo delle ipotesi su cui si deve fondare la tecnica; a tal proposito, è necessario soddisfare alcune condizioni: le variabili esplicative devono essere indipendenti tra loro e si devono distribuire congiuntamente come una variabile normale multivariata ed, infine, i gruppi, definiti dalle modalità della variabile dipendente, devono risultare omoschedastici (ovvero devono avere matrici di varianze e covarianze uguali).

Dopo aver stimato la funzione, occorre valutare il livello di significatività, generalmente fissato ad un valore pari a 0,05. Quando i gruppi sono più di due, oltre la significatività statistica della capacità discriminare fra tutti i gruppi, va valutata anche quella relativa alle singole funzioni. È importante ricordare che nell'analisi discriminante bisogna stimare tante funzioni quanti sono i gruppi tranne uno. Dopo la stima, il primo passo da compiere consiste nel calcolare i punteggi discriminanti Z, calcolati rispetto ad ogni unità statistica. Si considerano simili le unità statistiche che assumono analoghi valori dei punteggi Z; tale somiglianza dipende, essenzialmente, dal fatto che le osservazioni presentino valori molto vicini tra loro e, quindi, le variabili, inserite nell'equazione, assumano valori molto simili con pesi uguali.

Un buon criterio per valutare l'adattamento globale del modello ai dati si ottiene determinando le differenze tra i casi di ogni gruppo, espresso nei termini dei punteggi Z. Il successo dell'analisi si basa, quindi, sulla capacità di definire una funzione che produca come risultato differenze significative tra i centroidi; queste ul-

time, in genere, vengono misurate mediante la distanza di Mahalanobis[2], espressa attraverso la formula seguente:

$$D_{ij} = (\boldsymbol{X}_i - \boldsymbol{X}_j)' (\boldsymbol{X}_i - \boldsymbol{X}_j)/\boldsymbol{S}^2.$$

Nell'analisi discriminante la variabile dipendente non è metrica, quindi per valutare l'accuratezza previsiva del modello non è possibile ricorrere ad una misura come R^2, usata generalmente nella regressione multipla. Prima di effettuare il controllo sul grado di accuratezza, è necessario costruire delle matrici di classificazione e, poi, procedere al calcolo del tasso di corretta classificazione delle osservazioni. E necessario calcolare il valore soglia, ovvero, quel valore rispetto al quale viene confrontato il punteggio di ciascuna osservazione per determinare in quale gruppo deve essere classificata. La formula di calcolo di questo valore soglia o valore critico *Z* varia in funzione dell'uguaglianza o della diversità nell'ampiezza dei gruppi.

Quando i gruppi hanno uguale dimensione, il valore soglia che meglio discrimina tra due gruppi A e B, ad esempio, è definito da:

$$Z = \frac{Z_a + Z_b}{2}$$

dove:

Z = punteggio critico per la separazione tra i gruppi .

Z_a = il centroide del gruppo A.

Z_b = il centroide del gruppo B.

Se i due gruppi A e B, invece, hanno dimensioni diverse, si considera come punteggio ottimo per la separazione tra i due, la media ponderata dei centroidi, ovvero:

$$Z = \frac{N_a Z_a + N_b Z_b}{N_a + N_b}$$

dove, alle posizioni prima espresse, si aggiungono le seguenti:

N_a = numerosità del gruppo A.

N_b = numerosità del gruppo B.

Il valore soglia discriminante deve tener conto anche degli oneri derivanti da un'errata classificazione; se questi ultimi sono approssimativamente uguali per

[2] Metodo preferibile perché non impone che le variabili siano standardizzate, ed essendo espresso in termini di unità di variazione elimina gli effetti delle ridondanze, cosa che invece non accade usando la consueta distanza euclidea.

tutti i gruppi, il valore critico ottimo è quello che classifica in modo errato il minor numero possibile di unità rispetto a tutti i gruppi; invece, quando gli oneri sono diversi, il valore soglia ottimo sarà quello che rende minimi i costi dell'errata classificazione.

Per confermare la validità della funzione discriminante usando le matrici di classificazione, si può dividere casualmente il campione in due parti; in questo modo, si usa una delle due parti per determinare la regola discriminante e l'altra per stimare e verificare la stessa funzione discriminante. Per misurare l'accuratezza predittiva della funzione discriminante, inoltre, si potrebbe seguire una strada abbastanza pratica; fissare il tasso di corretta classificazione, indicato come quel valore che sia superiore di almeno un quarto della percentuale di osservazioni corrette classificate per effetto del caso.

Dopo aver svolto tutte le procedure atte a verificare la significatività delle funzioni discriminanti, si passa all'interpretazione dei risultati, evidenziandone il contributo alla separazione tra i gruppi, apportato da ciascuna variabile indipendente, giungendo alla loro convalida.

3.2 – La regressione logistica

La regressione logistica è un modello lineare facilmente interpretabile, che permette di studiare l'andamento di una variabile dipendente qualitativa. In questo modello le variabili esplicative sono combinate secondo una funzione lineare, utilizzata per stimare il logaritmo del rapporto tra la probabilità che un evento accada e la probabilità che lo stesso non accada. Tale logaritmo può essere espresso nel modo seguente:

$$\log\frac{P_j}{1-P_j} = \log(R_j) = w_0 + w_1 X_{1j} + w_2 X_{2j} + \ldots\ldots\ldots + w_p X_{pj},$$

(con j = 1,2,3,….., n), derivante dalla seguente funzione di distribuzione cumulativa di tipo logistico

$$\mathrm{Prob}\left(\mathrm{y}=1\right) = \frac{\mathrm{e}^{\mathrm{w}_0+\mathrm{w}_1\mathrm{X}_{1\mathrm{j}}+\ldots\ldots\ldots+\mathrm{w}_\mathrm{p}\mathrm{X}_\mathrm{p}}}{1+\mathrm{e}^{\mathrm{w}_0+\mathrm{w}_1\mathrm{X}_{1\mathrm{j}}+\ldots\ldots\ldots+\mathrm{w}_\mathrm{p}\mathrm{X}_\mathrm{p}}}$$

Tale logaritmo consente di prevedere l'appartenenza di un'osservazione ad un determinato gruppo definito a priori. Più specificatamente, per ogni osservazione, viene stimata la probabilità di appartenere ad una delle due categorie definite dalla variabile dipendente, dato il valore assunto dalle variabili esplicative; l'osser–

vazione viene quindi collocata nel gruppo con la più elevata probabilità stimata. Quanto più rilevanti e significative sono le variabili esplicative utilizzate, tanto più precisa è l'assegnazione operata dall'equazione logit sulla base delle variabili esplicative stesse; la verifica di tale significatività avviene tramite opportuni test, quali il test G^2 (di massima verosimiglianza) o il test di Wald. La regressione logistica è più vantaggiosa dell'analisi discriminante per la minore gravosità delle ipotesi precedentemente espresse nel primo modello; infatti, in questo modello, l'unica condizione per l'applicabilità è che per ogni variabile esplicativa le osservazioni siano indipendenti. Per tali motivi la regressione logistica risulta essere più robusta e più facile da applicare, dato che non si basa su ipotesi di normalità e di uniformità delle matrici di varianza e covarianza nei gruppi (Delvecchio, 2010; Ecchia, 1996).

4. - Valutazione e analisi del modello Eurocredit

L'obiettivo iniziale dell'analisi è stato quello di valutare l'affidabilità del modello Eurocredit nella determinazione dei rischi connessi agli affidamenti commerciali. La verifica è stata eseguita sull'intero database fornito da Eurocredit. Il collettivo di aziende pugliesi esaminato riguarda aziende del commercio all'ingrosso ed al dettaglio con dati relativi al 2015; nello specifico sono state analizzate 1392 società, di cui 695 appartenenti al commercio all'ingrosso e 697 al dettaglio. Tali aziende del collettivo sono state inizialmente classificate secondo i cinque punteggi di rischio assegnati da Eurocredit e così ripartite: aziende con un rischio molto elevato sono risultate in totale il 15,9%, quelle con un rischio elevato il 3,9%, quelle con un rischio leggermente elevato il 19,8%, quelle con un rischio normale il 42,4% e quelle con un rischio debole il 18,0% (Tabella 1).

La correttezza dello *score* assegnato a ciascuna azienda del collettivo esaminato è stata verificata mediante l'analisi discriminante multivariata lineare. Questo pri-

Tabella1. Distribuzione delle aziende classificate secondo il punteggio di rischio Eurocredit, per tipologia commerciale.

	Punteggio di rischio Eurocredit *(score)*					
Tipologia Commerciale	Molto elevato *score 1*	Elevato *score 2*	Legg. elevato *score 3*	Normale *score 4*	Debole *score 5*	Tot.
Commercio all'ingrosso	111	21	134	297	132	695
Commercio al dettaglio	115	30	136	292	124	697
Totale	226	51	270	589	256	1392
Distribuz. percentuale	*15,9*	*3,9*	*19,8*	*42,4*	*18,0*	*100,0*

mo passo ha avuto essenzialmente lo scopo di verificare l'attendibilità dell'intero processo elaborato da Eurocredit. Tale modello è il risultato pratico di anni di lavoro e dell'esperienza acquisita nella determinazione dei rischi commerciali e non è determinato, invece, da analisi e metodologie anche statistiche. Nonostante tutto, a detta del management Eurocredit, i risultati operativi conseguiti in questi anni sono risultati soddisfacenti.

La verifica empirica, effettuata sull'intero *database*, è stata condotta analizzando separatamente il collettivo per tipologia commerciale. Sia nel collettivo di aziende del commercio all'ingrosso che in quello del commercio al dettaglio, i cinque punteggi di *score* sono stati impiegati come variabili discriminanti, mentre, come variabili indipendenti, sono state considerate quelle utilizzate dal management Eurocredit per calcolare il rischio nelle esposizioni commerciali; più precisamente, le variabili indipendenti considerate sono state: i protesti, i mancati pagamenti, il recupero crediti, la eventuale presenza di negatività ed il coefficiente di redditività, ottenuto dal rapporto tra il risultato di esercizio ed il fatturato. In questa analisi i rischi legati al territorio e al settore di attività non sono stati considerati, dato che, in entrambi i collettivi, le aziende operano nel territorio pugliese ed appartengono al settore del commercio.

Tutte le 695 aziende del commercio all'ingrosso sono state studiate attraverso l'analisi discriminante canonica, assumendo che le probabilità a priori siano tutte uguali per ognuno dei cinque gruppi. L'analisi effettuata ha evidenziato che le variabili indipendenti con un maggiore potere discriminante sono "i protesti" ed "il recupero crediti", il valore test delle quali supera quello teorico della distribuzione F al livello di significatività del 5%. Delle 695 unità inserite, 89 sono state escluse dall'analisi a causa di dati mancanti e per incompletezza di informazioni; quindi, complessivamente, la verifica è stata compiuta su un collettivo di 606 unità.

Analizzando la Tabella 2, si osserva che solo il 36,2% delle unità del *database* sono state classificate correttamente dal modello Eurocredit, mentre il 21,3% risultano classificate erroneamente. Da questo primo approccio si evidenzia una grossa area grigia nella quale si vanno a collocare il 42,5% delle unità.

La verifica sulle aziende del commercio al dettaglio è stata effettuata su 696 unità, anziché su 697, dato che una unità è stata esclusa per incompletezza di informazioni. Come per il collettivo di aziende del commercio all'ingrosso, anche in questo caso la verifica è stata effettuata con la medesima metodologia e con lo stesso procedimento; tra le variabili indipendenti (i protesti, il recupero crediti, i mancati pagamenti, le eventuali perdite ed il coefficiente risultato d'esercizio/fatturato) quelle con maggiore potere discriminante sono i protesti e i mancati pagamenti.

Tabella 2. Riclassificazione delle aziende secondo l'analisi discriminante

Classificazione	Aziende	%
(Commercio all'Ingrosso)		
Aziende classificate correttamente	219	36,2
Aziende non classificate correttamente	129	21,3
Incerte ("Area grigia")	258	42,5
Totale	606	100
(Commercio al dettaglio)		
Aziende classificate correttamente	166	23,8
Aziende non classificate correttamente	238	34,2
Incerte ("Area grigia")	292	42,0
Totale	696	100

La verifica effettuata ha evidenziato che solo il 23,8% (166 unità) delle unità sono state classificate correttamente, che il 34,2% (238 unità) sono state classificate erroneamente ed, infine, che il 42,0% (292 unità) sono state posizionate nell'area grigia di indecisione (Tabella2).

La Tabella 3 riporta il confronto tra la classificazione Eurocredit e quella ottenuta con l'analisi discriminante secondo i cinque punteggi; dalla stessa emerge che in entrambi i collettivi esaminati gli errori di classificazione sono risultati rilevanti nel raggruppamento con punteggio di rischio "score1", "score3" e per il commercio al dettaglio per lo "score 4". In particolare, per lo "score3" (rischio leggermente elevato), ad eccezione della sola unità per il commercio al dettaglio, non si registrano unità assegnate a tale gruppo; si tratta di una modalità di confine, e quindi di incertezza tra quelle adiacenti.

Per le aziende del commercio al dettaglio, si evidenziano forti anomalie nel raggruppamento con "score 1", dove secondo l'analisi statistica, su un totale di 114 unità classificate con un punteggio di rischio molto elevato, solo 16 di esse sono risultate tali e ben 84 unità hanno presentato, invece, un rischio molto basso .

La verifica effettuata ha, quindi, evidenziato le discrepanze con il modello Eurocredit; la nuova classificazione presenta una forte concentrazione delle aziende nei gruppi a rischio più basso.

In merito a detto modello, peraltro, è importante sottolineare il fatto che tutte le analisi sono state condotte analizzando ogni singola azienda in termini relativi, ossia confrontandola con le relative medie di settore; purtroppo, i risultati derivati dalla verifica impongono una revisione generale dell'intero processo, al fine di individuarne i punti deboli per migliorarlo ulteriormente.

Tabella 3. Distribuzione delle aziende di commercio (*database* Eurocredit), riclassificate secondo il rischio con analisi discriminante multivariata.

	Rischio con analisi discriminante						
Score Eurocredit	Molto Elevato 1	Elevato 2	Leggerm. elevato 3	Normale 4	Debole 5	Totale Aziende	Unità classificate correttamente (%)
			Commercio all'ingrosso				
1	**3**	2	-	7	7	19	15,8
2	-	**15**	-	3	3	21	71,4
3	3	37	-	36	58	134	-
4	3	12	-	**113**	165	293	38,6
5	1	-	-	50	**88**	139	63,3
Totale	10	66	-	209	321	606	
(%)	*1,6*	*10,9*	-	*34,5*	*53*	*100*	
			Commercio al dettaglio				
1	**16**	1	-	13	84	114	14,0
2	-	**18**	-	1	11	30	60,0
3	8	39	**1**	10	78	136	0,7
4	23	10	-	**38**	221	292	13,0
5	10	-	-	21	**93**	124	75,0
Totale	57	68	1	83	487	696	
(%)	*8,2*	*9,8*	*0,1*	*11,9*	*70,0*	*100*	

Infine, da una prima analisi condotta sul modello Eurocredit, appare necessario e fondamentale determinare un giusto equilibrio tra la componente casuale, prettamente statistica, e quella soggettiva, poiché l'influenza dell'operatore non può e, forse, non deve essere eliminata, ma risulta necessario ridurne il peso.

5. - Classificazione delle aziende in due aree di rischio

In seguito ai risultati ottenuti dalla verifica effettuata e data l'evidente area grigia, si è preferito raggruppare le aziende del collettivo in esame in due grandi aree di rischio: *rischio elevato* (score 1) e *rischio normale* (score 0). Tale passo è stato effettuato, sia per una maggiore sintesi dei risultati e sia per il fatto che le società raggruppate nelle tre classi di rischio elevato e nelle due di rischio normale presentano rispettivamente caratteristiche molto simili tra loro. Dall'ultimo raggruppamento effettuato, si evince che le aziende con un rischio elevato sono risultate il 35% del totale e quelle con un rischio normale il 65%.

Il database di aziende pugliesi è stato diviso in due parti: il *training set*, utilizzato per stimare il nuovo modello, ed il *validation set*, utilizzato per effettuare una valutazione sul modello stimato. Le metodologie impiegate sono state l'analisi di-

scriminante e la regressione logistica. Come già spiegato nel terzo paragrafo, la seconda metodologia è stata preferita alla prima per la sua maggiore flessibilità e robustezza, soprattutto in presenza di due soli gruppi. Con buoni risultati la regressione logistica ha permesso di mettere in relazione variabili discrete, di tipo dicotomico (0-1), ed un insieme di variabili esplicative.

Anche per la realizzazione di questa seconda analisi si è preferito separare, ancora una volta, le aziende del commercio all'ingrosso da quelle al dettaglio, basando tale decisione sul fatto che il volume d'affari dei due settori è general–mente diverso e gli stessi rischi assumono connotazioni differenti.

5.1 - Le aziende del commercio all'ingrosso

Dal collettivo di 606 aziende del commercio all'ingrosso è stato inizialmente estratto un *training set*, attraverso un campionamento casuale stratificato per tipologia commerciale e classi di fatturato; si è così ottenuto un campione bilanciato di aziende, costituito da 175 unità con "rischio normale" e 175 unità con "rischio elevato". Si è preferito costruire il modello su un insieme bilanciato, per evitare che un gruppo assuma un peso maggiore rispetto all'altro.

Gli indici utilizzati nello studio sono i seguenti:

X_1 = Indice di copertura delle immobilizzazioni;
X_2 = Grado di indebitamento a breve;
X_3 = Indice corrente;
X_4 = Indice di liquidità;
X_5 = R.O.S.;
X_6 = Giacenza media delle scorte;
X_7 = Durata media dei crediti verso i clienti;
X_8 = Durata media dei debiti verso i fornitori;
X_9 = Incidenza dei proventi (oneri) finanziari sul fatturato;
X_{10}= Incidenza del costo del personale sul fatturato;
X_{11}= R.O.E.;
X_{12}= R.O.I;
X_{13}= Rotazione del capitale investito (turnover);
X_{14}= Redditività delle vendite;
X_{15}= Rotazione del magazzino;
X_{16}= Costi del personale sui costi di produzione;
X_{17}= Indici di rigidità;
X_{18}= Indice di elasticità;
X_{19}= Indice di autonomia finanziaria;
X_{20}= Indice di protezione del capitale;
X_{21}= Protesti;
X_{22}= Mancati pagamenti;
X_{23}= Indice di disponibilità del magazzino;

X_{24}= Quoziente di consolidamento del passivo;
X_{25}= Onerosità media dei finanziamenti;
X_{26}= R.O.E. al lordo;
X_{27}= Tasso di indebitamento (leverage);
X_{28}= Grado di incidenza dei proventi (oneri) extra gestione;
X_{29}= Intensità di capitale;
X_{30}= Indice di liquidità totale;
X_{31}= Indice di liquidità immediata;
X_{32}= Grado di finanziamento;
X_{33}= Indice di indebitamento permanente;
X_{34}= Margine operativo lordo;
X_{35}= Risultato operativo;
X_{36}= Cash flow;
X_{37}= Capitale circolante netto;
X_{38}= Fatturato per dipendente;
X_{39}= Costo del lavoro per dipendente;
X_{40}= Coefficiente di inizio attività;
X_{41}= Fido con le banche.

Prima di procedere alla scelta degli indici da inserire nel modello sarebbe necessaria una verifica preliminare delle ipotesi sottintese dall'analisi discri–minante, ovvero la *multinormalità* e l'*omoschedasticità* (ossia l'uguaglianza tra le matrici di varianza e covarianza dei due gruppi d'imprese). Nelle applicazioni pratiche, tuttavia, si utilizza spesso il modello lineare anche in assenza di tali condizioni, cercando al massimo di verificare la normalità dei singoli indici attraverso il test di Kolmogorov-Smirnov; questo perché l'analisi discriminante lineare, oltre ad essere più semplice e facilmente interpretabile, si dimostra in genere abbastanza valida anche in situazioni che contraddicono le condizioni di applicabilità del modello (Forestieri, 1986).

In effetti, alcuni degli indici utilizzati risultano collineari, manifestando tra loro una evidente ed inevitabile dipendenza economico-finanziaria, e notevolmente asimmetrici (e pertanto non normali). Tali indici sono stati comunque inseriti nell'analisi per il loro significativo apporto all'analisi economica, nella considerazione che la previsione delle insolvenze si basa non solo su metodologie statistiche ma anche su considerazioni economico-finanziarie (Summo, 1999). In particolare, si evidenzia che l'eventuale mancanza di omoschedasticità dovrebbe avere come conseguenza la non linearità della funzione discriminante, per cui la linea di demarcazione delle nuvole di punti relativi agli insiemi delle aziende dei due gruppi non sarebbe una retta, ma una curva, comunque determinata in modo da ridurre al minimo la sovrapposizione dei due gruppi[3].

[3] Si veda Forestieri, 1986; Laviola e Trapanese, 1997; Lachenbruch, 1979.

Tabella4 - Distribuzione delle aziende del commercio all'ingrosso riclassificate secondo le due metodologie considerate (valori percentuali).

	Rischio Normale	Rischio Elevato	Totale
	(Analisi discriminante)		
Rischio Normale	***85,1***	*14,9*	*100,0*
Rischio Elevato	*44,5*	***55,5***	*100,0*
Totale	64,8	35,2	100,0
	(Regressione logistica)		
Rischio Normale	***82,3***	*17,7*	*100,0*
Rischio Elevato	*35,3*	***64,7***	*100,0*
Totale	*58,9*	*41,1*	*100,0*

La Tabella 4 riporta i risultati complessivi della classificazione delle aziende secondo l'analisi discriminante[4] e la regressione logistica per le due classi di rischio.

Dalla lettura delle percentuali posizionate lungo la diagonale principale si rilevano le percentuali di giusta classificazione tra quella effettuata empiricamente dalla Società Eurocredit (classificazione effettiva) e quelle che derivano dalla applicazione delle due metodiche utilizzate nel presente lavoro (classificazione prevista): 85,1% e 55,5% per l'analisi discriminante e 82,3% e 64,7% per il modello logistico. Le percentuali al di fuori della diagonale principale rappre–sentano le classificazioni errate.

Si fa osservare che dalla elaborazione è emerso un più elevato tasso com–plessivo di corretta classificazione con la regressione logistica 73,6% rispetto a quello ottenuto con l'analisi discriminante (70,4%) A tal proposito si fa osservare che la regressione logistica non poggia su assunzioni forti come quelle dell'analisi discriminante e risulta essere maggiormente flessibile e capace di adattarsi meglio agli stessi dati impiegati nell'analisi[5].

La selezione delle variabili esplicative, necessarie alla costruzione del modello di rischio mediante la regressione logistica, è stata effettuata attraverso il metodo

[4] Nell'analisi discriminante lineare le aziende del *training set* sono state studiate selezionando gli indicatori più significativi attraverso il metodo *stepwise*; dopo aver costruito la funzione di discriminazione, sono state analizzate le unità classificate correttamente e quelle non correttamente e si è determinato la capacità di classificazione complessiva. Nello specifico, l'analisi *stepwise* ha selezionato le seguenti variabili come quelle con la maggiore capacità di discriminazione, ovvero: l'*indice corrente*, il *margine operativo lordo*, il *coefficiente d'inizio attività*, i *protesti* ed il *fido*.

[5] Nell'analisi discriminante multivariata, invero, se le variabili indipendenti fossero normalmente distribuite, le stime prodotte sarebbero le più efficienti, in quanto nessun'altra metodologia potrebbe determinare minori errori di classificazione; cfr. Maddala, 1983.

Tabella 5. Modello di regressione logistica finale per le aziende di commercio all'ingrosso e test di Wald per le variabili esplicative.

Variabili	Coeff. β_i	Test di Wald	Significatività
X_6 = Giacenza media delle scorte	0,0010	3,7896	0,0516
X_{21}= Protesti	-3,2559	24,0060	0,0000
X_{36}= Cash flow	-0,0007	3,6828	0,0550
X_{40}= Coefficiente di inizio attività	-0,3850	10,8677	0,0010
X_{41}= Fido con le banche	-1,5440	21,7877	0,0000
Costante	14,3497	28,0059	0,0000

stepwise ad inserimento, con test di massima verosimiglianza per la determinazione del modello finale, ossia quello che ottimizza il criterio di selezione[6].

Tutte le variabili, inserite ad una ad una nel processo di elaborazione sulla base del test G^2 per valutare il miglioramento del modello secondo la statistica di massima verosimiglianza, sono poi analizzate sulla base della statistica di Wald[7], impiegata per valutare se aggiungere una nuova variabile nel modello in costruzione (Tabella 5). Le variabili esplicative selezionate dal modello sono state: la *giacenza media delle scorte*, il *cash-flow*, il *coefficiente di inizio attività*, i *protesti* ed i *fidi.*

Il passo successivo è stato quello di costruire una funzione logistica, capace di discriminare le aziende del commercio all'ingrosso nelle diverse classi di rischio. La funzione logistica, di seguito riportata, è stata sviluppata considerando come variabili esplicative quelle selezionate dal modello descritto in Tabella 5.

$$\text{Prob}(y_i = 1) = \frac{e^{14,3497+0,0010X_6-3,2559X_{21}-0,0007X_{36}-0,385X_{40}-1,544X_{41}}}{1+e^{14,3497+0,0010X_6-3,2559X_{21}-0,0007X_{36}-0,385X_{40}-1,544X_{41}}}.$$

Il modello di regressione così ottenuto mette in risalto che i coefficienti negativi delle variabili cash-flow, inizio attività, protesti e fido contribuiscono positivamente nella riduzione del rischio di insolvenza; infatti, un elevato flusso di cassa, gli anni di esercizio di un'attività commerciale, la non presenza di protesti ed una flessibile linea di credito con il proprio istituto creditizio sono tutti elementi che contribuiscono a ridurre i rischi e, quindi, ad abbassare la probabilità che si possa verificare un mancato pagamento nell'affidamento commerciale.

[6] L'applicazione della procedura è stata effettuata con software SPSS; per quanto riguarda i riferimenti metodologici della stessa, si rimanda a testi didattici come, ad es., Delvecchio (2015); Fabbris (1997).

[7] Il test di Wald, espresso dalla relazione $W_e(y) = n(\hat{\theta} - \theta_0)^2 / (\hat{\theta})$, è impiegato per verificare il livello di approssimazione dei parametri di massima verosimiglianza stimati, misurando quindi lo scarto tra la frequenza osservata θ e quella attesa $\theta 0$, opportunamente standardizzate; in questa forma, esso si distribuisce asintoticamente secondo un χ^2.

Al contrario, la variabile relativa alla giacenza media delle scorte presenta un coefficiente positivo che, nonostante il valore ne riduce notevolmente il peso, contribuisce ad accrescere i rischi. Infatti, l'elevata presenza di scorte di magazzino ed un lungo periodo di giacenza della merce sono indubbiamente elementi negativi che evidenziano un mancato ricambio della merce e conseguentemente un basso cash-flow. Nette differenze si evidenziano, inoltre, analizzando la media e la deviazione standard relative alle variabili inserite nel modello appena costruito (Tabella 6) e calcolate distintamente per i due gruppi di aziende.

Tabella 6. Distribuzione della media e della deviazione standard relativa alle variabili significative per le aziende di commercio all'ingrosso.

Gruppo	Variabili	Media	Deviazione std.
Aziende a Rischio normale (score 0)	Giacenza media scorte	130,76	215,66
	Cash-flow	78,06	539,67
	Inizio attività	3,52	1,05
	Protesti	0,398	0,02
	Fido	0,9081	1,60
Aziende a Rischio elevato (score 1)	Giacenza media scorte	168,35	249,82
	Cash-flow	44,76	148,59
	Inizio attività	3,13	1,16
	Protesti	0,351	0,08
	Fido	0,24	0,48

Le aziende a "rischio normale" evidenziano una giacenza media delle scorte in media più bassa rispetto a quelle a "rischio elevato", un cash-flow mediamente più alto e presentano una maggiore anzianità nell'esercizio dell'attività commerciale. In merito ai protesti, si osserva che tra le aziende del primo gruppo non si evidenziano segnalazioni di rilievo. Significativi sono i dati relativi al fido; infatti si fa notare come le aziende a fido più basso sono quelle più a rischio.

Dopo aver individuato la funzione logistica è risultato necessario realizzare delle classi di *rating*, mediante la suddivisione in classi di intervalli della distribuzione di probabilità; a tale proposito, sono state individuate quattro aree di rischio, due all'interno del rischio normale e due in quello elevato, ovvero: *rischio basso, normale, elevato e molto elevato*. Tra le due aree di rischio è stata anche inserita un'*area di incertezza* (comunemente chiamata area grigia), con lo scopo di raggruppare tutte quelle aziende che presentano un probabilità di insolvenza compresa tra 0,45 e 0,55 e che, in particolare, presentano caratteristiche di rischio comuni ad entrambe le due classi (Tabella 7).

L'area d'incertezza costituisce sicuramente quella con le maggiori difficoltà di classificazione per la tipologia di aziende che vengono raggruppate in essa.

Tabella 7. Distribuzione degli intervalli di probabilità nelle quattro classi di rischio.

Tipologia di rischio	Classi di rischio	Probabilità
Rischio Normale	Basso	0≤P< 0.25
	Normale	0,25≤P< 0,45
Area di incertezza		0,45≤P< 0,55
Rischio Elevato	Elevato	0,55≤P< 0,75
	Molto elevato	0,75≤P< 1

È, comunque, doveroso sottolineare che, nel monitoraggio costante effettuato sulle aziende affidate, essa deve essere considerata anche come l'area che segnala l'aggravarsi o il miglioramento delle condizioni economico-finanziarie di una particolare azienda, preannunciando quindi in anticipo una probabile futura insolvenza o, alternativamente, una riduzione generale dei rischi connessi agli affidamenti verso la stessa azienda in precedenza classificata a rischio elevato.

Il passo successivo è stato quello di testare e validare il modello appena scelto attraverso l'individuazione di un nuovo collettivo di aziende; è stato quindi individuato un *validation set,* costituito da un insieme di 256 aziende, composto da 154 aziende, inizialmente classificate da Eurocredit come aziende a "rischio normale" e 102 a "rischio elevato". Su tale insieme di controllo è stata operata una riclassificazione secondo le cinque classi di rischio: *rischio basso, normale, area di incertezza, rischio elevato e molto elevato* (Tabella 7). Il validation set è stato esaminato mediante la funzione ottenuta dal modello di regressione logistica, calcolando per ciascuna azienda analizzata la probabilità di insolvenza. Si è pervenuti al nuovo punteggio di rischio, inserendo nella funzione i dati *relativi alla giacenza media delle scorte, al cash-flow, al coefficiente di inizio attività, ai protesti e al fido* (variabili indipendenti selezionate dal modello di regressione logistica).

Tabella 8. Distribuzione delle aziende del *validation set* del commercio all'ingrosso riclassificate secondo la regressione logistica (percentuali)

	Classi di rischio					Totale
	Basso	Normale	Incertezza	Elevato	Molto Elevato	
Basso	**57,6**	13,6	12,1	1,5	15,2	100,0
Normale	8,0	**72,7**	8,0	4,5	6,8	100,0
Incertezza	5,6	27,8	**46,3**	3,7	16,7	100,0
Elevato	-	9,5	14,3	**52,4**	23,8	100,0
Molto elevato	-	-	18,5	-	**81,5**	100,0
Totale	18,8	35,2	18,8	7,0	20,2	100,0

I risultati di questa nuova classificazione sono stati riportati nella Tabella 8; dalla stessa si evincono le percentuali di giusta classificazione, lungo la diagonale

principale, tra la iniziale classificazione Eurocredit e quella ottenuta con il modello di regressione logistica. Le più alte percentuali si riscontrano per le modalità *molto elevato e normale* (rispettivamente 81,5% e 72,7%), mentre intorno al 50% si posizionano le altre modalità riferite alla classi di rischio ad eccezione dell'area d'incertezza che rappresenta quella con più bassa percentuale di classificazione essendo quest'ultima quella con caratteristiche comuni alle modalità adiacenti.

5.2 - Le aziende del commercio al dettaglio.

Come per le aziende del commercio all'ingrosso, anche per quelle al dettaglio si è seguito lo stesso procedimento per la costruzione di un modello di rischio. Anche in questo caso, il database di 696 aziende pugliesi del commercio al dettaglio è stato diviso in training set e validation set; da tutte le unità è stato, quindi, selezionato un campione bilanciato di 400 aziende, composto da 200 aziende inizialmente classificate a "rischio normale" e 200 a "rischio elevato"[8]. In questo secondo studio sono state usate nuovamente entrambe le due metodologie in esame: l'analisi discriminante e la regressione logistica.

Nell'analisi discriminante lineare le aziende del *training set*, ancora una volta vengono studiate selezionando gli indicatori più significativi attraverso il metodo *stepwise*. Nello specifico l'analisi ha individuato le seguenti variabili come quelle con la maggiore capacità discriminante: il *fido*, la *durata media dei debiti verso i fornitori*, la *rotazione del capitale investito*, l'*indice di rigidità*, l'*indice di autonomia finanziaria*, l'*indice di liquidità immediata*, il *coefficiente di inizio attività*, i *protesti* ed, infine, i *mancati pagamenti*.

I risultati della procedura classificatoria sono presentati in forma matriciale nella Tabella 9.

Tabella9. Distribuzione delle aziende del commercio al dettaglio riclassificate secondo le due metodologie considerate (valori percentuali)

	Rischio Normale	Rischio Elevato	Totale
	(Analisi discriminante)		
Rischio Normale	***79,4***	*20,6*	*100,0*
Rischio Elevato	*33,0*	***67,0***	*100,0*
Totale	56,0	44,0	100,0
	(Regressione logistica)		
Rischio Normale	***73,6***	*26,4*	*100,0*
Rischio Elevato	*11,5*	***88,5***	*100,0*
Totale	*42,5*	*57,5*	*100,0*

[8] Anche questo *training set* è stato costruito con un campionamento casuale stratificato, classificando le unità per tipologia commerciale e per classi di fatturato.

I valori sulla diagonale principale corrispondono al numero di aziende attribuite correttamente; al contrario al di fuori della diagonale principale tali valori rappresentano le classificazioni errate. Nel dettaglio si nota una capacità del modello discriminante di classificare correttamente il 79,4% del gruppo di aziende a rischio normale ed una capacità del 67% nella classificazione di quelle a rischio elevato; per il modello logistico dette percentuali sono pari a 73,6% e 88,5% rispettivamente per il rischio normale e quello elevato (Tab. 9).

Anche in questo caso, in complesso, dalla analisi dei dati è emerso che il modello logistico porti a classificare correttamente una più elevata percentuale di aziende.

Dati i migliori risultati ottenuti dal modello logistico e la maggiore flessibilità, oltre alla migliore capacità di adattamento ai dati disponibili, si è preferito nuovamente applicare tale modello al data base delle aziende del commercio al dettaglio. Nel modello finale, ancora una volta definito attraverso una procedura stepwise con test di massima verosimiglianza, le variabili esplicative significative sono: *l'indice di rotazione del capitale investito, il coefficiente di inizio attività, i protesti, i mancati pagamenti e i fidi* (Tabella10).

Tabella 10. Modello di regressione logistica finale per le aziende di commercio al dettaglio e test di Wald per le variabili esplicative.

Variabili	Coeff. β_i	Test di Wald	Significatività
X_{13}= Rotazione del capitale investito	0,2284	38,6380	0,000
X_{21}= Protesti	-3,4790	22,8376	0,0000
X_{22}= Mancati pagamenti	-0,9742	5,7067	0,0169
X_{40}= Coefficiente di inizio attività	-0,2783	8,2823	0,0000
X_{41}= Fido con le banche	-2,7388	3,0908	0,0787
Costante	18,9000	24,2246	0,0000

Il modello logistico relativo alle aziende di commercio al dettaglio è quindi

$$\text{Prob}(y_i = 1) = \frac{e^{18,90+0,2284X_{13}-3,4790X_{21}-0,9742X_{22}-0,2783X_{40}-2,7388X_{41}}}{1+e^{18,90+0,2284X_{13}-3,4790X_{21}-0,9742X_{22}-0,2783X_{40}-2,7388X_{41}}}$$

Analizzando il modello così ottenuto, si evidenzia che i coefficienti negativi delle variabili *inizio attività, protesti, mancati pagamenti e fido* contribuiscono ad accrescere o diminuire il rischio d'insolvenza, accrescendo o diminuendo lo stesso denominatore. Un'attività commerciale che svolge la propria attività da molti anni, che non presenta protesti o mancati pagamenti e che ha una flessibile linea di credi-

to con la propria banca, presenterà sicuramente un punteggio di rischio molto basso negli affidamenti commerciali[9].

In particolare, analizzando la media e la deviazione standard calcolate per ciascuna classe di rischio (Tabella 11), si osserva che il *turnover* è in media più basso nel gruppo di aziende a rischio normale e che i protesti, calcolati in termini di numeri di anni trascorsi dal verificarsi del fenomeno, presentano un valore medio più alto tra le aziende a rischio normale (in media risultano essere trascorsi 10 anni della rilevazione di un protesto nelle aziende a rischio normale, mentre poco più di tre anni tra le aziende a rischio elevato).

Tabella11 - Distribuzione della media e della deviazione standard relativa alle variabili significative per le aziende di commercio al dettaglio.

Gruppo	Variabili	Media	Deviazione std.
Aziende a rischio normale (score 0)	Rotazione capitale investito	1,65	1,47
	Inizio attività	3,57	0,94
	Protesti	9,98	0,12
	Mancati pagamenti	3,98	0,14
	Fido	0,96	1,66
Aziende a rischio elevato (score 1)	Rotazione capitale investito	2,10	1,96
	Inizio attività	3,23	1,12
	Protesti	3,69	0,64
	Mancati pagamenti	3,86	0,58
	Fido	0,16	0,37

In definitiva, anche in questo secondo modello si evidenzia una netta differenza tra i due gruppi di aziende (a rischio normale ed a rischio elevato) per quanto riguarda l'ammontare medio del fido bancario.

Per verificare l'attendibilità del modello ottenuto, è stato costruito un *validation set* formato da 297 aziende (costituite da 217 classificate inizialmente a "rischio normale" e 80 a "rischio elevato"). Per tutte queste aziende è stata operata una riclassificazione sulla base della funzione logistica, sopra riportata, e sulla base delle cinque classi di rischio (rischio basso, normale, area d'incertezza, rischio elevato e molto elevato). Per ciascuna azienda del *validation set* è stata calcolata la probabilità di insolvenza e, quindi, il relativo punteggio di rischio sulla base delle sole variabili esplicative inserite nel modello statistico (Tabella12).

[9] Comparando i due modelli (discriminante e logistico), si nota che da entrambi sono state selezionate le variabili esplicative: *fidi*, *protesti*, *mancati pagamenti*, *coefficiente di inizio attività* e *indice di rotazione del capitale investito*.

Le aziende inserite nel *validation set*, già classificate da Eurocredit, risultano essere riclassificate in maniera sostanzialmente diversa rispetto alla analoga classificazione riscontrata per le aziende del comparto all'ingrosso; infatti, il nuovo raggruppamento assegna una percentuale di giusta classificazione del 39,1% per la modalità a "basso rischio", del 22,4% per quella "normale", 46,7% per "l'area di incertezza" e percentuali più alte nelle restanti due modalità: "più elevato" 69,8% e "molto elevato" 68,2%.

Tabella 12. *Distribuzione delle aziende del* validation set *del commercio al dettaglio riclassificate secondo la regressione logistica (valori percentuali).*

	Classi di rischio					Totale
	Basso	Normale	Incertezza	Elevato	Molto elevato	
Basso	**39,1**	24,5	17,3	15,5	3,6	100,0
Normale	15,9	**22,4**	21,5	35,5	4,7	100,0
Incertezza	26,7	20,0	**46,7**	6,7	-	100,0
Elevato	2,3	4,7	2,3	**69,8**	20,9	100,0
Molto elevato	-	-	9,1	22,7	**68,2**	100,0
Totale	21,9	18,9	17,5	30,6	11,1	100,0

6 - Sintesi e considerazioni conclusive

Le analisi di bilancio vengono effettuate per ottenere informazioni sulla situazione finanziaria ed economica dell'azienda; queste informazioni servono per valutare lo stato di salute e prevedere un eventuale stato di crisi aziendale. Tuttavia risulta alquanto problematico tenere conto in un giudizio sintetico delle indicazioni alquanto contraddittorie che emergono dagli indici di bilancio presi singolarmente. I modelli di previsione delle insolvenze forniscono a tal riguardo una risposta che può essere considerata abbastanza soddisfacente. Nel presente lavoro sono state utilizzate sia l'analisi discriminante che la regressione logistica.

Dai risultati ottenuti è emerso innanzitutto che le variabili con il maggiore potere discriminante sono risultate: il fido, il turnover, i protesti, i mancati pagamenti, il cash-flow ed il coefficiente d'inizio attività. In entrambi gli studi, si è ritenuto necessario inserire il coefficiente d'inizio attività perché nel commercio l'anzianità d'esercizio può essere considerata, in molti casi, un segno di sicurezza e d'affidabilità.

I riscontri empirici hanno rilevato che tra le due metodologie quella che ha presentato la maggiore efficacia e la maggiore adattabilità, soprattutto in presenza di

due sole classi, è risultata la regressione logistica. L'analisi discriminante lineare, pur in assenza delle due condizioni di base, è stata in ogni modo utilizzata perché nelle applicazioni empiriche consente comunque di conseguire risultati accettabili ed affidabili.

I modelli elaborati presentano una buona flessibilità e si adattano pienamente ai dati; inoltre, essi riescono a discriminare in modo soddisfacente le aziende a "rischio normale" da quelle a "rischio elevato", pur presentando un'evidente area d'incertezza, nella quale sono sostanzialmente inserite tutte quelle aziende che presentano un punteggio di rischio medio.

Questi modelli rappresentano un utile strumento per verificare velocemente le condizioni economiche–finanziarie delle aziende da affidare e consentono inoltre di poter assegnare un punteggio di rischio. Tra i due modelli elaborati, quello realizzato per il commercio al dettaglio risulta essere più robusto e, oltre a classificare correttamente il maggior numero di aziende (81,1%), riesce a classificare il 73,6% di aziende a rischio normale e circa l' 88,5% di quelle a rischio elevato.

Confrontando i due modelli statistici elaborati mediante la regressione logistica con quello empirico sviluppato da Eurocredit, si è riscontrato che tutti sono basati pressoché sulle stesse variabili esplicative. Come già osservato in precedenza, il modello empirico Eurocredit, pur non essendo sviluppato attraverso metodologie prettamente statistiche, manifesta alcuni punti di forza, dato che in esso il punteggio di rischio si determina su poche variabili realmente discriminanti, si analizza sempre ciascuna azienda in termini relativi confrontandola con il mercato in cui la stessa è inserita, si tiene conto non solo di variabili economico-finanziarie ma anche di variabili di tipo "ambientale" come sono i rischi connessi al territorio.

Le nuove riclassificazioni, elaborate sia con l'analisi discriminante multivariata che con la regressione logistica, hanno presentato una distribuzione complessivamente diversa da quella precedentemente effettuata da Eurocredit, mostrandone quindi una evidente discrepanza.

Per migliorare il modello Eurocredit sarebbe necessario rivedere le tabelle impiegate per assegnare alle varie variabili il relativo coefficiente di rischio o di rettifica, sarebbe necessario un continuo monitoraggio sull'andamento delle aziende del database ed, inoltre, sarebbe necessario determinare un giusto equilibrio tra la componente casuale, prettamente statistica, e quella soggettiva. E' indubbio che l'influenza dell'operatore non può e, forse, non deve essere eliminata, ma è necessario ridurne il proprio peso.

Note bibliografiche

Alberici A. (1986). *La previsione delle insolvenze aziendali. Profili teorici ed analisi empiriche*, Giuffrè Editore.

Airoldi G.; Brunetti G.; Coda V.(1994). *Economia Aziendale*, il Mulino, Bologna.

Appetiti S. (1984). *L'utilizzo dell'analisi discriminatoria per la previsione delle insolvenze: ipotesi e test per un'analisi dinamica*, in "Temi di discussione" n.27, Banca d'Italia.

BANCA D'ITALIA (2000). *Modelli per la gestione del rischio di credito: i rating interni*, in "Temi di discussione", aprile, Roma.

Box, G. E. P. (1949). *A general distribution theory for a class of likelihood criteria,* in "Biometria"*,* Vol.36.

Brasini, S.; Tassinari, F. (2002). *Lezioni di statistica aziendale*, Editrice Esculapio, Bologna.

Cavrini, G.; Mignani, S.; Soffritti, G.(2001). *Esercizi di analisi statistica multivariata risolti con SPSS per Windows*, Editrice Esculapio, Bologna

CENTRALE RISCHI (1998). *Alberi decisionali ed algoritmi genetici nell'analisi del rischio di insolvenza*, Milano.

Corigliano, R., (1998). *Rischio di credito e pricing dei prestiti bancari*, Editrice Bancaria, Roma.

de Helguero, F. (1909). Sulla rappresentazione analitica delle statistiche abnormali. *Atti del IV congresso internazionale dei matematici*, III, 288–299.

De Laurentis, G. (1998). *I processi di Rating ed i Modelli di scoring*, Bancaria Editrice, Roma.

De Laurentis, G. (1998). *La misurazione e la gestione del rischio di credito bancario*, Bancaria Editrice, Roma.

Delvecchio, F. (2010). *Statistica per l'analisi di dati multidimensionali*. Cleup, Padova.

Delvecchio, F. (2015). *Statistica per l'analisi dei fenomeni sociali*. Cleup, Padova.

Desario, V. (a cura di) (2000). *Modelli per la gestione del rischio di credito. I rating interni* in "Tematiche istituzionali", Banca d'Italia, Roma.

Dobson, A. J. (1990). *An Introduction to Generalized Linear Models*, Chapman & Hall Editrice.

Ecchia, S. (a cura di) (1996). *Il rischio di credito: metodologie avanzate di previsione delle insolvenze*, Giappichelli editrice, Torino.

EUROCREDIT (1999), *La Eurocredit nella new-economy* .

Fabbris, L. (1997). *Analisi Esplorativa di Dati Multidimensionali*, Mc Graw-Hill Editore, Milano.

Fisher R. A. (1936). The use of multiple measurements in taxonomic problems, *Annals of Eugenics*, 7: 179–188.

Forestieri, G. (1986). *La previsione delle insolvenze aziendali*. *Profili teorici ed analisi empiriche*, Giuffrè, Milano.

Gabbi, G. (1998). L'utilizzo delle reti neurali per la misurazione del rischio di credito. In: Sironi A. e Marsella M. (a cura di). *La misurazione e la gestione del rischio di credito. Modelli, strumenti e politiche*, Editrice Bancaria, Roma.

Green, W. H. (2000). *Econometric Analysis*, Prentice Hall Editrice, quarta edizione, New York.

McLachlan, G. J. (2004). *Discriminant Analysis and Statistical Pattern recognition*, John Wiley & Sons.

Laviola S.; Trapanese M. (1997). Previsione delle insolvenze delle imprese e qualità del credito bancario: un'analisi statistica. *Temi di discussione*, Servizio Studi della Banca d'Italia, n. 318, Roma.

Maddala, G. S. (1983). *Limited-Dependent and Qualitative Variable in Econometrics*, Cambridge University Press, Cambridge, MA.

Maiano, R. (2000). *La gestione del rischio di credito: esperienze e modelli nelle grandi banche italiane*, Edibank Roma.

Masciandaro, D.; Porta, A.(1998). *Le sofferenze bancarie in Italia*, Editrice Bancaria, Roma.

Mignani, S.; Montanari, A.(2001). *Appunti di analisi statistica multivariata*, Esculapio Editore, Bologna.

Morrison, D.F. (1976). *Metodi di analisi statistica multivariata*, Ambrosiana Editrice, Milano.

Rao, C. R. (1952). *Advanced Statistical Methods in Biomedical Research*, John Wiley & Sons, New York, NY.

Ricciardi, A. (1996). La previsione delle insolvenze nel credito al consumo mediante l'applicazione della tecnica del credit scoring. In: Ecchia S. (a cura di) *Il rischio di credito. Metodologie avanzate di previsione delle insolvenze*, Giappichelli, Torino.

Savona, P.; Sironi, A. (2000). *La gestione del rischio di credito: esperienze e modelli nelle grandi banche italiane*, Edibank.

Sironi, A.; Marsella, M. (a cura di) (1998). *La misurazione e la gestione del rischio di credito*, Editrice Bancaria, Roma.

Scardovi I. (1998). *Statistica applicata alle scienze sociali*, Enciclopedia delle scienze sociali, Treccani.

Summo, D. (1999). *Un modello empirico per l'analisi delle insolvenze aziendali*, Quaderno n.3, Dipartimento di Scienze Statistiche dell'Università degli Studi di Bari.

Vitali, O (1993). *Statistica per le Scienze Applicate*, vol. I e II, Cacucci Editore, Bari.

UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO *Metodi e analisi statistiche 2017*
DIPARTIMENTO DI ECONOMIA E FINANZA
ISBN: 978-88-6629-013-1
pp. 211-232

Dai conti economici ai conti satellite ambientali: Basilicata, un caso di studio

Agata Maria Madia Carucci[1], Flora Fullone[1], Giovanni Vannella[2]*

[1] *Istat, Ufficio territoriale della Basilicata*
[2] *Università degli studi di Bari, Dipartimento di Economia, management e diritto dell'impresa*

Riassunto: Con il presente articolo si descrivono le principali innovazioni introdotte dal SEC 2010 in merito ai conti satellite; si analizza il sistema produttivo lucano mettendolo in relazione con le prime risultanze ottenute da una sperimentazione dei conti satellite fatta su input della regione Basilicata, evidenziando le prime intercon–nessioni che è possibile individuare tra attività economiche e pressioni ambientali.

Keywords: Contabilità nazionale, Conti Regionali, Conti satellite ambientali, Conti dei flussi di materia, PIL, DMC, pressioni ambientali.

1. Introduzione

È da tempo palese come le politiche economiche liberiste che ritenevano che la libera fluttuazione dei prezzi avrebbe garantito un equilibrio di tutti i sistemi compreso quello ambientale, abbiano manifestato i loro evidenti limiti; limiti dovuti ad una pluralità di fattori tra cui si possono evidenziare la forte anelasticità che l'offerta dei beni ambientali presenta rispetto al loro prezzo, la non perfetta coincidenza territoriale tra territori inquinanti ed aree inquinate, la difficile determinazione del valore economico dell'ambiente.

Si aggiunga poi come raramente in passato siano state programmate analisi sul territorio che, partendo da valide rilevazioni aventi continuità spazio-temporale, permettessero di esprimere con agilità lo stato dell'inquinamento nei vari ecosistemi.

* Autore corrispondente: giovanni.vannella@uniba.it
Il lavoro è frutto del lavoro congiunto dei tre autori, ma a A.M.M Carucci sono attribuiti i paragrafi 2.1, 2.2, 3, 3.1, a F. Fullone i paragrafi 3.2, 3.2.1, 3.2.2, 3.2.3, 4 ed a G. Vannella i paragrafi 1, 2 e 5.

La forte interconnessione tra attività economiche (o più in generale antropiche) e l'utilizzo delle risorse ambientali ha quindi visto il proliferare di progetti che permettessero di integrare gli schemi di contabilità nazionale con schemi di contabilità ambientale.

Tra questi spicca il SEEA (System of *Environmental - Economic Accounting*) promosso dall'Ufficio Statistico dell'ONU, strutturato come un conto satellite in stretta relazione col SNA, nonché i vari progetti comunitari in cui l'Istat è impegnata tra cui si ricordano (Vannella, 2001, 2003):

- la NAMEA *(National Accounting Matrix including Environmental Accounts);*
- il SERIEE *(Système européen pour le rassemblement des informations économiques sur l'environnement);*
- i SIP *(Sectoral Infrastructure projects).*

Tali progetti, in ottemperanza al SEC, si stanno indirizzando anche nell'ottica della produzione dei conti regionali ed in tale ambito particolare rilievo assume il caso della Basilicata per diversi ordini di motivi.

In primo luogo in quanto, come si illustra di seguito, tale area rappresenta un unicum in Italia essendo caratterizzata da una distribuzione del valore aggiunto per attività economica completamente diversa rispetto alla distribuzione media del paese.

In secondo luogo in quanto tale area è caratterizzata da una forte presenza di attività agricole ed attività estrattive.

Infine la Basilicata è una delle poche regioni in Italia per la quale sia stato prodotto il Conto Satellite dei flussi di materia, progetto pilota attualmente inserito nel PSN per garantirne l'*"esportazione"* nelle altre regioni italiane.

2. Dai conti economici ai conti satellite: quali sono i conti principali e come sono inseriti nel SEC2010

Nel corso del settembre 2014, l'Istat ha diffuso i primi risultati della revisione completa dei conti nazionali programmata in occasione dell'introduzione del nuovo Sistema europeo dei conti (SEC 2010).

Se da un lato i nuovi conti incorporano tutte le innovazioni metodologiche proposte dalla nuova versione delle regole di contabilità nazionale, dall'altro l'introduzione del nuovo SEC ha rappresentato l'occasione per inserire ulteriori innovazioni e miglioramenti metodologici. Inoltre, le basi informative della contabilità sono state integrate con nuove fonti che si sono rese disponibili negli anni recenti e che non potevano essere utilizzate se non introducendo una discontinuità temporale (Istat, 2014).

Per quanto concerne l'oggetto del presente lavoro, occorre sottolineare come le principali novità, solo apparentemente non collegate, siano il maggior utilizzo dei dati amministrativi nonché l'ulteriore implementazione dei conti satellite.

Per quanto concerne il primo aspetto, la più evidente utilizzazione si è avuta con la costruzione della base dati (FRAME-SBS), per le statistiche di impresa, da cui derivano le stime dell'attività produttiva dei settori di mercato (Carucci, Vannella, 2016). La maggiore disponibilità ed utilizzo di dati amministrativi ai fini della produzione di informazione statistica, ha agevolato anche l'implementazione dei conti satellite su cui da anni l'Istat lavora. Ed è proprio a tali conti satellite che per la prima volta il SEC dedica un intero capitolo all'argomento, presentando come attraverso i conti satellite si possa contare su "maggiori dettagli di analisi, riorganizzando alcuni concetti del quadro centrale oppure fornendo informazioni supplementari, per esempio i flussi e le consistenze non monetari. Essi possono discostarsi dai concetti del quadro centrale. Modificando i concetti, è possibile migliorare il collegamento con concetti economici teorici come il benessere o i costi delle operazioni, concetti amministrativi come il reddito imponibile o gli utili nella contabilità aziendale, e concetti politici come le industrie strategiche, l'economia della conoscenza e gli investimenti economici utilizzati nella politica economica nazionale o europea" (Eurostat, 2013.a).

I conti, presentati nel capitolo 22 del SEC 2010, sono i seguenti:

- conti dell'agricoltura;
- conti ambientali;
- conti sanitari;
- conti relativi alla produzione delle famiglie;
- conti relativi al lavoro e matrici di contabilità sociale;
- conti relativi alla crescita e alla produttività;
- conti relativi alla ricerca e sviluppo;
- conti della protezione sociale;
- conti del turismo.

La varietà dei conti satellite dimostra come i conti nazionali, pur essendo un indispensabile quadro di riferimento per le statistiche nazionali, necessitano di una maggiore flessibilità concettuale e classificatoria propria dei conti satellite. Il SEC 2010 riporta l'esempio dei conti ambientali che, ampliando il quadro dei conti economici, riescono a valutare le esternalità ambientali, permettendo di giungere ad una misura più esaustiva del benessere, meno legata al concetto strettamente economico. In Fig. 1 si riporta una panoramica dei conti satellite e delle interazioni degli stessi con i conti economici nazionali.

Figura 1. *Panoramica sui conti satellite e sulle loro principali caratteristiche*

	Otto caratteristiche dei conti satellite								
		Conti per settore specifico							
	Conti funzionali	Collegamenti con le branche di attività economica o i prodotti	Collegamenti con i settori industriali	Inclusione di dati non monetari	Dettagli aggiuntivi	Concetti supplementari	Concetti fondamentali diversi	Risultati sperimentali e uso più ampio della modellizzazione	Parte del programma di trasmissione dell'Unione europea
1. *Conti satellite descritti nel presente capitolo*									
Agricoltura		X			X	X			X
Ambiente	X	X		X	X	X	X	X	X
Sanità	X	X		X	X		X		X
Produzione familiare			X	X	X		X	X	
Lavoro e SAM		X	X	X	X				
Produttività e crescita		X		X	X	X	X	X	X
R & S	X	X		X	X		X	X	
Protezione sociale	X			X	X				X
Turismo	X	X		X	X	X			

Fonte: Eurostat

2.1 I conti ambientali

Il SEEA individua un complesso quadro di contabilità per descrivere e analizzare l'ambiente e le sue interazioni con l'economia. I conti ambientali, essendo fra l'altro, conti satellite di contabilità nazionale, utilizzano, in generale, gli stessi concetti e le stesse classificazioni della contabilità nazionale.

L'integrazione tra conti economici e conti ambientali, consente di analizzare il contributo dell'ambiente all'economia e l'impatto dell'economia sull'ambiente.

Sebbene il quadro centrale dei conti economici, rappresenti il punto di partenza per la contabilità ambientale, ad esso si aggiungono ulteriori elementi e riclassificazioni al fine di superare le principali criticità legate agli aggregati economici (PIL, investimenti e risparmi). Il SEC2010 individua, in particolare, due criticità. In primo luogo la scarsità e il progressivo esaurimento delle risorse naturali possono minacciare la produttività dell'economia. In secondo luogo, i conti economici non coprono in modo esaustivo il degrado della qualità ambientale, i conseguenti danni per la salute umana ed il benessere.

Il regolamento SEEA prevede i seguenti conti:

- conti di flusso fisici e ibridi;
- conti economici per le operazioni ambientali;
- conti del patrimonio ambientale in termini fisici e monetari;
- conti per le spese di protezione e l'esaurimento delle risorse;
- modifica degli aggregati dal quadro centrale per tener conto del degrado.

Attualmente l'Italia, per adempiere al Regolamento UE 691/2011 sulle statistiche ambientali e su base volontaria, produce le tipologie di conti ambientali (Tudini, 2015) riportate in Fig. 2.

Figura 2. *Attività corrente relativa ai conti economici ambientali*

CONTI FISICI	***Conti delle emissioni atmosferiche*** ***Conti dei flussi di materia a livello di intera economia***
CONTI MONETARI	***Imposte ambientali ripartite per attività economica*** ***Spesa ambientale***

Al fine inoltre di adempiere al Reg UE n. 538/2014 che modifica il regolamento UE n. 691/2011 relativo ai conti economici ambientali europei, sempre in Italia si intende sviluppare e perfezionare nel medio periodo le attività presentate in Fig. 3.

Figura 3. *Attività da sviluppare e perfezionare relativa ai conti economici ambientali*

CONTI FISICI	***Conti dei flussi fisici dell'energia***
CONTI MONETARI	***Conti del settore dei beni e dei servizi ambientali*** ***Conti delle spese per la protezione dell'ambiente***

I conti fisici dell'ambiente si fondano sulla necessità di disporre di una descrizione delle interazioni tra sistema economico e sistema ambientale, fornendo dunque ai decisori politici un indispensabile strumento di analisi.

2.2 I conti fisici: Conti dei flussi di materia a livello di intera economia

I conti dei flussi di materia in termini fisici registrano quattro diversi tipi di flussi (Eurostat, 2013.b):

- risorse naturali: risorse minerarie, energetiche, del suolo e biologiche;
- input dell'ecosistema;
- beni e servizi prodotti nell'ambito della sfera economica e impiegati al suo interno;

- residui: prodotti accidentali e indesiderati dell'economia che hanno un valore pari a zero o negativo per chi li genera.

I flussi fisici vengono misurati in quantità, utilizzando unità che riflettono le caratteristiche fisiche del materiale, solitamente tonnellate.

In particolare, i conti dei flussi di materia consentono anche di calcolare indicatori relativi alle pressioni del sistema economico sull'ambiente naturale, quali:

- il Direct Material Input (DMI): quantità totale di risorse naturali estratte all'interno della regione e di risorse importate che entra nel sistema economico per essere successivamente trasformata e/o commercializzata, è dato dalla somma delle biomasse, dei combustibili fossili e degli altri minerali estratti dal suolo nazionale, nonché dei beni di ogni tipo importati dall'estero, inclusi nell'aggregato secondo il loro peso effettivo, indipendentemente dal grado di lavorazione.
- il Domestic Material Consumption (DMC): consumo interno di materiali, calcolato a partire dal DMI al netto delle esportazioni, include i materiali, sia estratti internamente che importati, che rimangono nel paese e che sono accumulati in stock o trasformati in rifiuti, emissioni, usi dissipativi ecc;
- il Physical Trade Balance (PTB): bilancia commerciale fisica calcolata come differenza tra la quantità di materia importata e quella esportata, ovvero tra peso totale dei beni importati e peso totale dei beni esportati. Esso fornisce una prima indicazione sul ruolo giocato dal paese nella divisione internazionale dell'estrazione delle risorse e del loro uso – e delle relative pressioni sull'ambiente naturale – che accompagnano la divisione internazionale del lavoro. I paesi esportatori netti di materia sopportano infatti le pressioni sull'ambiente derivanti dalle attività di estrazione e produzione che svolgono "in favore" dei paesi che sono importatori netti di beni materiali, mentre in questi ultimi l'output verso l'ambiente naturale locale e/o l'accumulo di capitale fisico è maggiore di quanto sarebbe permesso dalle sole risorse del paese.

La metodologia di analisi utilizzata, Material Flow Analysis (MFA) si basa sul principio fisico di conservazione della massa: "nulla si crea e nulla si distrugge", vale a dire che la massa entrante in un sistema socioeconomico si bilancia in maniera esatta con la materia uscente a meno delle variazioni degli stock.

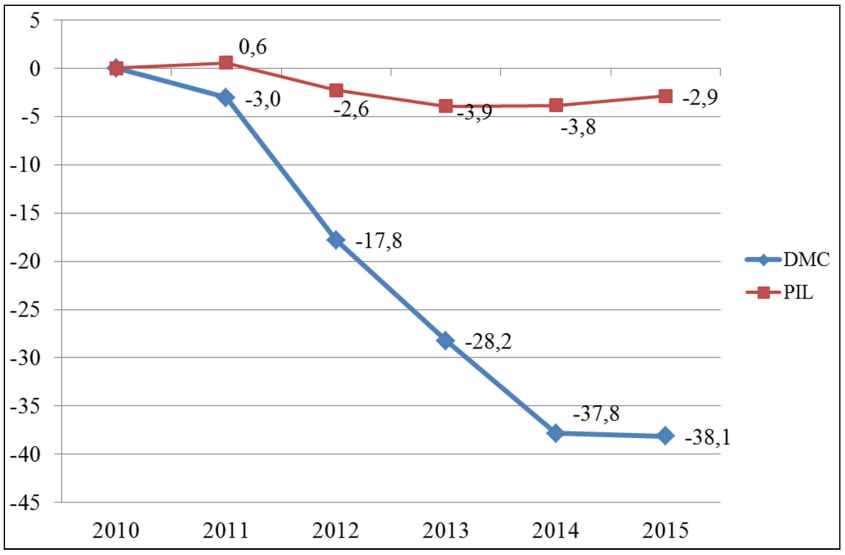
L'analisi del DMC e del PIL (Tab. 1, Fig. 4), permette di verificare se vi è o meno una disgiunzione tra lo sfruttamento di risorse naturali e la crescita economica. Tra il 2010 e il 2015, il PIL a prezzi costanti dell'Italia è diminuito del 2,9%.

Tabella 1. *Indicatori relativi alle pressioni del sistema economico sull'ambiente naturale e PIL. Anni 2010-2015*

Indicatori	2010	2011	2012	2013	2014	2015
DMI - input materiale diretto *(migliaia di tonnellate)*	828.297	806.817	710.790	635.651	562.411	573.452
DMC - consumo materiale interno *(migliaia di tonnellate)*	681.741	661.035	560.429	489.336	423.717	421.785
PTB - bilancia commerciale fisica *(migliaia di tonnellate)*	198.409	191.293	163.767	151.281	150.180	153.240
PIL - prodotto interno lordo. Valori concatenati 2010 (milioni di euro)	***1.604.514***	***1.613.766***	***1.568.274***	***1.541.171***	***1.542.924***	***1.558.317***

Fonte: Elaborazione su dati Istat

Figura 4. *Andamento PIL e DMC Italia. Anni 2010-2015. Variazioni percentuale rispetto al 2010.*



Fonte: Elaborazione su dati Istat

Guardando al consumo delle risorse fisiche, si può evidenziare una tendenza di decrescita di quasi il 40%.

In particolare, dal 2013 il PIL, sebbene in modo non marcato, è cresciuto e, nello stesso periodo, il DMC ha continuato il suo trend di decrescita.

Questo è il tipico esempio di sganciamento del consumo di materia dalla "crescita", che può considerarsi indipendente dalla pressione ambientale.

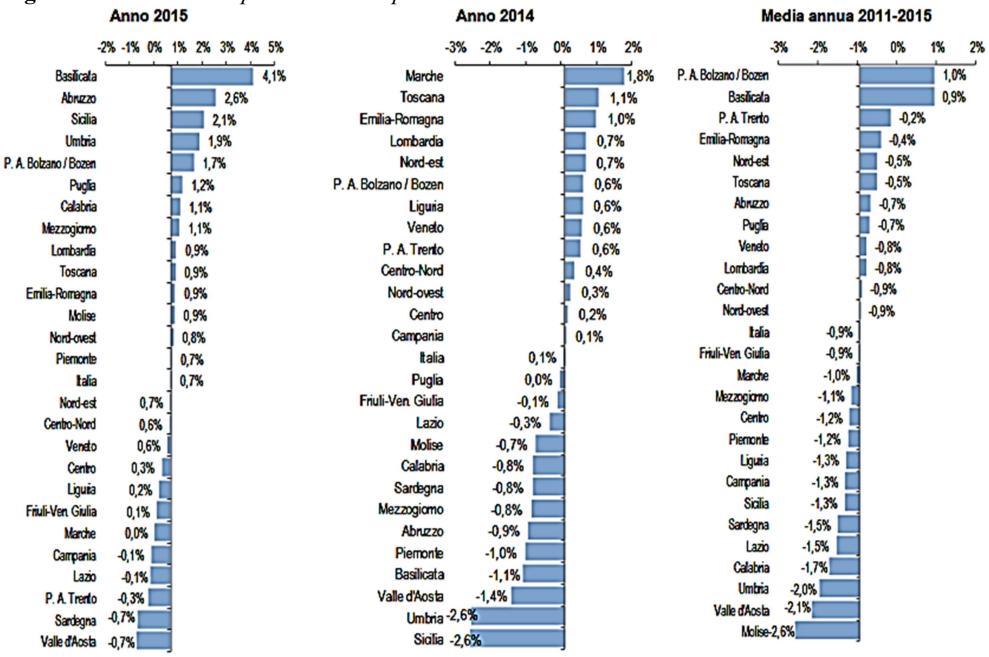
3. Caso di studio: la Basilicata, perché e come integrare i conti economici ed i conti satellite.

La struttura produttiva della Basilicata è un unicum in Italia poiché caratterizzata da una distribuzione del valore aggiunto per attività economica completamente diversa rispetto alla distribuzione media del Paese. Aspetti economici che portano la regione a differenziarsi dal quadro economico generale e del Mezzogiorno.

3.1 I conti economici

L'analisi dei dati economici, pubblicati dalla Contabilità Nazionale nel novembre 2016, rileva che nel 2015, il Pil in volume a livello nazionale aumenta dello 0,7% rispetto all'anno precedente. La migliore performance dell'ultimo anno a livello di macroaree (Fig. 5) è quella del Mezzogiorno, che ha segnato una crescita dell'1,1% rispetto al 2014, trainata da Basilicata (+4,1%), Abruzzo (+2,6%), Sicilia (+2,1%) e Puglia (+1,2%). In Basilicata, la crescita del valore aggiunto è trainata dall'industria, +18,5%, le costruzioni invece perdono in volume il 4,8%.

Figura 5. *Variazioni percentuali del prodotto interno lordo in volume. Anni 2011-2015.*



Fonte: Istat, Conti economici regionali

La crescita della Basilicata è confermata anche in termini di occupati dipendenti, +3% rispetto allo 0,9% dell'Italia sebbene l'aumento del reddito da lavoro di-

pendente non copra l'incremento occupazionale. Infatti il reddito da lavoro dipendente per occupato diminuisce di 0,7% rispetto all'anno precedente. (Prospetto 1)

Prospetto 1. *Redditi da lavoro dipendente, occupati dipendenti e redditi da lavoro dipendente per occupato, per regione. Variazioni percentuali*

REGIONI	2015/2014			2014/2013			Media annua 2011/2015		
	Redditi da lavoro dipendente	Occupati dipendenti	Redditi da lavoro dipendente per occupato	Redditi da lavoro dipendente	Occupati dipendenti	Redditi da lavoro dipendente per occupato	Redditi da lavoro dipendente	Occupati dipendenti	Redditi da lavoro dipendente per occupato
Piemonte	1,7	0,4	1,3	0,0	-0,8	0,8	-0,2	-0,5	0,3
Valle d'Aosta	0,2	0,0	0,2	-1,7	-0,9	-0,8	-1,2	-0,3	-0,8
Lombardia	2,0	0,5	1,5	1,0	0,5	0,5	0,5	0,0	0,6
Provincia Autonoma Bolzano / Bozen	2,7	1,9	0,8	0,2	0,7	-0,5	1,2	0,7	0,5
Provincia Autonoma Trento	1,6	0,7	0,9	-2,3	-0,9	-1,5	-0,5	0,0	-0,5
Veneto	1,4	-0,5	1,9	1,4	0,8	0,6	0,3	-0,4	0,7
Friuli-Venezia Giulia	1,1	0,1	1,0	-0,9	-0,1	-0,8	-0,9	-0,7	-0,2
Liguria	0,8	-0,3	1,1	0,8	0,5	0,3	0,3	-0,3	0,6
Emilia-Romagna	2,2	1,1	1,1	0,5	-0,5	1,0	0,6	0,0	0,6
Toscana	2,2	2,1	0,1	-0,1	0,4	-0,5	0,2	0,5	-0,4
Umbria	2,7	3,6	-0,9	-2,4	-2,3	-0,1	-0,8	-0,4	-0,3
Marche	0,9	-1,5	2,4	0,3	-0,4	0,7	-0,2	-1,1	0,9
Lazio	1,1	0,6	0,5	0,6	0,9	-0,3	-0,2	0,1	-0,3
Abruzzo	3,4	3,1	0,4	0,1	-1,3	1,3	-0,4	0,1	-0,5
Molise	1,1	1,5	-0,4	-1,8	0,4	-2,2	-2,2	-1,0	-1,1
Campania	2,0	2,3	-0,3	-0,1	1,3	-1,4	-0,6	-0,1	-0,5
Puglia	2,6	3,1	-0,5	-0,3	-0,1	-0,2	-0,5	-0,1	-0,4
Basilicata	2,2	3,0	-0,7	0,9	-0,6	1,5	-1,0	-0,4	-0,6
Calabria	-0,2	-2,5	2,3	-0,4	-0,8	0,4	-1,5	-2,0	0,4
Sicilia	1,4	1,1	0,3	-1,6	-0,5	-1,1	-1,3	-0,9	-0,4
Sardegna	0,9	2,1	-1,1	-1,5	-0,4	-1,0	-0,7	-0,1	-0,6
Nord-ovest	1,8	0,4	1,4	0,7	0,1	0,6	0,3	-0,2	0,5
Nord-est	1,8	0,4	1,4	0,6	0,1	0,4	0,3	-0,2	0,5
Centro	1,5	1,0	0,5	0,2	0,4	-0,2	-0,1	0,1	-0,2
Centro-nord	1,7	0,6	1,0	0,5	0,2	0,2	0,2	-0,1	0,2
Mezzogiorno	1,8	1,8	0,0	-0,6	0,0	-0,6	-0,9	-0,5	-0,4
Italia	1,7	0,9	0,8	0,2	0,2	0,1	-0,1	-0,2	0,1

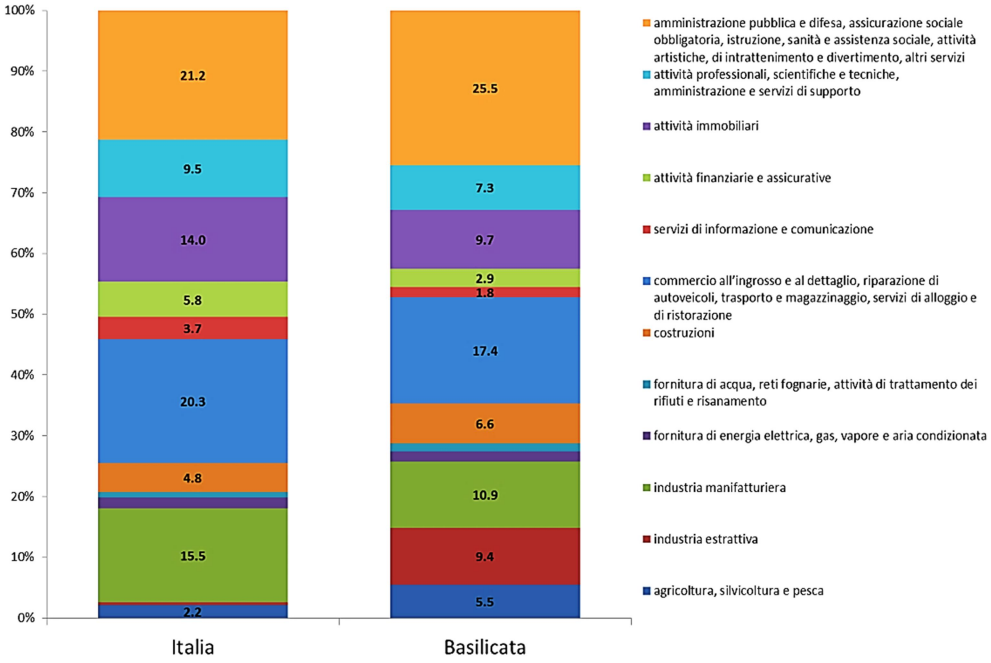
Fonte: Istat, Conti economici regionali

Dall'analisi della distribuzione di valore aggiunto e addetti per attività economica in Italia e Basilicata (Fig. 6), si evincono facilmente le specificità della regione, utili a spiegare le dinamiche di crescita. In Basilicata, più importante rispetto all'intera economia è il contributo, in termini di valore aggiunto, dell'agricoltura, dell'attività estrattiva e della pubblica amministrazione.

In termini occupazionali, la distribuzione delle attività economiche (Fig. 7) è particolarmente diversa: le attività estrattive pur rappresentando quasi il 10% del valore aggiunto, rappresentano meno dell'1% degli occupati; di contro il 5,5% del valore aggiunto dell'agricoltura, deriva da più del 10% di occupati nel settore.

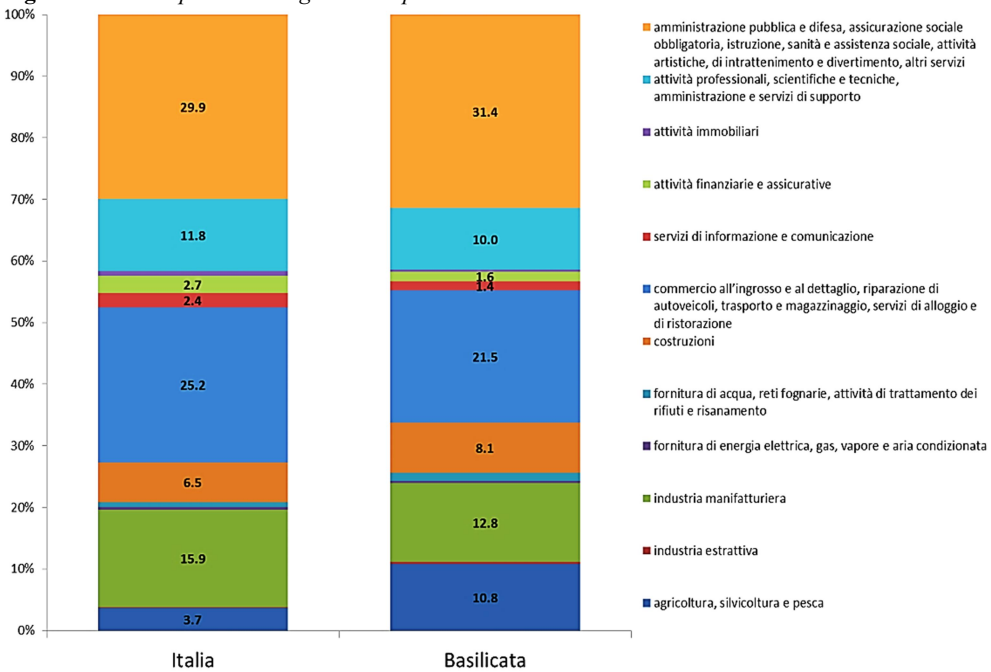
In questo contesto, affiancare alla tradizionale analisi economica, delle analisi più prettamente ambientali risulta particolarmente utile.

Figura 6. *Composizione del valore aggiunto per attività economica: Italia e Basilicata. Anno 2015.*



Fonte: Elaborazioni su dati Istat

Figura 7. *Composizione degli addetti per attività economica: Italia e Basilicata. Anno 2015.*



Fonte: Elaborazioni su dati Istat

3.2 I conti dei flussi fisici di materia[1]

La Regione Basilicata in collaborazione con la sede territoriale dell'Istat, nell'ambito di un protocollo di convenzione, ha previsto di compilare i conti dei flussi di materia relativi alla regione. La finalità è di costruire un indicatore regionale relativo al consumo interno di materiale e di risorse naturali da confrontare con l'andamento dell'economia regionale, che integri il set di indicatori regionali del "Benessere Equo e Sostenibile", nella dimensione Ambiente.

Il lavoro è stato svolto in base alla metodologia europea EW_MFA (Eurostat 2001), alla più recente versione della guida pratica alla compilazione (Eurostat 2013.b) e alle pubblicazioni Istat in materia. Le fonti sono i microdati delle indagini Istat sulle produzioni agricole, commercio con l'estero, trasporto merci, i dati sulle estrazioni dei combustibili fossili del Ministero dello Sviluppo Economico e alcuni dati amministrativi regionali opportunamente trattati.

Sono state pertanto computate:

- le estrazioni interne di materiale suddivise in quattro macro categorie: Biomassa, Minerali metalliferi intesi come minerali grezzi, Minerali non metalliferi, Materiali e vettori energetici fossili;
- i flussi di ingresso nella regione di materiali provenienti dall'estero e dalle altre regioni italiane. Le importazioni fisiche comprendono tutti i prodotti a qualunque stadio della trasformazione da materia prima a prodotto finito. I flussi di import sono suddivisi in 6 macro categorie: Biomassa e prodotti da biomassa; Minerali metalliferi concentrati, grezzi e trasformati; Minerali non metalliferi grezzi e trasformati; Materiali e vettori energetici fossili grezzi e trasformati; Rifiuti importati per il trattamento definitivo e materie prime secondarie; Altri materiali che non rientrano espressamente nelle precedenti;
- il dettaglio dei flussi di ingresso nella regione di materiali provenienti dalle economie estere;
- i flussi in uscita dalla regione, di materiali diretti alle altre regioni italiane e alle economie estere. Le esportazioni fisiche comprendono tutti i prodotti a qualunque stadio della trasformazione da materia prima a prodotto finito. I flussi di export sono suddivisi nelle stesse 6 macro categorie dei flussi di import;
- il dettaglio dei flussi di uscita dalla regione di materiali destinati alle economie estere.

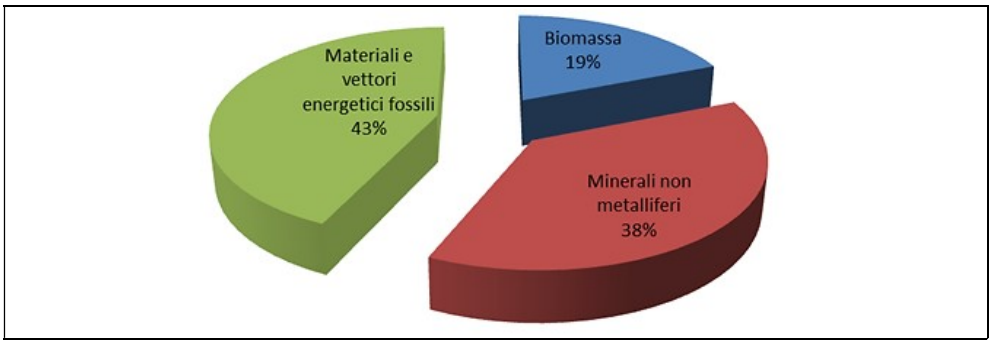
[1] Le analisi e le elaborazioni sono tratte dal Rapporto su "Conti dei flussi di materia a livello di intera economia della regione Basilicata" stilato dalla sede ISTAT per la Basilicata ed in particolare da Flora Fullone, Salvatore Cariello e Antonella Bianchino.

3.2.1. Estrazione interna

Nel 2013, anno per cui è attualmente disponibile l'intero set di fonti utili alla compilazione del conto dei flussi di materia per la Basilicata, sono stati prelevati dal sistema naturale della regione 11,1 milioni di tonnellate di risorse naturali vergini, di cui 2,1 da biomassa, 4,2 da minerali e 4,8 da materiali e vettori energetici fossili.

Il 43% dei materiali estratti è costituito da petrolio e gas, il 38% da minerali non metalliferi, per lo più argilla, calcare, gesso, etc., che confluiscono nel settore delle costruzioni. Infine il 19% dell'estrazione interna è rappresentato da biomasse, per la maggior parte cereali, residui utilizzati di colture e foraggio (Fig. 8, Tab. 2).

Figura 8. *Composizione percentuale dell'estrazione interna di materiale. Anno 2013*



Fonte: Rapporto Regione Basilicata – ISTAT

Tabella 2. *Estrazione interna di materiale. Anni 2008-2013 (migliaia di tonnellate)*

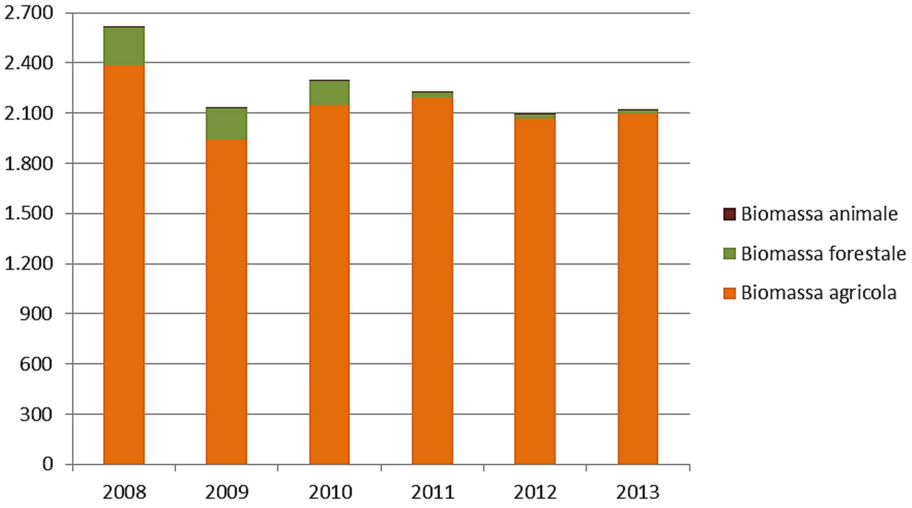
	2008	**2009**	**2010**	**2011**	**2012**	**2013**
Biomassa	2.615	2.133	2.294	2.227	2.093	2.121
Minerali non metalliferi	10.500	10.300	9.700	7.700	5.200	4.206
Materiali e vettori energetici fossili	4.692	3.800	4.227	4.557	4.944	4.836
Totale	**17.807**	**16.233**	**16.221**	**14.484**	**12.237**	**11.163**

Fonte: Rapporto Regione Basilicata - ISTAT

Su 2,1 milioni di tonnellate di biomassa estratta nel 2013 (Fig. 9), quella di origine forestale e animale ammonta a poco più di 20 mila tonnellate (1% del totale della biomassa). Le coltivazioni agricole costituiscono la quota di gran lunga preponderante di biomassa estratta nella regione, con un peso che passa dal 91,3% del 2008 al 99% del 2013.

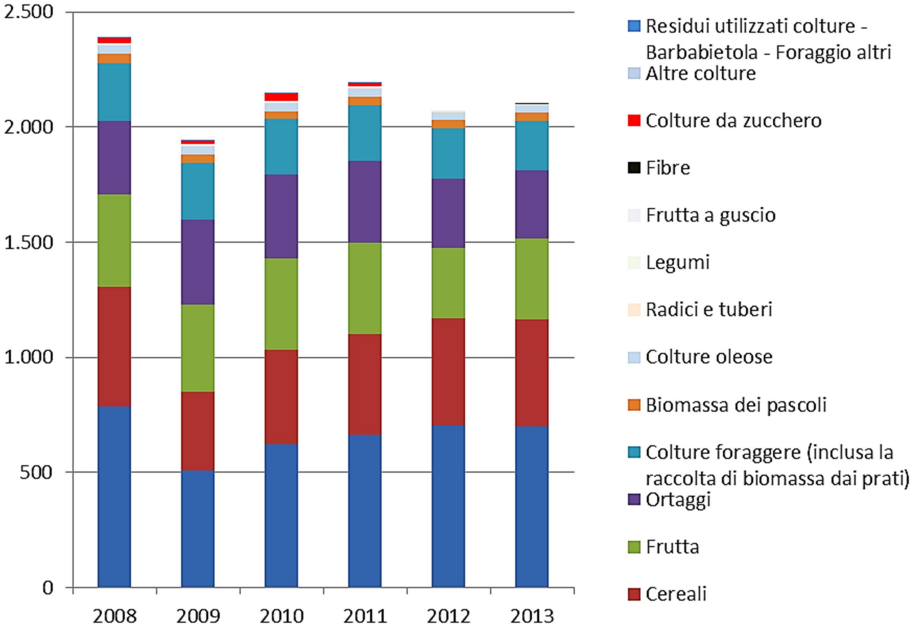
Le biomasse forestali fanno registrare una considerevole riduzione negli anni 2008-2010 (-35,5%) e un vero crollo nel triennio successivo: si passa dalle 225 mila tonnellate del 2008, alle 145 mila del 2010 e alle 19 mila del 2013.

Figura 9. *Estrazione di biomassa. Anni 2008-2013 (migliaia di tonnellate)*



Fonte: Rapporto Regione Basilicata - ISTAT

Figura 10. *Estrazione di biomassa agricola. Anni 2008-2013 (migliaia di tonnellate)*



Fonte: Rapporto Regione Basilicata – ISTAT

Le biomasse agricole (Fig. 10) sono costituite essenzialmente da cereali (464 mila tonnellate nel 2013, pari al 21,9% del totale), ortaggi (296 mila tonnellate, 14%), frutta (350,4 mila tonnellate, 16,5%), residui utilizzati di colture (701 mila tonnellate, 33,1%) e colture foraggere (214 mila tonnellate, 10,1%).

Tabella 3. *Composizione dei materiali e vettori energetici fossili. Anni 2008-2013 (migliaia di tonnellate)*

	2008	2009	2010	2011	2012	2013
Petrolio greggio	3.930	3.156	3.443	3.731	4.033	3.940
Gas Naturale	761	644	784	826	911	896
Materiali e vettori energetici fossili	4.692	3.800	4.227	4.557	4.944	4.836

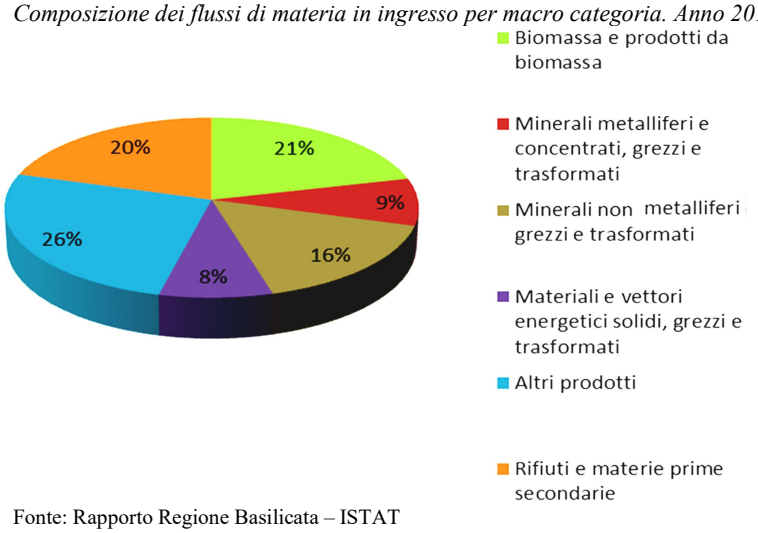
Fonte: Rapporto Regione Basilicata – ISTAT

Nel 2013 sono stati estratti, in Basilicata, 4,8 milioni di tonnellate di materiali e vettori energetici fossili, circa 108 mila in meno rispetto ai 4,9 milioni di tonnellate del 2012 (Tab. 3). Il greggio, 3,9 milioni di tonnellate estratte nel 2013, rappresenta l'81,5% delle estrazioni, il gas naturale il restante 18,5%. Il 2012 è stato l'anno in cui si è registrato il più alto livello di estrazione di risorse energetiche fossili: 4,9 milioni di tonnellate di cui 4 di petrolio e 0,9 di gas naturale.

3.2.2 Flussi di materia in ingresso nella regione Basilicata

I flussi di materia in ingresso nella regione Basilicata sono valutati in base al contributo dell'import dall'estero e al contributo del materiale importato dalle altre regioni italiane. I valori rappresentati si riferiscono ai soli prodotti importati senza considerare i flussi generati "a monte", cioè i materiali trasformati in emissioni e rifiuti per la produzione dei prodotti importati ("flussi indiretti").

Figura 11. *Composizione dei flussi di materia in ingresso per macro categoria. Anno 2013*



Fonte: Rapporto Regione Basilicata – ISTAT

I flussi di materia in ingresso nella regione sono stati aggregati in 6 macro-categorie (Fig. 11): biomassa e prodotti da biomassa, minerali metalliferi e concen-

trati, grezzi e trasformati, minerali non metalliferi, grezzi e trasformati, materiali e vettori energetici solidi, grezzi e trasformati, altri prodotti e rifiuti e materie prime secondarie.

Nel 2013 in Basilicata (Tab. 4) si è avuto un flusso di materia in ingresso di 4,3 milioni di tonnellate, di cui 911 mila (21,1%) costituiti da biomasse, 690 mila (16%) di minerali non metalliferi e 876 mila (20,3%) classificate come rifiuti per il trattamento e smaltimento definitivo e materie prime secondarie. Il picco dell'im–portazione di biomasse si raggiunge nel 2009 con 1,9 milioni di tonnellate; l'anno successivo, invece, si tocca il punto più basso con 536 mila tonnellate.

Tabella 4. *Flussi di materia in ingresso per macro-categoria. Anni 2008-2013 (migliaia di tonnellate)*

	2008	**2009**	**2010**	**2011**	**2012**	**2013**
Biomassa e prodotti da biomassa	780	1927	536	762	939	911
Minerali metalliferi e concentrati, grezzi e trasformati	724	423	471	512	1137	366
Minerali non metalliferi, grezzi e trasformati	1.445	3.652	2.009	2.939	1.540	690
Materiali e vettori energetici fossili, grezzi e trasformati	322	424	985	482	370	352
Altri prodotti	1.633	1.892	2.497	1.163	949	1.118
Rifiuti e materie prime secondarie	80	188	1409	47	202	876
Totale	**4.984**	**8.506**	**7.907**	**5.906**	**5.137**	**4.314**

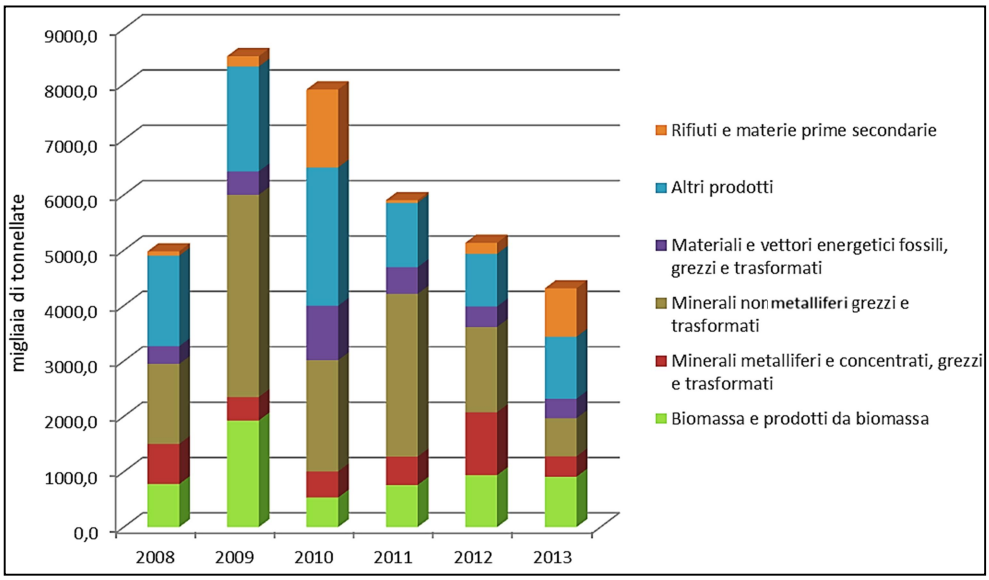
Fonte: Rapporto Regione Basilicata - ISTAT

Diminuiscono sensibilmente sia l'import di minerali non metalliferi che di quelli metalliferi. I primi passano da 1,4 milioni di tonnellate del 2008 a 690 mila tonnellate del 2013; i secondi da 724 mila tonnellate in ingresso del 2008 a 366 mila del 2013. L'import di materiali e vettori energetici fossili, eccezion fatta per il 2010 quando si toccano le 985 mila tonnellate, si mantiene intorno alle 400 mila tonnellate per anno.

Come si evince dalla Fig. 12, dopo aver toccato gli 8 milioni e mezzo di tonnellate nel 2009, l'import di materia registra negli anni successivi un costante decremento, fino a dimezzarsi nel 2013 (4,3 milioni di tonnellate).

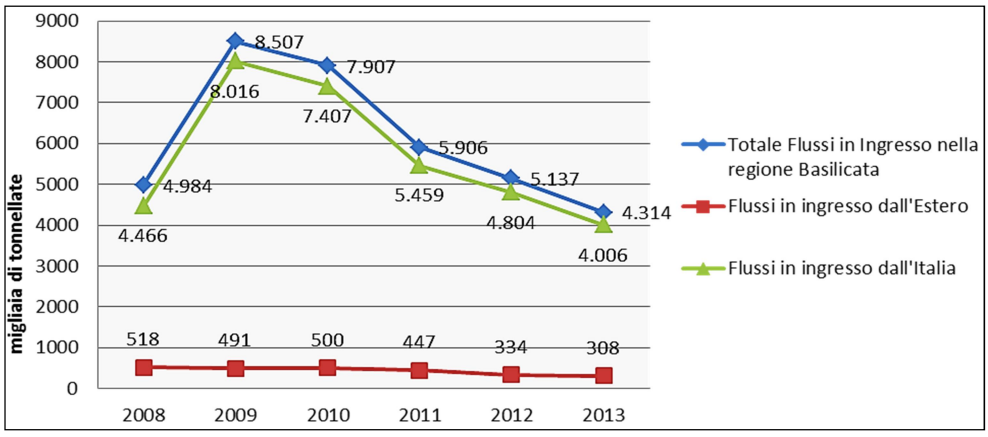
Rispetto all'origine dei flussi, la Fig. 13 evidenzia il ruolo preponderante dell'import da altre regioni italiane. Infatti i flussi dall'estero contribuiscono mediamente solo per il 7% sul totale della materia. Sia i flussi in ingresso dalle regioni italiane che quelli dal–l'estero hanno un andamento decrescente nel periodo 2009-2013.

Figura 12. *Flussi di materia in ingresso per macro categoria. Anni 2008-2013*



Fonte: Rapporto Regione Basilicata - ISTAT

Figura 13. *Flussi di materia in ingresso per provenienza. Anni 2008-2013*



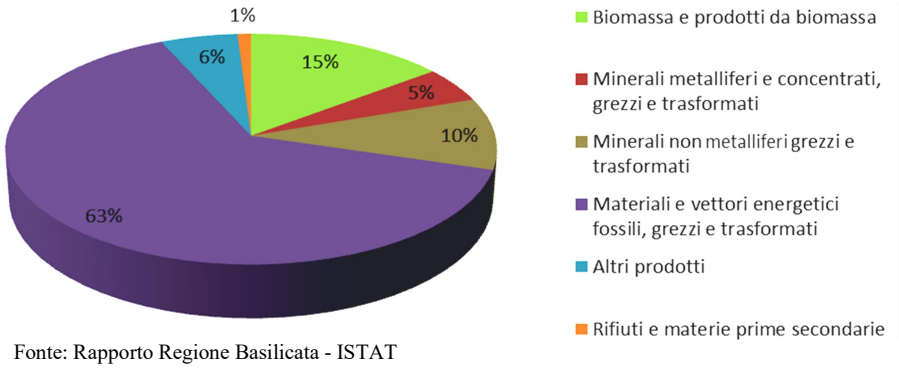
Fonte: Rapporto Regione Basilicata - ISTAT

3.2.3 Flussi di materia in uscita dalla regione Basilicata

I flussi di materia in uscita dalla regione Basilicata sono stati determinati computando il contributo dell'export verso l'estero e il contributo del materiale esportato verso le altre regioni italiane. I valori rappresentati si riferiscono ai soli prodotti esportati senza considerare i flussi generati in emissioni e rifiuti, per la produzione delle materie esportate. I flussi in uscita nel 2013 sono stati pari a 8,2 milioni di tonnellate, di cui 5,2 (63,3%) costituiti da materiali e vettori energetici fossili.

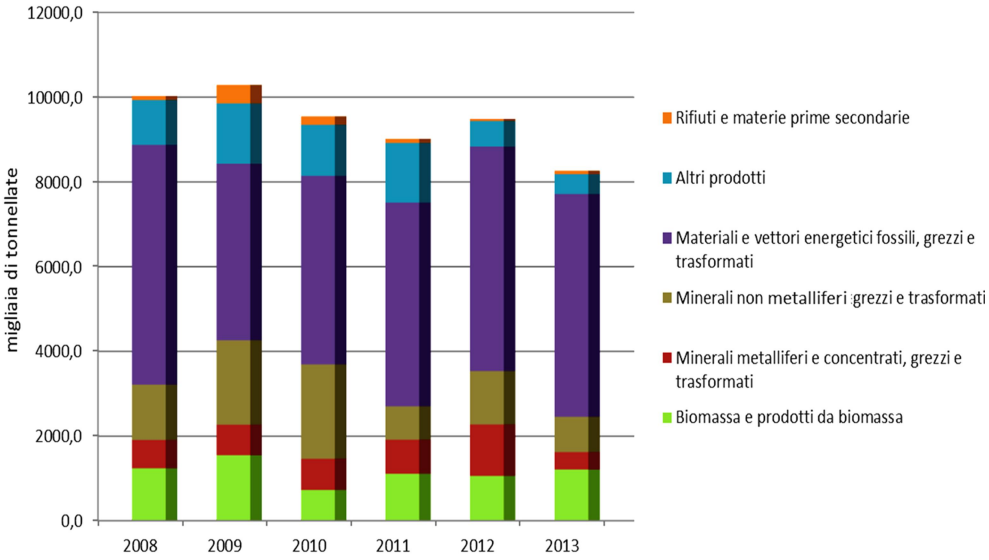
L'export di biomassa è stato di 1,2 milioni di tonnellate (14,7%), quello di minerali non metalliferi è stato di 830 mila tonnellate (10%). (Fig. 14 e Fig. 15)

Figura 14. *Composizione dei flussi in uscita. Anno 2013*



Fonte: Rapporto Regione Basilicata - ISTAT

Figura 15. *Flussi di materia in uscita per macrocategoria. Anni 2008-2013*

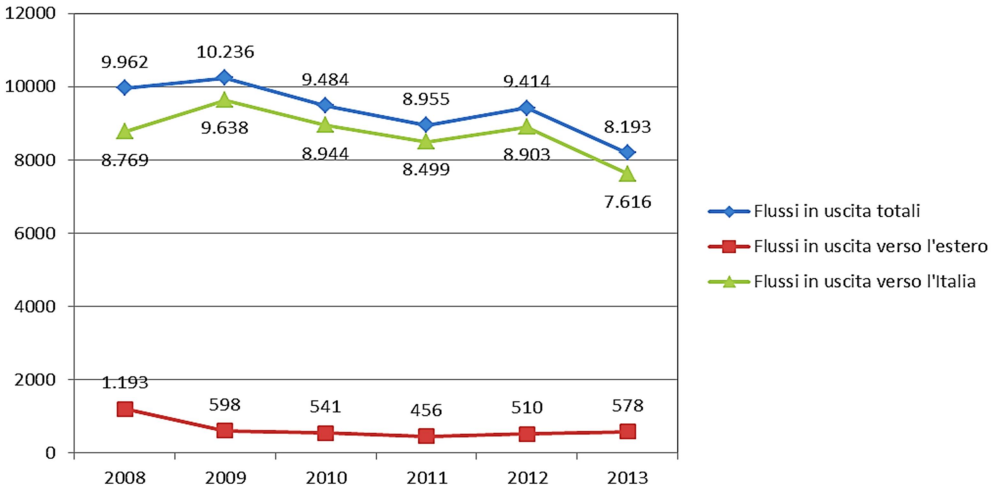


Fonte: Rapporto Regione Basilicata – ISTAT

Tra il 2008 e il 2013 il volume dell'export complessivo di materia cala del 17%, soprattutto per effetto della contrazione dell'export di minerali metalliferi e non metalliferi (-38,4% per i minerali metalliferi e -36,5 per i non metalliferi). Come per i flussi in ingresso, anche per quelli in uscita il contributo maggiore è fornito dallo scambio con le altre regioni italiane. La serie storica dei flussi in uscita mostra un trend decrescente, interrotto solo nel 2009 (+2,6% rispetto all'anno precedente) e nel 2012 (+5,1%). La dinamica dell'export di materia è determinata dai

flussi verso le altre regioni italiane, che rappresentano mediamente il 93% dell'export di materia dell'economia regionale. Come si evince dalla Fig. 16, la massa in uscita verso l'estero nel 2013 (578 mila tonnellate) rappresenta poco meno della metà di quella esportata nel 2008. La serie presenta un andamento decrescente fino al 2011 e una ripresa nei due anni successivi.

Figura 16. *Flussi di materia in uscita per destinazione. Anni 2008-2013 (migliaia di tonnellate)*



Fonte: Rapporto Regione Basilicata – ISTAT

4. Analisi degli indicatori e delle interazioni tra conti economici e conti dei flussi di materia

L'analisi degli indicatori derivanti dai conti dei flussi di materia, rappresenta il primo passo per verificare contestualmente gli effetti economici e ambientali dei flussi di materia. L'indicatore DMI – Direct Material Input - misura la quantità totale di risorse naturali estratte all'interno della regione a cui si aggiungono i prodotti in ingresso nella regione dal resto del mondo.

Tabella 5. *Input di materiale diretto (DMI). Anni 2008-2013 (migliaia di tonnellate)*

	2008	2009	2010	2011	2012	2013
Estrazioni Interne	17.806,7	16.232,6	16.220,8	14.484,0	12.237,4	11.162,6
Importazioni	4.984,0	8.506,7	7.906,7	5.906,2	5.137,5	4.314,2
DMI	**22.790,7**	**24.739,3**	**24.127,5**	**20.390,2**	**17.374,9**	**15.476,8**
DMI pro capite(t)	*39,0*	*42,4*	*41,5*	*35,2*	*30,1*	*26,8*
% estrazioni domestiche su DMI	*78,1*	*65,6*	*67,2*	*71,0*	*70,4*	*72,1*
% Importazioni su DMI	*21,9*	*34,4*	*32,8*	*29,0*	*29,6*	*27,9*

Fonte: Rapporto Regione Basilicata - ISTAT

Nella Tab.5 è riportata la serie storica dell'indicatore DMI, utilizzato per misurare l'input di materia in una economia. Nel 2013 l'input di materia nell'economia regionale è stato pari a 15,5 milioni tonnellate, di cui 11,2 milioni di tonnellate estratte all'interno del territorio regionale e 4,3 importate. Tra il 2008 e il 2013, le estrazioni interne hanno rappresentato, mediamente, il 70% del DMI regionale, contro il 60% registrato a livello nazionale e nello stesso periodo il DMI pro capite in Basilicata è sceso da 39 a 27 tonnellate (-31,2%).

L'indicatore di consumo di materiale interno (DMC) indica quanta parte del materiale estratto dalla regione e importato è "consumato" internamente, cioè trasformato in nuovi stock utili (edifici, infrastrutture, macchinari, beni durevoli, ecc.) del sistema antropico locale, in nuovi stock indesiderati (accumulo di rifiuti in discariche controllate) o in emissioni atmosferiche, reflui, sversamenti in discariche abusive o comunque fuori dal controllo dell'uomo. Si calcola come differenza tra l'input materiale diretto e le esportazioni. Nel 2013 il consumo di materiale nella regione Basilicata è stato pari a 7,3 milioni di tonnellate, con un valore pro capite di 12,5 tonnellate (Tab. 6).

Tabella 6. *Consumo interno di materiale (DMC). Anni 2008-2013(migliaia di tonnellate)*

	2008	2009	2010	2011	2012	2013
DMI	22.790,7	24.739,3	24.127,5	20.390,2	17.374,9	15.476,8
Esportazioni	9.962,1	10.236,2	9.484,4	8.955,1	9.413,9	8.193,4
DMC	**12.828,6**	**14.503,1**	**14.643,1**	**11.435,1**	**7.961,0**	**7.283,4**
% esportazioni su DMI	*43,7*	*41,4*	*39,3*	*43,9*	*54,2*	*52,9*
DMC pro capite(t)	*21,9*	*24,9*	*25,2*	*19,8*	*13,8*	*12,6*

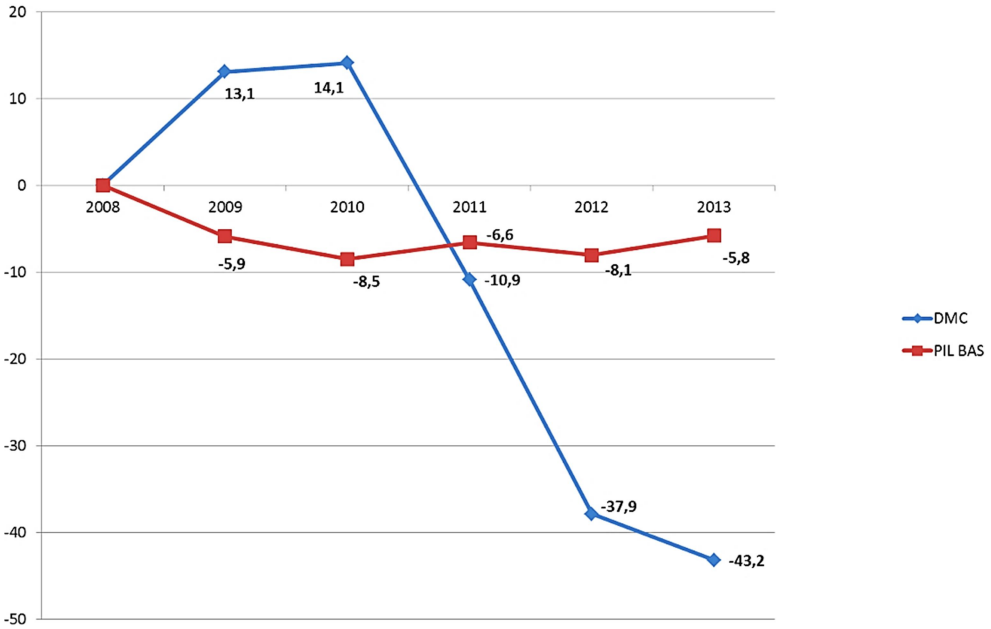
Fonte: Rapporto Regione Basilicata - ISTAT

Nel 2013 le esportazioni rappresentano il 52,9% del DMI regionale, in media nel periodo in esame le esportazioni hanno rappresentato il 46% dell'input di materiale nella regione. Tra il 2008 e il 2013 il valore del DMC pro capite nella regione scende da 21,9 a 12,6 tonnellate (-42,5%). Come si evince dalla Fig.17, tra il 2008 e il 2013, il PIL in volume della Basilicata è diminuito del 5,8%.

Guardando al consumo delle risorse fisiche, si può evidenziare una tendenza di decrescita di oltre il 40%. In particolare, dal 2012 il PIL, sebbene in modo non marcato, è cresciuto e, nello stesso periodo, il DMC ha continuato il suo trend di decrescita.

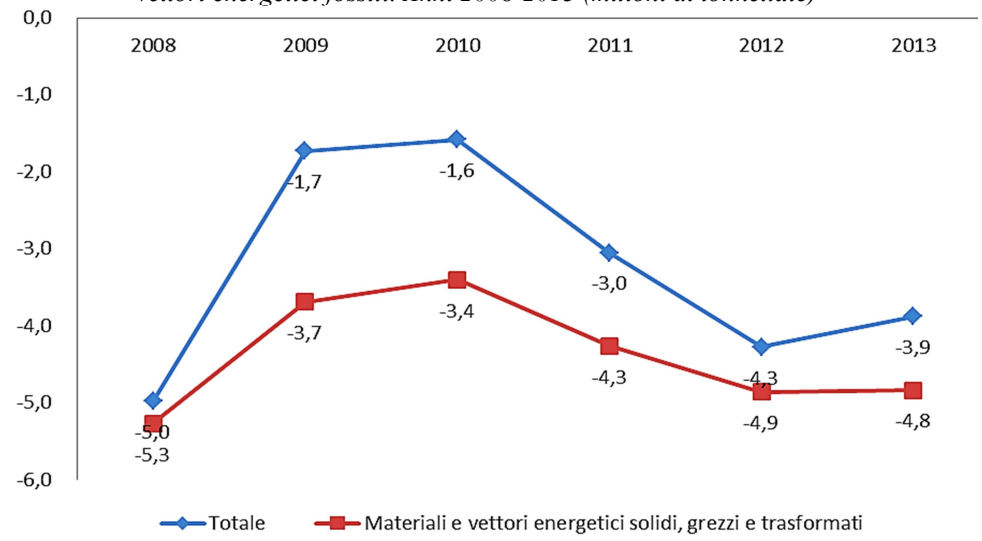
La bilancia commerciale fisica (*PTB: Physical Trade Balance*) misura il deficit o il surplus dei flussi di materia diretti del commercio di una economia. L'indicatore si ottiene sottraendo le esportazioni dalle importazioni, al contrario di quanto accade per la bilancia commerciale dei conti nazionali monetari.

Figura 17. *Andamento PIL e DMC regione Basilicata. Anni 2008-2013. Variazioni percentuali rispetto al 2008*



Fonte: Rapporto Regione Basilicata - ISTAT

Figura 18. *Bilancia commerciale fisica in complesso e per la macro categoria materiali e vettori energetici fossili. Anni 2008-2013 (milioni di tonnellate)*



Fonte: Rapporto Regione Basilicata – ISTAT

Come si evince dalla Fig. 18, la Basilicata è essenzialmente una regione produttrice di risorse che vengono consumate in economie esterne. Infatti, pur con signi-

ficative fluttuazioni, il saldo della bilancia commerciale fisica è sempre negativo, con valori che oscillano dai 5 milioni di tonnellate di surplus di materiale esportato nel 2008 all'1,6 milioni del 2010. Nel 2013 il saldo è stato pari a -3,9 milioni di tonnellate. Contrariamente a quanto accade a livello nazionale, dove si rileva un deficit di materia, compensato dalle importazioni, la Basilicata si caratterizza come territorio con prevalenza di materia esportata, sopportando quindi forti pressioni sull'ambiente a beneficio di altre economie.

5. Alcune considerazioni conclusive

Alla luce di quanto emerso, si può ritenere che l'esperienza lucana di implementazione dei conti ambientali permetta di estrapolare aspetti di particolare rilievo altrimenti non evidenziabili. In particolar modo si è avuto modo di notare come in una economia negli ultimi tempi in crescita, esista una forte incidenza nella formazione del Pil regionale della Pubblica amministrazione da un lato, e delle attività estrattive ed agricole dall'altro. Tale incidenza non sembra riprodursi un modo analogo per quanto concerne la distribuzione degli occupati.

Infatti le attività estrattive, come ci si poteva aspettare, pur presentando quasi il 10% del valore aggiunto lucano rappresentano meno dell'1% degli occupati, mentre l'agricoltura a fronte di una incidenza del 5,5% in termini di valore aggiunto, presenta una incidenza di oltre il 10% di occupati. Inoltre, per quanto concerne la quantità di estrazione interna di materiale, il settore agricolo forestale ha contributo con una partecipazione pari al 19% (biomassa), i minerali non metalliferi in misura del 38% ed i materiali e vettori energetici fossili (petrolio e gas naturale) in misura del 43%.

In riferimento a tale aspetto, mentre per i flussi di estrazione interna regionale si è assistito ad un calo del 37,3%, l'estrazione di materiali e vettori energetici fossili ha visto in controtendenza un aumento del 3%. A fronte di tali risorse prelevate dall'ambiente naturale, i flussi in ingresso importati risultano essere sempre significativamente più ridotti, e sostanzialmente non provenienti dall'estero.

È infine emerso come le risorse prelevate siano poco utilizzate internamente poiché destinate prevalentemente ad altre regioni italiane.

Tali rilevanze, unite alla considerazione che l'agricoltura a fronte di una produzione di biomassa minore ed in calo presenta una partecipazione più significativa al mercato del lavoro ancorché meno rilevante alla formazione del valore aggiunto, potrebbero far pensare ad una regione le cui risorse fossili siano impiegate prevalentemente per il fabbisogno di altre regioni italiane con impatto solo sul valore aggiunto regionale ma non sul mercato del lavoro.

Tuttavia queste conclusioni risulterebbero con tutta evidenza essere parziali e non validabili, stante la assoluta non longevità dei dati rilevati e la necessità di valutare le modalità attraverso cui le attività estrattive incidano sull'indotto regionale.

Inoltre, trattandosi della prima sperimentazione condotta in tal senso, i dati economici sono riferiti a periodi temporali diversi, e di conseguenza potrebbe non esserci una totale coerenza fra la contabilità economica e la contabilità ambientale.

Emerge infine come l'implementazione di strumenti di contabilità ambientale, possa permettere di condurre analisi complesse relative alle interrelazioni tra economia ed ambiente ed in tal senso sicuramente fanno ben sperare gli sforzi attualmente in atto mirati a perfezionare tale strumento.

Riferimenti bibliografici

Carucci, A.M.M.; Vannella, G. (2016), *Sull'integrazione tra fonti amministrative e statistiche per le imprese, in Metodi ed analisi statistiche*, Dipartimento di Scienze economiche e metodi matematici, Università degli Studi di Bari Aldo Moro.

Eurostat (2001), *Economy-wide material flow accounts and derived indicators. A methodological guide*, Luxembourg.

Eurostat (2008), *The 2008 SNA*, European Commission, IMF, OECD.

Eurostat (2013.a), *Sistema europeo dei conti 2010*, Luxembourg.

Eurostat (2013.b), *Economy-wide Material Flow Accounts (EW-MFA). Compilation Guide 2013*, Luxembourg.

Istat (2010*), I flussi di materia del sistema socio-economico italiano, Tavole di dati, 17 maggio 2010* (http://www.istat.it/it/archivio/2243).

Istat (2014) *Il passaggio al Sec 2010 e la revisione generale dei conti nazionali*, atti del seminario del 16 dicembre 2014.

Istat (2016) *Conti economici regionali*, Comunicato stampa pubblicato il 12 dicembre 2016, http://www.istat.it/it/archivio/193916.

Regione Basilicata (2015), *Rapporto preliminare: Conti dei flussi di materia a livello di intera economia della Regione Basilicata,* Potenza.

Tudini, A. (2015), *Introduzione alla contabilità ambientale. Seminario sui conti economici ambientali*, Istat, Roma.

Vannella, G. (2001), *Un approccio multivariato all'analisi statistica economica sull'inquinamento marino Lucano*, Statistica nr. 2/01, CLUEB, Bologna.

Vannella, G. (2004), *Problems related to the development of integrated systems for economic and environmental accounting: a preliminary analysis of the economic-environmental impact of human activities on the marine pollution in the Basilicata region*, Statistica nr. 2/04, CLUEB, Bologna.

UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO
Metodi e analisi statistiche 2017
DIPARTIMENTO DI ECONOMIA E FINANZA
ISBN: 978-88-6629-013-1
pp. 233-251

Dynamiques démographiques en Méditerranée: tendances, défis, enjeux

Maria Carella[1*], Roberta Pace[1], Alain Parant[2]

[1]*Università degli Studi di Bari Aldo Moro,* [2]*Futuribles International, Paris*

Riassunto: L'article résume les tendances (en termes d'intensité et de calendrier) de la fécondité, de la mortalité et de la mobilité des personnes à l'œuvre depuis les années 1950 dans les différents pays du Bassin méditerranéen et analyse l'impact de ces tendances sur la structure par sexe et par âge. Il conclut sur la tendance générale au vieillissement des populations considérées et sur les défis et les enjeux de société associés à ce phénomène.

Keywords: bassin méditerranéen; fécondité; mortalité; mobilité; vieillissement.

1. Introduction

Le Bassin méditerranéen est un espace où se superposent des dynamiques très contrastées de croissance démographique et de développement socio-économique ; des dynamiques dont les effets croisés s'avèrent, au fil du temps, de plus en plus prégnants et délicats à gérer.

Cela se traduit, notamment, par un grand contraste de modèles démographiques. Le premier de ces modèles concerne les pays de la rive nord, plus particulièrement les pays membres de l'Union européenne, qui ont achevé le processus transitionnel depuis plusieurs années, expérimentent la phase du «post-modernisme démographique» (Van de Kaa, 1987) et se caractérisent par une baisse décisive de la fécon-

[*] Auteur correspondant: maria.carella1@uniba.it.

Cet article est le résultat d'un projet conjoint; toutefois les paragraphes 3 et 4 sont attribués à M. Carella, les paragraphes 1, 2, 5 à R. Pace, alors que le paragraphe 6 a été rédigé conjointement par R. Pace et A. Parant et le paragraphe 7 par M. Carella et A. Parant.. Le texte a été révisé, tant au plan thématique qu'au plan formel, par A. Parant.

dité et un vieillissement très marqué de leurs populations. Le deuxième régime intéresse les pays des rives sud et est, qui se situent à un stade moins avancé de la transition démographique.

Les différentiels de croissance humaine ayant été préalablement mis en évidence, l'article traite tout d'abord, à grands traits, de l'évolution des phénomènes clés de la dynamique démographique – fécondité, mortalité, migrations – au cours des 60 dernières années. Il se focalise ensuite sur la transformation de la structure par âge des populations des pays observée sous l'effet de l'évolution des composantes du renouvellement démographique. La tendance étant très clairement au vieillissement généralisé des populations méditerranéennes, la dernière partie de l'article s'articule autour de deux types d'interrogation : de quels enjeux ce vieillissement est-il porteur pour les sociétés concernées ? Quelles adaptations ou mutations leur impose-t-il à plus ou moins long terme ?

2. Population et croissance démographique d'un espace très hétérogène

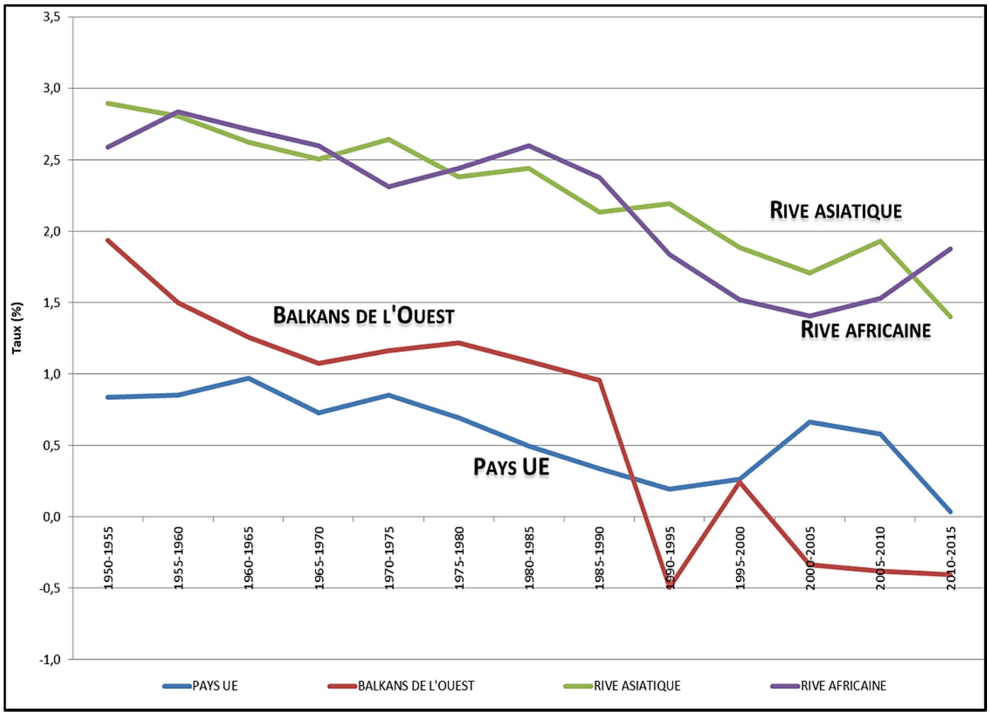
Le Bassin méditerranéen, tel que défini ici[1], est un espace dont la population totale s'élevait en 2015 à quelque 530 millions d'habitants. Par le niveau de développement des pays qui le composent autant que par leur peuplement, c'est un ensemble hétéroclite. Se côtoient, en effet, des pays très petits et très densément peuplés, comme Malte, la Palestine ou le Liban, et des pays très vastes mais très faiblement peuplés, tels que l'Algérie ou la Lybie; des sous-ensembles géographiques –rives asiatique et africaine– dont la population s'est très fortement accrue depuis 1950 et d'autres –ensemble des pays membres de l'Union européenne (UE) et Balkans de l'Ouest– où la population n'a que très faiblement augmenté.

Globalement, sur la période 1950-2015, la population du Bassin méditerranéen a augmenté de plus de 307 millions, l'accroissement étant aux trois-quarts le fait des rives est et sud. Si l'on excepte les pays de l'UE, un ensemble où la croissance démographique a assez peu fluctué dans une fourchette de taux faiblement positifs et en net regain depuis le début des années 2000, le Bassin méditerranéen est au-

[1] Dans ce travail, le champ d'étude porte sur 25 pays, les 21 pays qui ont une ouverture reconnue sur la Méditerranée et quatre pays –Portugal, Serbie (avec Kosovo), Ancienne république yougoslave de Macédoine (ARYM), Jordanie– qui se situent dans leur immédiat voisinage et entretiennent avec eux des liens privilégiés; il est subdivisé comme suit:
- rive européenne : Pays de l'Union européenne (Portugal, Espagne, France, Italie, Slovénie, Croatie, Grèce, Malte, Chypre); Balkans de l'Ouest (Albanie, Monténégro, Bosnie-et-Herzégovine, Serbie, ARYM);
- rive asiatique : Turquie, Syrie, Liban, Israël, Palestine, Jordanie;
- rive africaine : Egypte, Libye, Tunisie, Algérie, Maroc.

jourd'hui un espace où la croissance des effectifs s'opère à un rythme significativement plus ralenti qu'au sortir de la Deuxième Guerre mondiale. C'est plus particulièrement le cas dans les Balkans de l'Ouest où les crises des années 1990 ont marqué un très sévère coup d'arrêt à une croissance jusqu'alors supérieure à 1 % par an (Fig. 1).

Figure 1. *Bassin méditerranéen, 1950-2015. Taux de croissance démographique (en %).*



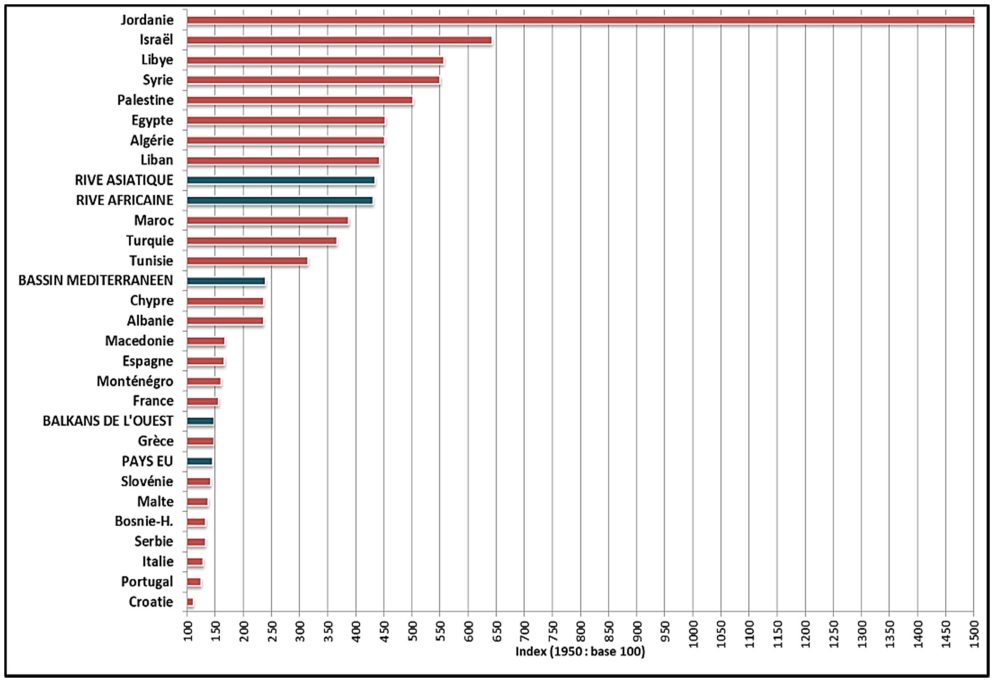
Source: *World Population Prospects: The 2017 Revision*.

Dans un espace dont la population a été globalement multipliée par un facteur 2,4 de 1950 à 2015, la situation apparaît fort contrastée entre une rive nord, où les effectifs n'ont progressé que de 50 % en 60 ans, et les rives est et sud où la croissance a été 7 à 8 fois plus vive (Fig. 2). À l'intérieur de ces grands sous-ensembles géographiques, la diversité des rythmes de croissance est tout aussi accusée. Au Nord, au sein des pays membres de l'Union européenne, le rapport est de 2 à 1 entre Chypre (indice 236) et la Croatie (111); dans les Balkans de l'Ouest, entre l'Albanie (236) et la Serbie (133), il est de 1,7 à 1. Au Sud, la croissance de la population a été 1,7 fois plus rapide en Lybie (557) qu'en Tunisie (315). À l'Est, par rapport au Liban (442) et à la Turquie (367), le différentiel de croissance a approché 1,4 avec Israël (643) et la Syrie (551), mais il a culminé à 4,4 avec la Jordanie (1911).

La diversité de la croissance démographique dans le Bassin méditerranéen ne tient pas cependant au seul rythme à long terme de celle-ci ; elle réside également dans sa nature, plus ou moins intrinsèque et/ou étrangère.

Sur la période 2010-2015, la différenciation est forte, sur la rive nord, entre les États membres de l'Union européenne et ceux des Balkans de l'Ouest. Ces derniers ont tous enregistré, en effet, un déficit migratoire qui, soit s'est ajouté au déficit naturel (Bosnie-et-Herzégovine), soit a limité son impact sur la croissance globale (Monténégro, ARYM), soit l'a surcompensé et entrainé un déclin général (Albanie, Serbie). *A contrario*, les pays de l'UE ont tous bénéficié – hormis la Croatie qui, par-delà son adhésion récente à l'espace communautaire (1er juillet 2013), partage encore beaucoup avec les autres États des Balkans de l'Ouest – d'un apport net de population qui a contrebalancé leur déficit naturel (Italie, Portugal) ou l'a peu ou prou amplifié, à l'instar de l'Espagne ou de la France. Sur les rives est et sud de la Méditerranée, tous les pays présentent un excédent naturel et seuls les différencient leurs soldes migratoires. Systématique au Sud, sans pour autant contrebalancer la croissance intrinsèque, l'émigration nette ne concerne, à l'Est, que la Palestine et la Turquie, avec un effet relativement limité sur la croissance globale.

Figure 2. *Bassin méditerranéen, 1950-2015.Croissance démographique indiciaire (1950: base 100).*



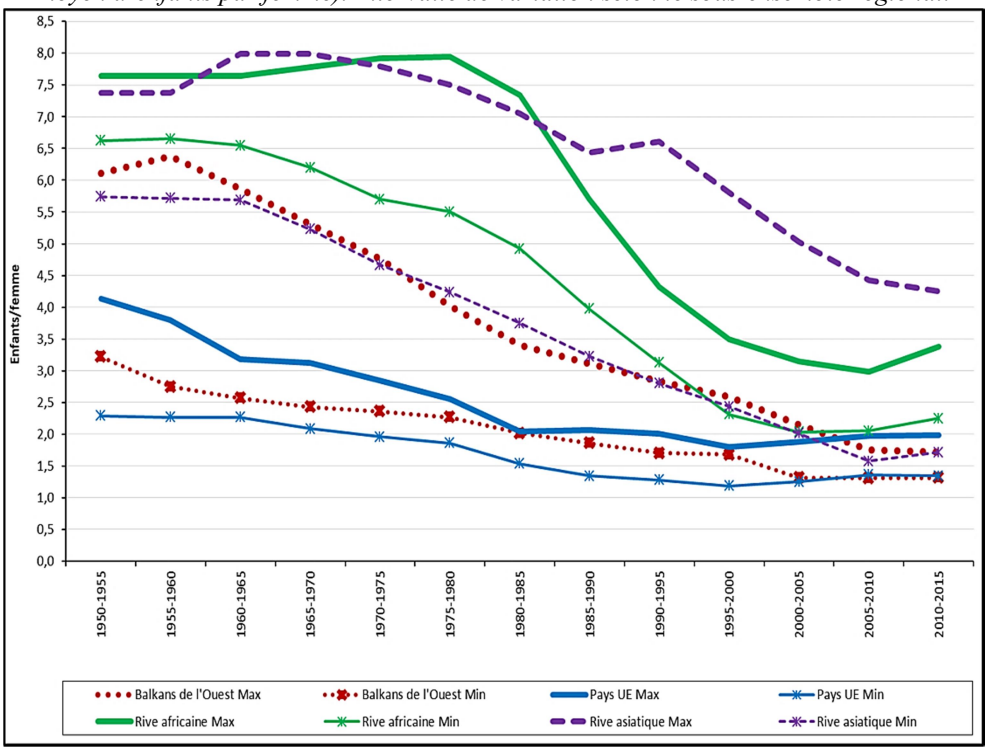
Source: *World Population Prospects: The 2017 Revision*.

3. Fécondité, mortalité, migrations : tendances lourdes et disparités

3.1 Fécondité

Sur la longue période, la tendance de la fécondité s'est nettement infléchie à la baisse et, à l'échelle du Bassin méditerranéen dans son entier comme à celle des sous-ensembles régionaux qui le composent, le spectre des niveaux atteints se resserre (Fig. 3). Débuté plus tardivement que sur la rive européenne, le processus de baisse s'est accéléré sur les rives asiatique et africaine depuis les années 1980 et si, dans des pays tels que la Palestine, la Jordanie, la Syrie, l'Egypte et Israël, l'indicateur conjoncturel de fécondité avoisine ou excède encore 3 enfants en moyenne par femme, au Liban, en Tunisie ou en Turquie, il se situe désormais à des niveaux qui ne diffèrent plus guère, sinon plus du tout, de ceux enregistrées dans les pays de l'UE ou des Balkans de l'Ouest.

Figure 3. *Bassin méditerranéen, 1950-2015. Indicateur conjoncturel de fécondité (nombre moyen d'enfants par femme). Intervalle de variation selon le sous-ensemble régional.*



Source: *World Population Prospects: The 2017 Revision*.

Partout, la baisse d'intensité de la fécondité s'accompagne, quand elle n'en est pas la conséquence, d'un allongement du calendrier. À peine plus précoce dans les

Balkans, en Turquie et en Égypte, l'âge moyen à la maternité est aujourd'hui très largement inclus dans la fourchette 29-31 ans, indépendamment du développement socio-économique des pays, de leur culture, de la prégnance de leurs coutumes ou de l'intensité des pratiques religieuses. La tendance à l'augmentation de la durée des études, des filles plus particulièrement, la crise durable des économies, les troubles politiques et les conflits plus ou moins ouverts sont autant de facteurs à l'origine de retards amples et nombreux dans la mise en couple des jeunes. Quand, par voie de conséquence, ils ne sont pas tout simplement annulés, les projets de descendance sont différés et peu ou prou revus à la baisse.

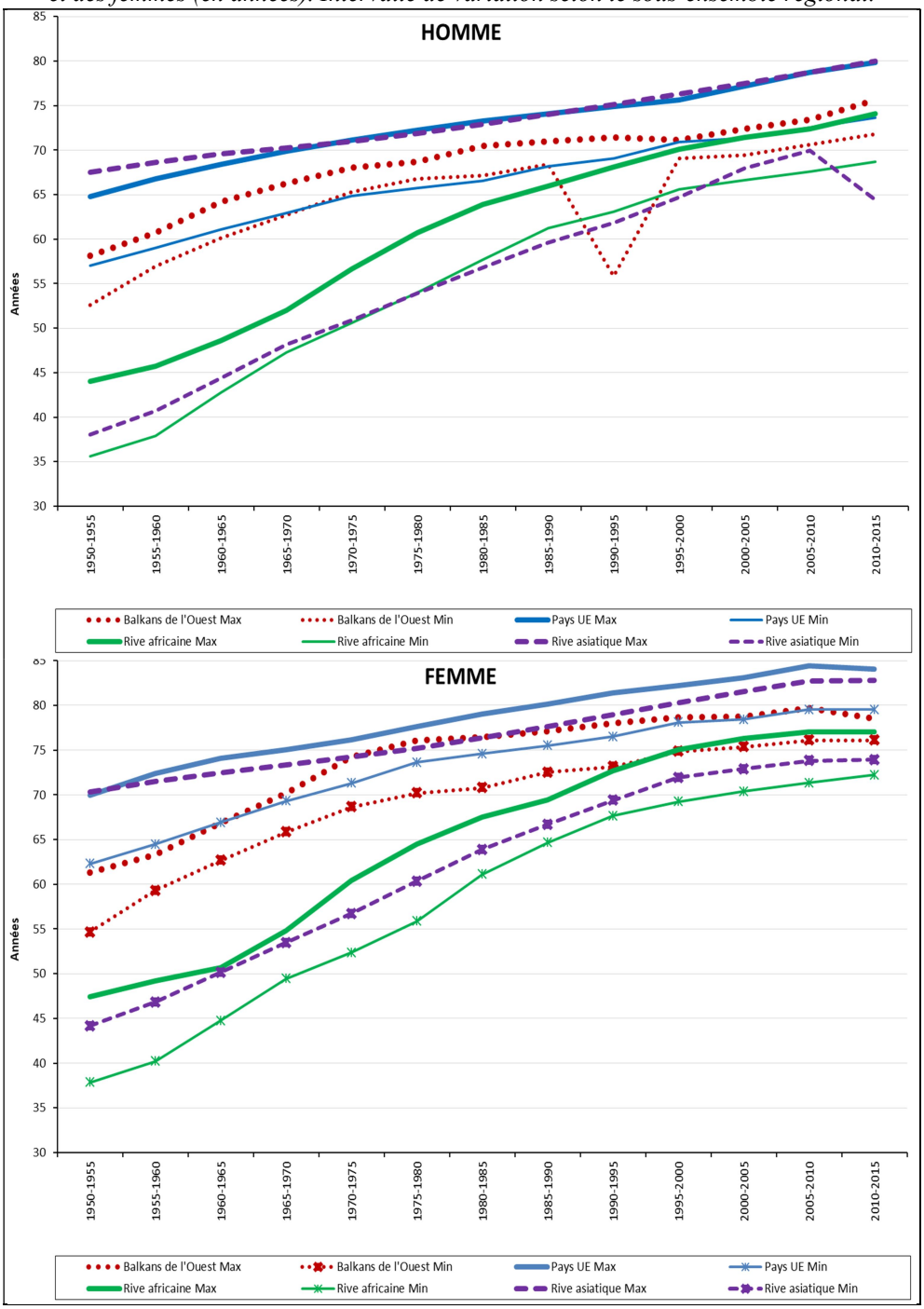
Telle qu'elle s'est déjà produite, la baisse de la fécondité a fortement entamé le potentiel de renouvellement des générations féminines en âge de procréer. Parmi les pays de la rive Nord de la Méditerranée, seule la France assurait encore en 2010-15, dans les conditions de survie des filles jusqu'à l'âge moyen à la maternité données pour la période, le remplacement nombre pour nombre de sa population de femmes âgées de 15 à 49 ans (taux net de remplacement égal à 1). Sur les rives asiatique et africaine, le Liban n'y parvenait plus, la Turquie, la Tunisie et le Maroc, qui y réussissaient à peine, étaient sous la menace d'une nouvelle baisse de fécondité non compensée par une amélioration de la survie dans les premiers âges de la vie des femmes.

3.2 Mortalité

Dans le Bassin méditerranéen, depuis les années 1950, la tendance est manifeste et générale à l'allongement des durées de vie moyennes et, comme pour la fécondité, au resserrement des écarts entre les valeurs extrêmes, y compris au sein des différents sous-espaces (Fig. 4). Celui des Balkans de l'Ouest est le seul où la progression vers des durées de vie plus longues a temporairement été perturbée, l'inflexion subite à la baisse de l'espérance de vie à la naissance, au début des années 1990, ayant essentiellement concerné la Bosnie-et-Herzégovine et l'Albanie et davantage prévalu pour les hommes que pour les femmes.

En 2010-2015, c'est sur la rive asiatique que les écarts absolus d'espérances de vie à la naissance entre les pays étaient les plus importants, Israël se démarquant nettement dans cette sous-région de la Turquie (pour les hommes) et de la Palestine (pour les femmes). Sur la rive africaine, où le calendrier de la mortalité était globalement le plus précoce, les écarts avoisinaient 6 ans. Au regard des calendriers de la mortalité, une profonde démarcation séparait l'Ouest et l'Est de la rive européenne. À l'Ouest, les durées de vie moyennes étaient, hors la Croatie, substantiellement plus longues et les écarts entre pays limités à 3 ans.

Figure 4. *Bassin méditerranéen, 1950-2015. Espérance de vie à la naissance des hommes et des femmes (en années). Intervalle de variation selon le sous-ensemble régional.*



Source: *World Population Prospects: The 2017 Revision*.

Dans les Balkans de l'Ouest, auxquels peut être rattachée en l'occurrence la Croatie, les écarts absolus d'espérances de vie à la naissance entre pays avoisinaient également 3 ans, mais dans un intervalle de valeurs nettement plus faibles.

D'une manière générale, la corrélation est nette entre durée de vie moyenne et taux de mortalité infantile[2]: plus celui-ci est faible et plus l'espérance de vie tend à être longue. Du fait cependant de sa baisse générale et rapide depuis les années 1950, le taux de mortalité infantile a atteint un peu partout en Méditerranée des niveaux relativement bas, sinon des niveaux très proches de l'incompressibilité, comme dans les pays de l'UE. À l'évidence, les réserves de survie sont désormais extrêmement limitées dans les premiers âges de la vie et, pour y puiser, les efforts seront de plus en plus difficiles et coûteux.

3.3 Migrations

La Méditerranée, espace carrefour s'il en est, a toujours été, aux dires des historiens pour les temps passés et des statisticiens pour les périodes plus contemporaines, le lieu de mouvements plus ou moins intenses de population. Au jeu de l'attractivité/répulsion, les cartes paraissent aujourd'hui clairement distribuées, comme on l'a précédemment noté, entre les pays de la rive africaine et des Balkans de l'Ouest, pourvoyeurs nets de migrants, et ceux de l'Union européenne ou de la rive asiatique (à l'exception de la Palestine et de la Turquie), hébergeurs nets de migrants.

On ne saurait insister, ici, sur le fait que ce tableau ne vaut que pour la période pour laquelle il a été dressé. Ainsi, n'est-ce que depuis le début des années 1970 que le Portugal, l'Espagne, l'Italie et la Grèce enregistrent des excédents migratoires ; jusque-là, ces pays perdaient des habitants et leurs déficits migratoires excédaient alors bien souvent, en valeurs absolues, leurs excédents actuels, pour partie constitués par d'anciens émigrants. Le phénomène migratoire s'avèrant très sensible aux modifications de l'environnement socio-économique et moins inerte que la fécondité ou la mortalité, sa propagation est soumise à des changements rapides, importants, encore plus sensible dans les petites territoires.

Par ailleurs, les bilans nets masquent des réalités bien plus complexes, tous les pays étant simultanément hébergeurs et pourvoyeurs de migrants ; des migrants qui, au demeurant, ne sont parfois qu'en transit là où ils sont recensés. Pour afficher des taux de migration nette négatifs, le Maroc, l'Algérie ou la Tunisie n'en

[2] Le taux de mortalité infantile mesure la probabilité pour un enfant né vivant de décéder avant son premier anniversaire.

sont pas moins des pays d'accueil pour de nombreux migrants subsahariens. La Turquie connaît une situation analogue vis-à-vis de migrants entrés par ses frontières de l'est. Et, parmi les pays à taux de migration nette positif, la France est quittée pour des raisons diverses par certains de ses résidents nationaux et par des résidents étrangers pour lesquels elle ne constitue qu'un lieu de passage vers le Royaume-Uni ou le continent nord-américain.

N'excédant que très exceptionnellement 10‰, les taux de migrations nettes estimés pour les pays du Bassin méditerranéen peuvent sembler relativement faibles. Ce n'est pas pour autant que leurs effets le sont. Il a, par exemple, été montré que sans l'immigration des années 1960-1998, la population de la France n'aurait pas été en 1999 plus nombreuse qu'en 1975, soit 11 % plus faible que la population effectivement recensée, et le nombre de naissances en 1998 aurait été abaissé de 738.000 à 542.000, soit un déficit de 26,5% (Tribalat, 2005).

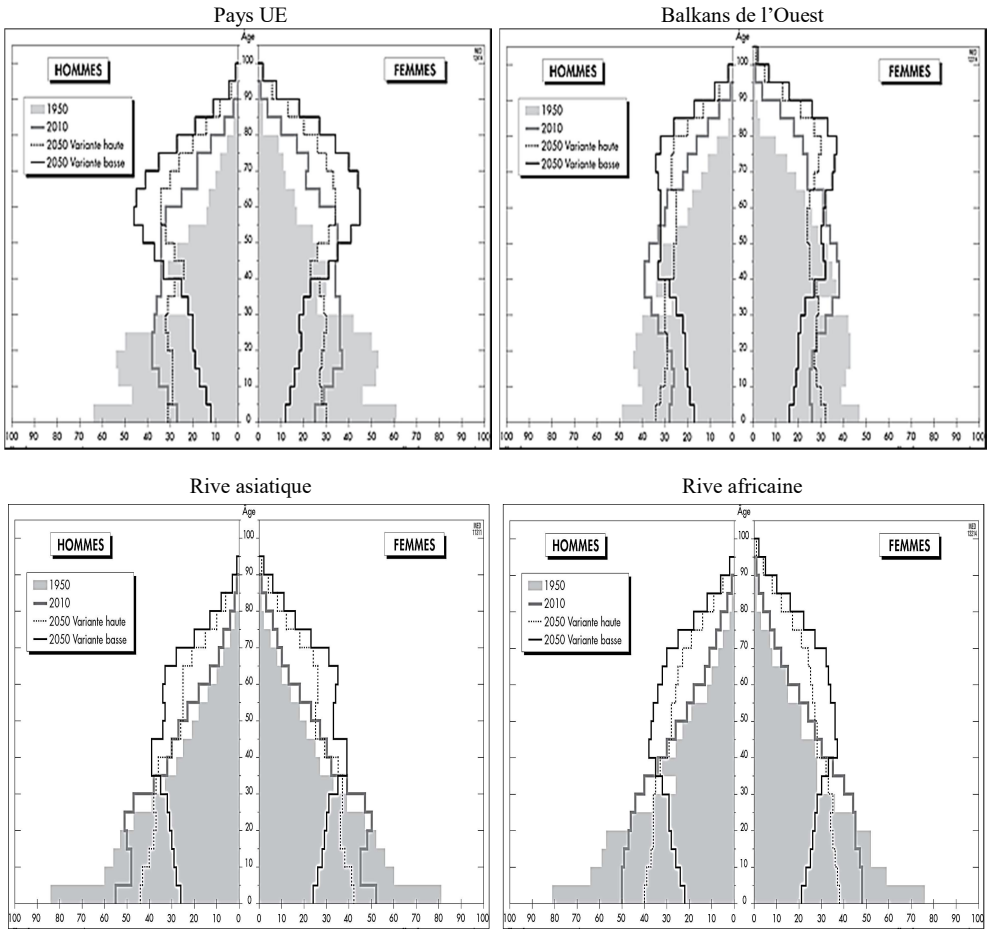
4. La transition de la structure par âge: vers un vieillissement généralisé

Confrontés à une baisse de la fécondité ou au maintien durable de celle-ci à des niveaux bas et à un allongement des espérances de vie, tous les pays du Bassin méditerranéen connaissent un vieillissement de leur population: baisse du poids des jeunes, hausse du poids des personnes âgées, plus ou moins marginalement atténué ou, au contraire, accentué par le jeu migratoire. La mécanique du vieillissement des ensembles humains, considérés sous l'angle des seuls facteurs démographiques directs –et non de ceux, économiques ou sociaux, qui les influencent–, est parfaitement connue (Nations unies, 1956; Parant, 1978).

À l'œuvre depuis plusieurs décennies dans les pays situés le plus à l'Ouest de la rive européenne, le processus d'inversion des pyramides gagne les pays des autres rives et la tendance va se poursuivre, comme en attestent les plus récentes perspectives élaborées par la Division de la population des Nations unies (Fig. 5).

La tendance à l'allongement des durées de vie étant par ailleurs censée se poursuivre, que l'on se place dans un environnement de fécondité soutenue (hausse dans les pays plus développés, déclin à pas ralenti dans les pays moins développés) –variante haute des projections– ou dans un environnement de fécondité universellement déprimée –variante basse des projections–, la croissance des effectifs sera dans tous les pays du Bassin méditerranéen plus rapide au sommet des pyramides des âges qu'à la base.

Figure 5. *Bassin méditerranéen 1950-2010-2050. Pyramides des âges des sous-ensembles régionaux, estimées et projetées selon deux variantes contrastées.*

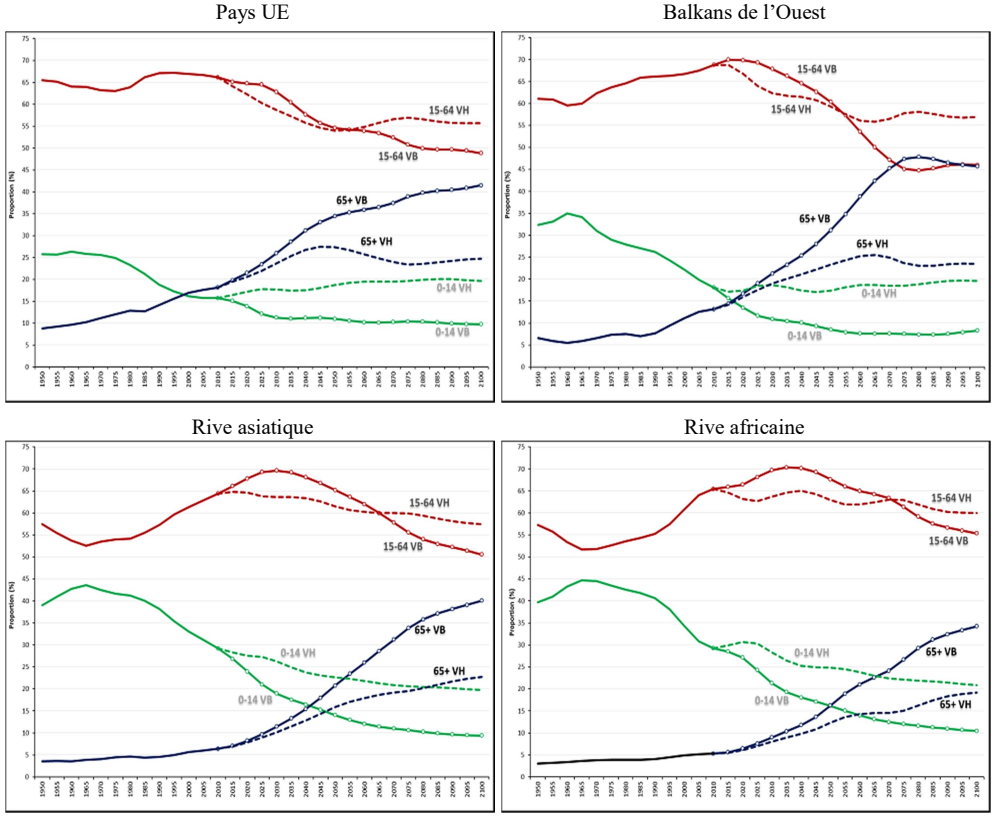


Source: *World Population Prospects: The 2017 Revision.*

Tandis que les proportions de jeunes de moins de 15 ans pourraient, au mieux, se maintenir à long terme (variante haute), celles des 65 ans ou plus devraient partout fortement progresser (Fig. 6).

Les personnes jeunes et âgées étant, dans leur très large majorité, à la charge des personnes adultes, l'évolution de leurs nombres et de leurs proportions au fil du temps en réponse à l'évolution de la fécondité (modifications d'intensité et de calendrier) et de la mortalité (modifications de calendrier) et, plus secondairement, des échanges migratoires, n'est pas neutre quant aux effets exercés, plus particulièrement au plan socio-économique.

Figure 6. *Bassin méditerranéen 1950-2010-2050. Proportion par groupes d'âge estimées et projetées selon deux variantes contrastées.*



Source: *World Population Prospects: The 2017 Revision*.

5. Les sociétés méditerranéennes à l'épreuve du vieillissement

Dans le bassin méditerranéen comme ailleurs, le vieillissement des populations est certes à appréhender comme le résultat de deux tendances communément estimées heureuses: la maitrise plus ou moins intégrale de la fécondité et l'allongement de la vie induit par la baisse des probabilités de décès jusque dans les âges les plus élevés. Mais cet antihasard par excellence est surtout à percevoir comme un redoutable défi que les sociétés de tous les pays –les plus développés comme les moins développés– doivent s'astreindre à relever au plus vite et avec d'autant plus de détermination qu'elles l'ont jusqu'à présent durablement ou trop profondément occulté.

Les enjeux qui lui sont associés débordent, en effet, de la question des retraites ou de celle de la dépendance du Grand Âge auxquelles on tend à les circonscrire dans les pays dotés de systèmes de protection sociale plus ou moins généreux mais

tous aujourd'hui fortement mis à mal par une crise économique durable et sévère. Les enjeux du vieillissement interfèrent avec ceux relatifs à l'emploi et à ses mutations, à l'évolution de la population active, sa formation, son propre vieillissement, son management. Ils renvoient aux modalités de fonctionnement des économies nationales et des leurs rapports avec le reste du monde, dans un contexte de grande dérégulation et d'émergence de nouveaux pôles. Ils interpellent sur le devenir de l'offre et de la demande des biens et services traditionnels (dont une population comptant une forte proportion de personnes âgées peut être partiellement ou totalement saturée) ou nouveaux (tout autant soumis que les précédents à des effets d'âge, de génération ou de période). Ils interagissent avec l'évolution des jeux de pouvoirs (déclin des corps intermédiaires, montée du radicalisme) et des rapports de force entre générations successives (effets d'éviction des jeunes adultes, des femmes en âge de procréer). Ils conduisent immanquablement à des questionnements d'essence plus philosophiques sur le sens de la vie et sur le droit de tout être humain à une fin de vie et à une mort dignes.

Dans les pays moins développés, le défi du vieillissement revêt ces différents aspects, mais il se pose en termes encore plus prégnants. Car dans ces pays, tardivement mais brutalement gagnés par la révolution démographique[3], un nouveau modèle familial va en outre se substituer à l'ancien à l'horizon d'à peine une vie. Initialement large et répartie tout au long de l'échelle des âges, la famille *largo sensu* va très vite se contracter et se concentrer sur des plages d'âges de plus en plus étroites (du fait de la réduction la taille des fratries et de l'écart d'âge entre les aînés et les cadets) et de plus en plus espacées (en raison de l'augmentation de l'âge moyen des mères à la naissance des enfants). Dans ces pays dépourvus de système de protection sociale susceptibles de jouer le rôle d'amortisseur des crises, les structures et l'entraide familiale vont aussi être profondément impactées par la migration massive des jeunes ruraux vers les villes, des villes qui ne leur offrent, le plus souvent, que chômage, emplois temporaires sous-qualifiés et sous-payés, logements de misère. Confrontés au sous-emploi d'une jeunesse très nombreuse et de plus en plus instruite, incapables de créer la richesse suffisante pour faire face à leurs besoins du moment, ces pays risquent de basculer dans le vieillissement sans

[3] Au sens où l'entendait Adolphe Landry de l'avènement d'un régime démographique caractérisé par une pratique généralisée de la limitation des naissances répondant au souci essentiel, non plus de maintenir un niveau de vie, mais de l'élever au profit des parents et de la progéniture. Cet avènement marque le basculement d'un monde où l'on "tendait vers une égalisation de la mortalité et de la natalité, vers un état de la population destiné à demeurer par la suite stationnaire" à un autre monde où il n'y a plus d'équilibre et où "on pourra même voir la population décroître, malgré les progrès, si remarquables seraient-ils, soit de la technique reproductive, soit de la médecine et de l'hygiène".

avoir pu constituer de réserves ; un scénario synonyme, s'il se réalisait, d'extrême détresse pour les futures populations âgées.

6. Plaidoyer pour une prospective du vieillissement et des questions de population en méditerranée

Pour des motifs divers, d'ordre technologique, économique ou financier, culturel ou politique, le changement s'accélère, les interdépendances s'accroissent, les risques de rupture se multiplient, l'incertitude et l'imprévisibilité se renforcent. Contrairement au passé, simple lieu de faits connaissables sur lesquels nous ne pouvons rien, et contrairement au présent, insaisissable et que nous ne pouvons que traverser plus ou moins bien, l'avenir n'est pas déjà fait. Il n'est pas prédéterminé, mais, au contraire, plus que jamais ouvert à plusieurs futuribles (pour user du concept forgé naguère par Bertrand de Jouvenel, par contraction de futur et possible).

Tant pis pour le sujet connaissant, qui doit faire l'apprentissage de l'incertitude. Tant mieux pour le sujet agissant, pour qui les plages d'indétermination constituent autant d'espaces de liberté, de marges d'autodétermination.

L'éventail des futurs possibles étant non seulement ouvert, mais se déformant sans cesse – des futuribles disparaissent tandis que d'autres émergent –, le sujet connaissant doit se forcer à un effort incessant de veille. Il lui faut repérer, analyser et évaluer les tendances lourdes, celles qui s'inscrivent dans le temps long passé et possèdent une forte inertie à court et moyen termes. Il lui faut également s'interroger sur les incertitudes majeures, qui ouvrent plus ou moins largement et, parfois, à très court terme, le spectre des évolutions futures de certaines variables. Il lui faut encore traquer ce que Pierre Massé nommait le fait porteur d'avenir (on parle plutôt aujourd'hui de signal faible), "signe infime par ses dimensions présentes, mais immense par ses conséquences virtuelles, qui annonce une mutation technique, économique ou sociale".

Pour cet effort de veille, le sujet connaissant doit:

- définir des indicateurs pertinents et user des sources d'information disponibles avec un minimum d'esprit critique,
- éviter de privilégier certains schémas mentaux, principes a priori et autres idées régnantes,
- croiser les points de vue,
- éviter de télescoper les horizons temporels et se garder d'occulter les calendriers, même totalement discordants, des phénomènes et des événements observés,

- s'abstraire au maximum de l'attraction du présent et de la préférence très (trop) marquée pour le court terme.

Mais un effort de veille est dépourvu de sens s'il n'est pas au service d'une intention, ce que Sénèque signifiait par : "Il n'y a pas de vent favorable pour celui qui ne sait où il va" (Lettres à Lucilius).

Au sujet agissant, il faut une raison motrice, un système de valeurs lui permettant de fixer des objectifs, de se forger une vision d'un futur souhaitable : une chose à faire, encore au stade de simple image jetée en avant et très éloignée d'un véritable projet. Un projet, au sens que lui confèrent les prospectivistes, est l'expression d'un vouloir qui, pour être accompli, s'inscrit nécessairement dans la durée, une durée d'autant plus longue que la mise en œuvre du projet implique une rupture avec l'ordre existant, une mobilisation de moyens pas forcément disponibles dans l'instant. Là intervient l'équation subtile entre le rêve et la raison, le premier engendrant des "visions" d'un avenir meilleur; visions qui, passées au crible de la raison (d'aucuns parlent d'études de faisabilité), deviendront les véritables moteurs de l'action.

Appréhender la réalité à travers ses multiples aspects ; intégrer la dimension du temps long (passé et à venir) pour révéler la dynamique profonde des systèmes ; postuler l'instabilité, la discontinuité et les ruptures (subies et volontaires) ; élaborer, à un horizon dûment choisi (en fonction de l'inertie du système, de l'échéan–cier des décisions à prendre, du pouvoir de décision et des moyens d'action, du degré de crispation et de motivation es acteurs), des scénarios contrastés de type exploratoire (qui défrichent le champ des futurs possibles et éclairent sur ce qui pourrait advenir) ou de type stratégique (qui visitent le champ des futurs désirés et dictent le compte à rebours des actions à entreprendre); comparer les avantages et les inconvénients des stratégies possibles ; enfin, arbitrer entre ces dernières: autant d'étapes obligées de "l'invention du futur".

Au-delà de la théorie, comment se présente dans les faits cette "invention du futur" en Méditerranée?

- *L'invention du futur implique une connaissance préalable la meilleure, c'est-à-dire la plus variée et la plus approfondie, de la situation passée et présente.*

Au strict plan démographique, cette connaissance est loin d'être toujours assurée, quand elle n'est pas lacunaire. Mal assurée, sinon lacunaire, car les estimations de populations, indispensables entre autres pour le calcul d'indicateurs robustes, et la collecte des événements qui forment les numérateurs des taux ne présentent pas toujours la qualité requise, souffrent d'une trop grande ancienneté, sont partiellement ou totalement défaillantes. Ainsi, en est-il, par exemple, de la connaissance

des flux de migrants internationaux ; des flux dont les effectifs de migrants déduits des données de recensements et d'enquêtes spécifiques permettent d'affirmer qu'ils ne sont pas négligeables dans l'espace méditerranéen. Mais quid de l'ampleur des différents flux en fonction du pays de provenance ou du pays d'arrivée? Quid de la distinction flux de natifs/flux de non natifs? La réponse à de telles interrogations implique de disposer d'informations – lieu de naissance des migrants, lieu de provenance, nationalité à la naissance – rarement collectées simultanément et, de manière encore plus exceptionnelle, à une échelle infranationale, celle où les effets des mouvements de population sur la dynamique démographique et socio-économique des espaces d'arrivée et des espaces quittés s'avèrent pourtant les plus décisifs.

La connaissance est tout aussi imparfaite et incomplète s'agissant, pour se limiter à nouveau au seul champ démographique, des facteurs qui déterminent l'évolution et le niveau des trois composantes clés du renouvellement des populations (fécondité, mortalité, migrations) ou des facteurs explicatifs des unions/désu–nions et des variations qui peuvent affecter ces phénomènes au fil du temps selon les territoires et groupes d'individus.

La réflexion sur l'avenir requiert la mise à disposition du plus large public possible du plus grand nombre de données de qualité. Si les producteurs de données ne peuvent nier ni le caractère incomplet ou défectueux de certaines informations mises à disposition ni même, parfois, certaines velléités plus ou moins fondées et transparentes de changer la donne de collectes ayant pourtant largement fait la preuve de leur utilité et de leur efficacité, la responsabilité de cet état de fait incombe aussi, pour une part plus ou moins grande, aux utilisateurs. La qualité d'une statistique publique en phase avec les besoins de connaissance est, en effet, directement fonction du degré d'exigence de ses utilisateurs. C'est pleinement leur rôle de convaincre les producteurs de la nécessité de maintenir une certaine ardeur pour la production d'informations fiables, variées et complexes. Et les pouvoirs publics seraient, eux aussi, bien inspirés de pousser fermement dans cette direction.

- *L'invention du futur est une démarche pluridisciplinaire, d'inspiration systémique, qui s'efforce de tenir compte des phénomènes de discontinuité et de rupture.*

  On restera, là encore dans le champ démographique.

Quand bien même la collecte statistique satisferait tous les besoins, il faudrait encore que dans les exercices d'anticipation démographique, la fécondité, la mortalité et la mobilité – internationale ou régionale – des personnes ne soient pas appréhendées de manière totalement exogène, mais plutôt comme les très complexes micro-systèmes qu'elles sont en réalité. Dans les projections de population existantes,

les niveaux atteints, à tel ou tel horizon temporel, par les indicateurs choisis (indicateurs conjoncturels de fécondité ou descendances finales, quotients de mortalité par sexe et âge et taux de migration nette par sexe et âge ou soldes migratoires) sont des produits d'extrapolations et de choix plus ou moins raisonnés, à l'intérieur de spectres généralement peu ouverts et arrêtés sans préoccupation des conditions de leur réalisation. Ils ne résultent pas de la confrontation d'évolutions contrastées des variables explicatives premières de chacun des trois sous-systèmes clés. Or, ces derniers incorporent des éléments fondamentaux en provenance des sphères sociale, culturelle, législative, économique, technologique, que l'on ne saurait négliger; et, peut-être, moins dans le Bassin méditerranéen qu'ailleurs.

Il est peu probable que les perspectivistes des offices statistiques nationaux ou des instances supranationales spécialisées soient spontanément à la veille de cesser de confondre un fonctionnement plus ou moins performant d'un sous-système donné – fécondité, mortalité ou mobilité des personnes, pour ce qui nous occupe ici – avec une transformation structurelle du sous-système en question ; à la veille d'appréhender un "significativement plus" ou un "significativement moins" de la même chose (enfant par femme, année de vie, migration nette) comme une chose radicalement différente de la chose initiale ; à la veille d'être dans une véritable posture de prospectiviste.

Sans doute faudra-t-il les "bousculer" plus ou moins rudement pour qu'ils adoptent ce parti-pris élémentaire de la prospective : "mieux vaut une approximation grossière mais juste, plutôt qu'une prévision très fine mais erronée" (Hugues de Jouvenel, 2004) et pour qu'ils puissent ne plus faire un usage abusif du terme "scénario contrasté". Les scénarios contrastés s'articulent sur des configurations morphologiques différentes. "Ce n'est pas un peu plus ou peu moins de la même chose mais une autre chose, une autre histoire bâtie à partir de transformations structurelles du système" (Hugues de Jouvenel, 2004).

- *L'invention du futur est l'expression d'une volonté politique forte.*

Depuis la phase d'expression première d'un projet d'action encore vague pour un territoire aux contours mal définis sur lequel les acteurs sont multiples, les pouvoirs partagés et les intérêts conflictuels, jusqu'à celle de l'évaluation itérative de la stratégie finalement proposée, cette volonté a maintes occasions de se manifester et d'être conséquemment appréciée.

Les projets ne naissent pas et n'ont *a priori* aucune chance d'aboutir là où n'existent pas de lieux d'expression citoyenne, là où la dissidence intellectuelle ne peut se faire entendre. Favoriser la multiplication de ces lieux constitue un premier devoir pour un politique soucieux de changer une donne territoriale. Son deuxième

consiste en la recherche et en la mutualisation d'un savoir, à faire remonter de sources multiples, à transformer. Le politique doit, en l'occurrence, composer avec les pesanteurs technocratiques et faire en sorte que toute la connaissance disponible, sinon nécessaire, puisse être organisée – aux bons soins d'une structure *ad hoc* – en une base de données rapidement opérationnelle pour permettre l'élaboration d'un diagnostic partagé puis celle des divers scénarios indispensables à l'esquisse d'une stratégie. Mais pour que le projet s'enracine dans son territoire, il faut encore que le politique, en dernier et ultime ressort, tranche en faveur d'une stratégie et qu'il accepte le principe de son évaluation future. C'est à cette double aune que sera finalement appréciée la profondeur de son attachement au projet initial et, d'une certaine manière aussi, son authenticité de politique.

Pour insuffler de la vie dans une société, il faut un chef d'orchestre. Le politique est naturellement fait pour endosser l'habit. Il n'a guère pour cela à témoigner de connaissances particulières, ni même d'un esprit tout spécialement visionnaire ; d'autres, dont il saura s'entourer, en auront pour lui et pourront les lui faire partager. Il doit, en revanche, manifester une attirance innée pour le bien public et une aptitude à trancher pour ce dernier toujours fermement et en temps opportun.

Bien conscients que du projet révélé à l'élaboration d'un plan puis d'un programme d'action, il puisse y avoir beaucoup, sinon trop –comme en attestent l'enlisement de la stratégie de Lisbonne et l'échec de l'Union pour la Méditerranée–, on ne nous fera pas accroire cependant que cette espèce d'hommes et de femmes est, aujourd'hui, en voie de disparition dans cet espace que l'on nommait jadis *Mare nostrum*.

7. Conclusions

En Méditerranée, la raréfaction des enfants –qu'on la subordonne, comme Adolphe Landry, à un principe individuel de rationalisation de la vie, ou bien, comme Frank Notestein et Kingsley Davis, théoriciens de la transition démographique, à l'émergence d'un mode moderne de développement économique– induit un vieillissement démographique d'autant plus intense et subit qu'elle est prononcée et rapide. Lorsque sur ce processus de raréfaction de l'enfance se greffe une tendance soutenue à l'allongement des durées de vie, le vieillissement démographique résultant peut alors très vite confiner à un redoutable défi pour les sociétés concernées.

La tendance au vieillissement n'est pas nouvelle en Méditerranée; elle affecte de longue date les populations de la rive Nord. Mais elle est aujourd'hui générale,

indépendamment du rythme auquel elle se réalise, de ses modalités et de ses effets, variables selon les États ; et elle est appelée à se renforcer au cours des prochaines décennies, fut-ce à des degrés divers. Se tonifiant, alors que la mondialisation est exacerbée par l'émergence de nouveaux pôles, que la concurrence est sans cesse plus âpre entre des économies plus ou moins sévèrement mises à mal par une succession de crises, que certains États sont en déficit de stabilité politique et de gouvernance, que d'autres s'ignorent quand ils ne se font pas la guerre, cette tendance requiert une attention toute spéciale.

Les sociétés méditerranéennes peuvent envisager d'instaurer des politiques ou des mesures d'accompagnement limité de ces différents processus perçus comme constituant une donne sur laquelle elles estiment n'avoir aucune prise ou qu'elles n'entendent pas modifier ; c'est l'option que l'on peut qualifier de «dictature des processus» ou «d'accompagnement au fil de l'eau». Les sociétés concernées peuvent au contraire adopter des politiques et des mesures beaucoup plus volontaristes, jugeant que les enjeux liés aux processus démographiques sont trop primordiaux et périlleux pour accepter qu'ils suivent librement leurs cours; c'est l'option «politiques interventionnistes» ou de «soumission du principe de plaisir au principe de réalité».

Dans les faits, les sociétés méditerranéennes ont d'abord joué sur le mode gestion douce de processus dont elles ne cherchaient guère par ailleurs à saisir les différents tenants et aboutissants en dépit de quelques écrits de démographes exerçant en cela pleinement leur fonction d'alerte. Cette réponse a minima n'ayant eu que des résultats très temporaires et/ou limités, l'obligation se fait aujourd'hui très vivement sentir de recourir à des politiques plus invasives. Ce ne sont pas les champs d'intervention qui manquent (Carella, Parant, 2018).

Bibliographie

Carella, M., Parant, A. (2016). Age-Structural Transition and Demographic Windows around the Mediterranean. In: R. Pace, R. Ham-Chande (eds.), *Demographic Dividends: Emerging Challenges and Policy Implications*, Springer.

Carella, M., Parant, A. (2018). La transition de la structure par âge dans les Balkans. In B. Kotzamanis, A. Parant (eds.), *Regards sur la population du Sud-Est d'Europe,* Presses Universitaires de Thessalie, Volos (en presse).

Davis, K. (1945). The World Demographic Transition. *Annals of the American Academy of Political and Social Science*, New York.

De Jouvenel, B. (1964). *L'art de la conjecture*. Éditions du Rocher, Monaco.

De Jouvenel, H. (2004). *Invitation à la prospective. An invitation to Foresight*, Futuribles, Collection Perspectives, Paris.

Landry, A. (1934). *La Révolution démographique. Études et essais sur les problèmes de population*. Éditions Sirey, Paris.

Lesthaeghe, R. (1995). The Second Demographic Transition in Western Countries: an interpretation. In: K.O. Mason, A-M. Jensen (eds.), *Gender and family changes in industrialized countries*. Clarendon Press, Oxford.

Lesthaeghe, R. (2010). The Unfolding Story of the Second Demographic Transition. *PSC Research Report*, 10: 696, University of Michigan.

Masse, P. (1962). Planification et prévision. *La Table ronde*, 117.

Malanima, P. (2013). *Rapporto sulle economie del Mediterraneo.* Istituto di Studi sulle Società del Mediterraneo (ISSN-CNR), Il Mulino, Bologna.

Notestein, F. (1944). Problems of Policy in Relation to Areas of Heavy Pressure. *Milbank Memorial Fund Quarterly*, 22 (4).

Notestein, F. (1945). Population: The Long View. In: T. Schultz Theodoren (ed.), *Food for the World*, University of Chicago Press, Chicago.

Pace, R.; Ham-Chande, R. (2016). *Demographic Dividends: Emerging Challenges and Policy Implications.* Springer.

Parant, A. (2000). La situation démographique de l'Europe du Centre-Sud : les lacunes de la connaissance. In: *Avoir 20 ans dans 20 ans en Méditerranée.* Marly-le-Roi, INJEP: 201-206.

Rowland, D.T. (2012). Population aging: the transformation of societies. Springer Dordrecht-Heidelberg-New York.

Van de Kaa, D.J. (1987). Europe's Second Demographic Transition. *Population Bulletin*, 42 (1).

Vallin, J. (2012). Faut-il une politique de population? *Population et sociétés*, 489.

Tribalat, M. (2005). Fécondité des immigrées et apport démographique de l'immigration étrangère. In: C. Bergouignan, C. Blayo, A. Parant, J.P. Sardon, M. Tribalat (eds.), *La population de la France. Évolutions démographiques depuis 1946*, tome 2, CUDEP: 727-767.

Tabutin, D. ; Schoumaker, B. (2005). La démographie du monde arabe et du Moyen-Orient des annees1950 aux années 2000. *Population*, 60 (5-6): 611-724.

Nations Unies, Division de la population (2017). *World Population Prospects : The 2017 Revision*, New York

UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO *Metodi e analisi statistiche 2017* ISBN: 978-88-6629-013-1
DIPARTIMENTO DI ECONOMIA E FINANZA pp. 253-262

Tecniche di Machine Learning per una previsione finanziaria

Najada Firza[1]*, Alfonso Monaco[2]

[1] *Università Nostra Signora del Buon Consiglio (Tirana, Albania)*
[2] *Istituto Nazionale di Fisica Nucleare (sezione di Bari)*

Riassunto: Il presente lavoro, propone un modello di previsione basato su classificatori di machine learning che utilizzano tra le variabili input anche i risultati ottenuti da una previsione statistica. Tutto ciò per migliorare ed efficientare la previsione di uno degli indici più rilevanti nel mercato finanziario degli Stati Uniti: il Dow Jones Average Index. Dapprima si utilizzerà un modello ARIMA che meglio si adatta ai dati a disposizione ed in seguito, da un set di variabili, verranno scelte quelle che più influenzano il Dow Jones Average Index tra variabili di commodity e altri indicatori finanziari principali. I classificatori di machine learning presi in considerazione per l'analisi e la previsione dei dati sono le Reti Neurali Artificiali e le Random Forest.

Keywords: Serie storiche, Indici finanziari, Reti Neurali Artificiali, Random Forest, Modelli ARIMA, Mercati finanziari.

1. Introduzione

Lo strumento finanziario è il bene scambiato in un mercato finanziario. In quest'ottica, tale bene è alla base di transazioni quotidiane ed è anche l'oggetto al quale, il mercato assegna un valore.

Il valore del bene finanziario non è mai costante nel tempo poiché frutto di una serie di variabili decisionali razionali e irrazionali, per tale motivo le aspettative del mercato rispetto al bene variano nel tempo. Questo meccanismo genera ed alimenta la volatilità del mercato finanziario.

* Autore corrispondente: nfirza84@gmail.com.

La volatilità del mercato quindi da una parte è costruita da:

- Logiche razionali derivate da calcoli scientifici e misure economiche corrette tratte dai bilanci aziendali;
- Variabili macroeconomiche di fondo.

Dall'altra parte, include le aspettative che il mercato riserva in un determinato arco temporale quindi rispecchia la fiducia che il consumatore ha negli strumenti finanziari in un determinato periodo temporale.

La previsione di un indicatore finanziario passa attraverso l'analisi e la previsione delle sue variabili che a loro volta possono essere esplicative oppure latenti nelle informazioni che il mercato ci fornisce.

Lo scopo di tale lavoro è l'analisi e la previsione di uno degli indici più importanti del stock market degli Stati Uniti: *Dow Jones Industrial Average*.

2. Dow Jones Average Index

Il Dow Jones Industrial Average oltre ad essere un indice storico e uno degli indici più noti della borsa di New York, è anche considerato dagli analisti finanziari un valido indicatore del mercato azionario americano.

L'indice è costruito come media ponderata dei prezzi di 30 azioni di società che appartengono alla categoria *Blue Chip*[1]. In tale categoria generalmente rientrano le società leader di ciascun settore di appartenenza.

La costruzione dell'indice inoltre è alquanto stabile poiché tiene conto di quelle particolari variazioni che possono destabilizzare un indicatore finanziario come ad esempio aumenti di capitale, fusioni e scissioni.

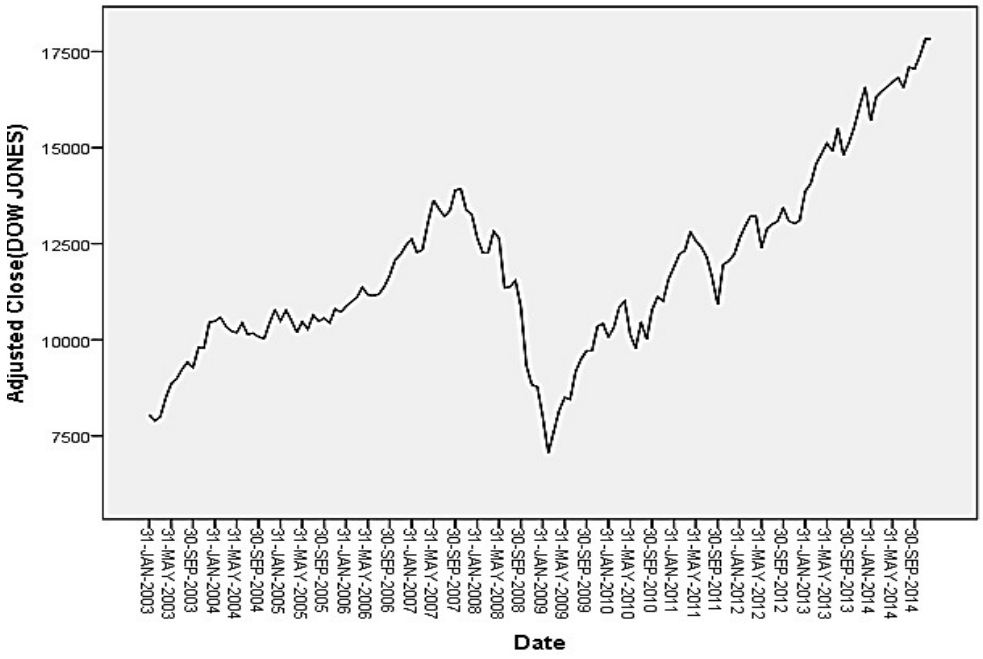
Le società che fanno parte dell'indice vengono selezionate da *The Wall Street Journal* anche e soprattutto a seconda delle condizioni del mercato. Sin dal 1928, il Dow Jones è un indice fondamentale per i mercati finanziari. Esso è considerato il barometro dell'economia e del mercato azionario degli Stati Uniti.

L'arco temporale preso da noi come riferimento va dal 2003 al 2014, i dati considerati sono medie mensili dello stesso indice.

La Fig. 1 mostra l'andamento tendenzialmente crescente dell'indice in un primo periodo (fino al 30/09/2007), una repentina caduta a partire da tale data fino al 31/01/2009 e poi una irregolare ma rapida risalita negli anni successivi.

[1] Il comparto Blue Chip accomuna le società ad alta capitalizzazione azionaria.

Figura 1. *Dow Jones Average Index 2004-2014*



3. Metodologia utilizzata

I metodi utilizzati per le analisi predittive sono i seguenti:

- ARIMA;
- Rete Neurale Artificiale;
- Random Forest.

3.1 Modello ARIMA

Il modello ARIMA[2], è un modello statistico che trae le sue origini dal modello classico ARMA (Box and Jenkins, 1970; Box *et al.*, 2015).

Nel modello generico ARIMA (p,d,q) si indica con:

p l'ordine dell'autoregressore;

d l'ordine dell'integrazione;

q l'ordine della media mobile.

In questo lavoro, abbiamo individuato nel modello ARIMA(3,1,0), la metodologia che meglio rileva le caratteristiche dell'indice per la previsione dello stesso.

[2] Acronimo di *AutoRegressive Integrated Moving Average.*

Inoltre, abbiamo deciso di utilizzare la *10-Fold Cross Validation Procedure*[3] affinché i risultati di previsione con il modello ARIMA (3,1,0) fossero più robusti.

Tale modello. così articolato, ha previsto i valori nel *range* 1/15 – 4/15 con un errore del 2,8% (MAPE, *mean absolute percentage error*).

Tabella 1. *Indice di Determinazione e MAPE per ogni campione di Cross Validation col modello ARIMA (3,1,0)*

10-Fold Cross Validation	R^2	MAPE
1	0.969	2.858
2	0.972	2.857
3	0.969	3.037
4	0.973	2.900
5	0.947	2.851
6	0.963	2.872
7	0.971	2.839
8	0.966	2.995
9	0.952	2.977
10	0.957	2.886

Data l'importanza che riveste l'indice DJA per il mercato finanziario, comprendiamo bene che più è corretta e tempestiva la previsione di tale indice, più valore acquisisce la previsione stessa.

Ora, l'indice DJA (come tutti gli altri indicatori di borsa) dipende dalle variabili macroeconomiche e la tempestività della previsione dell'indice dipende anche dal capire il legame che unisce l'indice con le variabili macro.

In questo lavoro utilizzeremo il coefficiente di correlazione per evidenziare il legame che esiste tra l'indice e le variabili macro. Abbiamo optato per il metodo più semplice perché in un secondo lavoro vorremmo dimostrare che per la legge dell'efficienza, a parità di condizioni, utilizzare un indice semplice che rileva il legame lineare tra i dati, si dimostra più utile (in termini di rapporto costo/ beneficio) ai fini predittivi.

Gli *asset* utilizzati in tale analisi di correlazione come variabili sono:

- USD GBP;
- CPI USA (Consumer Price Index);
- BRENT CRUDE OIL;

[3] La *10-Fold Cross Validation Procedure* è un metodo statistico-computazionale utilizzato per verificare la bontà dei risultati ottenuti. Si suddivide il dataset in 10 gruppi ugualmente numerosi e di volta in volta si esclude, se esiste, quello che fornisce risultati incoerenti con quelli degli altri gruppi.

- LOCKHEED;
- USD EUR;
- USD CNY.

Delle variabili su riportate, solo tre hanno mostrato una notevole correlazione con il nostro indice, ossia:

a) Lockheed ($\rho = 0{,}89$)
b) CPI USA ($\rho = 0{,}71$)
c) Brent Crude Oil ($\rho = 0{,}66$)

3.2 Reti Neurali Artificiali e Random Forest

Le Reti Neurali artificiali sono la modellazione algoritmica dei sistemi neurali biologici. Le abilità del cervello umano per eseguire simultaneamente compiti complessi, per imparare, memorizzare e generalizzare, hanno ispirato gli scienziati informatici a sviluppare algoritmi informatici basati sugli stessi principi del funzionamento del cervello umano. Le reti neurali artificiali sono i risultati di tali sforzi.

Le Reti Neurali Artificiali (ANN, *Artificial Neural Network*) sono reti stratificate di neuroni artificiali (AN). Ogni ANN riceve segnali da un altro AN o dall'ambiente, li raccoglie e forma un segnale di uscita che viene trasmesso ad un altro AN o all'ambiente. Un ANN è costituito da uno strato di input, uno o più livelli nascosti e uno strato di output di ANN. Ogni AN in uno strato è collegato, in tutto o in parte, ai ANN nel livello successivo. In alcune configurazioni ANN vengono introdotti collegamenti di feedback con gli strati precedenti.

Un neurone artificiale riceve un insieme di segnali di ingresso (x_1, x_2, ..., x_n) dall'ambiente o da un altro ANN. Un peso w_i ($i = 1, ..., n$) è associato a ciascun segnale di ingresso: se il peso è positivo allora il segnale è attivato, altrimenti il segnale è inibito. ANN raccoglie tutti i segnali di ingresso, calcola un segnale netto e trasmette un segnale di uscita; il segnale netto può essere calcolato come somma dei segnali di ingresso.

Il segnale in uscita viene calcolato utilizzando una funzione denominata funzione di attivazione. Sono possibili diversi tipi di funzioni di attivazione: funzione lineare, funzione a gradino, sigmoide, tangente iperbolica e così via. Mentre in linea di principio una rete neurale con n neuroni può avere n^2 connessioni direzionali, la complessità può essere ridotta organizzando i neuroni in strati e solo connessioni dirette da un dato livello al livello seguente. Questo tipo di rete neurale è chiamato Multi Layer Perceptron (MPL).

In questo articolo sono state utilizzate solo reti MLP poiché ottimale per gli obiettivi prefissati. La rete MLP è un meccanismo *feedforward* che propaga sem-

plicemente il segnale di ingresso attraverso tutti i livelli. L'algoritmo che ottimizza le prestazioni di classificazione di una rete neurale tramite la regolazione dei pesi, è definito *back propagation*.

Per quel che concerne l'altro metodo di *machine learning* utilizzato, ossia la *Random Forest*, esso è un *classificatore d'insieme* composto da molti alberi di decisione, il cui nome fu proposto per primo da Tin Kam Ho (1995); l'algoritmo attualmente utilizzato, tuttavia, è stato definito da Breiman (2001). Possiamo sintetizzarne il funzionamento sottolineando che combina l'idea della suddivisione-aggregazione dei Classification&Regression Trees (Breiman *et al.*, 1984) con la selezione casuale delle caratteristiche, introdotta prima da Ho e poi, indipendentemente, da Amit e Geman (1997) per costruire una raccolta di alberi di decisione con varianza controllata, fornendo in uscita il risultato che corrisponde al più frequente dei risultati degli alberi individuali.

La Random Forest è utilizzata perché è un metodo robusto rispetto a valori anomali o mancanti nei dati; inoltre, per sua costituzione, non incorre nel rischio di *overfitting*. Essa basa il proprio apprendimento sia su un training set (come il CRT), e sia su un vettore casuale di output reali. Ogni albero utilizza un diverso campione *bootstrap*[4] del dataset originale, e i dati non selezionati nell'albero vengono utilizzati per la stima degli errori di classificazione: in questo modo la Random Forest garantisce in automatico una valutazione dell'errore.

4. Procedura sperimentale e selezione delle variabili

Per migliorare la previsione ARIMA abbiamo implementato una procedura multivariata. Prima di tutto per aumentare la quantità di informazioni e di conseguenza migliorare il modello autoregressivo, abbiamo scelto le tre variabili macroeconomiche (ossia le serie di dati degli asset Lockheed, CPI USA e Brent Crude Oil) maggiormente correlate con l'indice Dow Jones e che dunque dovrebbero esserne, verosimilmente, alcune delle variabili antecedenti.

Per tener conto della stagionalità del Dow Jones abbiamo considerato per ogni mese le seguenti caratteristiche:

$$T_1 = \cos\left(\frac{2\pi \cdot M}{12}\right), \qquad T_2 = \sin\left(\frac{2\pi \cdot M}{12}\right).$$

All'interno delle formule, i mesi sono rappresentati dalla costante M, codificata in ordine cronologico (es. Gennaio = 1, Febbraio= 2,…, Dicembre = 12).

[4] Il metodo bootstrap consiste nell'estrarre numerosi campioni con sostituzione dall'insieme rilevato.

A questo punto, utilizzando le tre variabili su citate implementiamo le procedure multivariate con due diversi algoritmi:

- Reti Neurali MPL
- Random Forest

Una volta generate le analisi e le previsioni, i due metodi verranno comparati.

4.1 Rete Neurale MPL

Il meccanismo di *forecasting* di una Rete Neurale Artificiale MLP (*Multi-Layer Perceptron*) prende spunto dal meccanismo del cervello umano di apprendere dai propri errori: quindi la rete neurale cerca di fornire previsioni su nuovi dati in base all'esperienza che acquisisce tramite l'apprendimento sui dati e sui risultati già esistenti. Vi è dunque una prima fase di *memorizzazione delle regole* che si utilizzano per giungere ad un risultato.

Una rete neurale MLP, ai fini predittivi, adotta una forma di apprendimento supervisionato. Essa, infatti, considera in un primo momento un *training set*, ossia un insieme di dati di esempio con input e output ben definiti, i quali vengono posti in relazione tra loro tramite uno o più insiemi di nodi (che vengono detti "strati neuronali", in quanto i suoi componenti si comportano come centri di formulazione e trasmissione delle decisioni similmente ai neuroni del cervello); gli input vengono combinati dai neuroni con "pesi sinaptici" inizialmente casuali, ma che poi vengono definiti tramite dei cicli di addestramento della rete, in ognuno dei quali l'output (o gli output) risultante da tale combinazione viene confrontato con l'output reale, e le differenze tra i due vengono via via corretti, fin quando sia possibile (in base alla completezza e qualità dei dati) aggiustando i pesi sinaptici, in modo tale che alla fine dell'allenamento del MLP l'output della rete si avvicini il più possibile al– l'output reale.

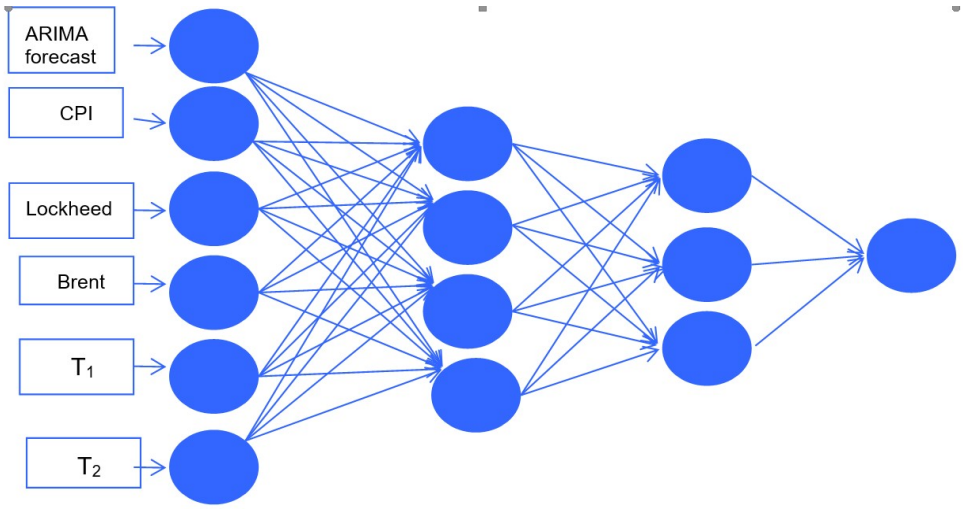
La seconda fase è quella della verifica e generalizzazione dei risultati tramite nuovi insiemi di dati (*test set*); in seguito si utilizza il modello ottenuto per analizzare nuovi input e fare una previsione di output.

Per il modello di Rete Neurale MPL preso in considerazione nel presente lavoro, il training set è stato costruito sui dati dell'arco temporale 2003–2014. Naturalmente i dati sono stati normalizzati prima di avviare il *forecasting*.

Come architettura di rete abbiamo optato per due strati nascosti poiché risulta l'architettura più efficace per casi di previsione quando si lavora con dati ad alta frequenza, minimizzando l'errore nel training set. In particolare, abbiamo identificato 6 neuroni nello strato di input (corrispondenti all'output del forecasting ARI-

MA, alle tre variabili macroeconomiche prescelte e alle due variabili trigonometriche di stima della stagionalità), rispettivamente 4 e 3 neuroni negli strati nascosti (*hidden layers*) ed infine un neurone per lo strato di output.

Figura 2. *Architettura MPL di previsione.*



Per quel che riguarda il numero dei neuroni utilizzati negli strati nascosti, vige il criterio della minimizzazione del rischio di *overfitting* che si corre quando, per il numero eccessivo dei neuroni, la rete si adatta troppo ai dati di training set e non è più in grado di generalizzare i risultati.

La funzione di attivazione utilizzata è la funzione sigmoidale, che fornisce valori sempre compresi fra 0 e 1, garantendo stime adeguate anche con relazioni non lineari:

$$\gamma(x) = \frac{1}{1+e^{-x}} \cdot$$

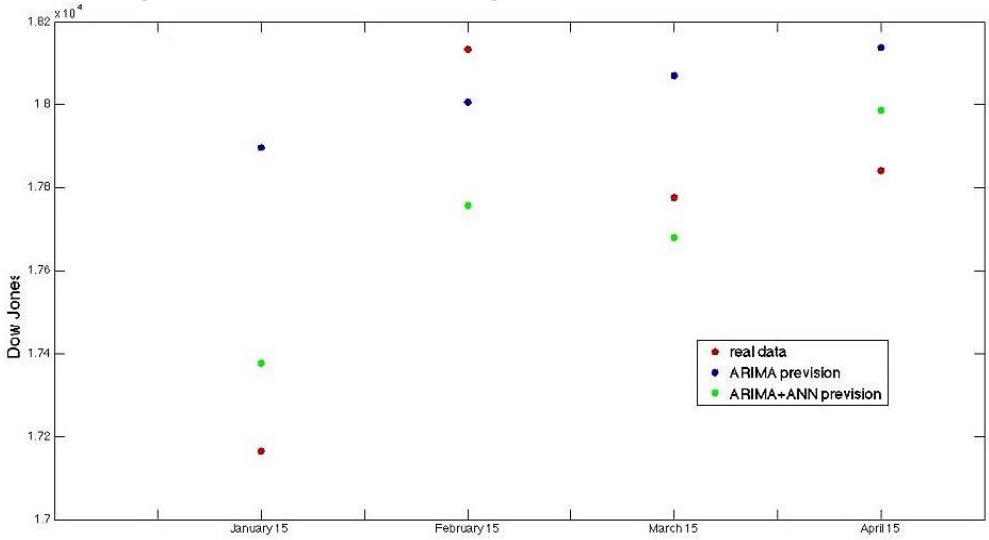
L'indicatore di bontà predittiva è il *Mean Absolute Percentage Error* (MAPE), ossia la media aritmetica degli errori relativi, in valore assoluto espresso in percentuale.

Si noti che con la procedura MPL utilizzata la previsione è stata più performante, evidenziando un MAPE ridotto del 50% rispetto al metodo ARIMA:

$$\text{MAPE (ARIMA+MPL)} = 1{,}4\%$$

Nella Figura 3 viene rappresentata la previsione sia col modello ARIMA che con il modello MPL.

Figura 3. *Previsioni Dow Jones Average Index 2015 con ARIMA e ANN-MLP*

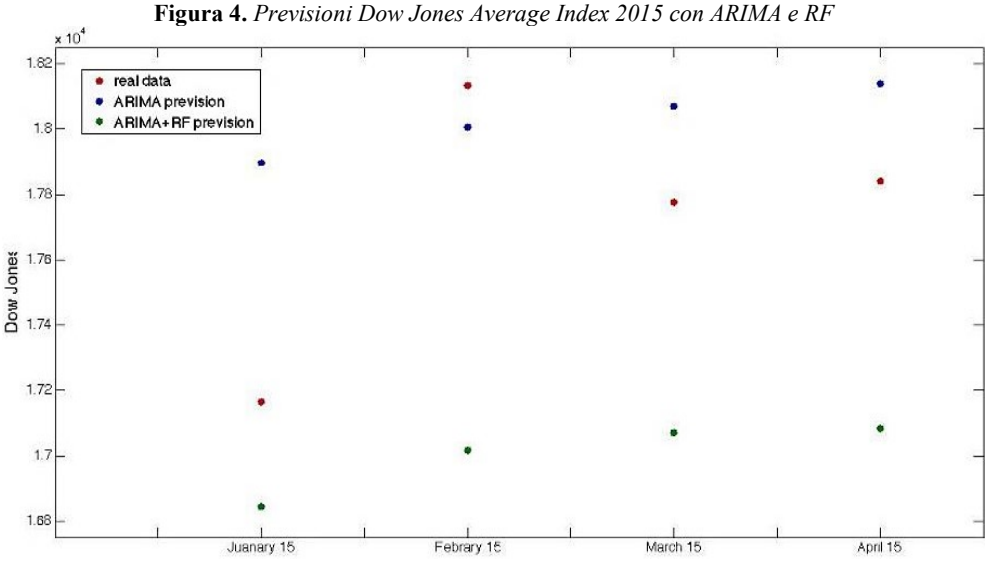


4.2 Random Forest

Come già affermato, per migliorare le previsioni ARIMA abbiamo preso in considerazione un altro metodo multivariato: la Random Forest. I criteri utilizzati per costruire l'algoritmo predittivo con 500 alberi sono identici ai criteri di costruzione delle Reti Neurali e inoltre abbiamo utilizzato le stesse variabili di input.

La procedura Random Forest utilizzata per la previsione è stata leggermente più performante rispetto al modello ARIMA, evidenziando un MAPE del 2,5%.

Figura 4. *Previsioni Dow Jones Average Index 2015 con ARIMA e RF*



5. Conclusioni

I risultati ottenuti dal lavoro svolto hanno fatto luce sulla importanza che riveste la previsione con metodi statistici arricchita delle funzionalità dei metodi di *machine learning*, non solo dal punto di vista dello schema predittivo ma soprattutto dalla adattabilità e specificità predittiva alla stessa variabile analizzata.

Utilizzando i dati mensili per il modello ARIMA, abbiamo ottenuto delle previsione che complessivamente hanno un errore (MAPE) del 2,8%.

La combinazione di metodi predittivi statistici e di *machine learning* , e precisamente di una Rete Neurale MLP, ha portato un innegabile miglioramento predittivo, abbattendo del 50% l'errore di previsione dovuto al metodo ARIMA. I risultati ottenuti con le Random Forest, invece, non hanno apportato miglioramenti predittivi rispetto alla previsione ARIMA.

In ogni caso, un errore predittivo pari all'1,4% sull'andamento futuro di un indice così rilevante quale il Dow Jones, è un rilevante traguardo che ci incoraggia a sperimentare nuove metodologie confinanti tra la metodologia statistica e i metodi avanzati di *machine learning*.

Riferimenti bibliografici

Amit, Y.; Geman, D. (1997). Shape quantization and recognition with randomized trees, *Neural Computation*. 9 (7): 1545–1588. doi: 10.1162/neco.1997.9.7.1545.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*, Springer-Verlag New York.

Box, G. E. P.; Jenkins, G. M. (1970). *Time Series Analysis: Forecasting and Control*, San Francisco: Holden-Day

Box, G. E. P.; Jenkins, G. M.; Reinsel, G. C.; Ljung, G. M. (2015). *Time Series Analysis: Forecasting and Control*, John Wiley & Sons.

Breiman, L.; Friedman, J.H.; Olshen, R.A.; Stone, C.J. (1984). *Classification and Regression Trees*, Chapman & Hall, New York-London.

Breiman, L (2001). Random Forests. *Machine Learning*, 45 (1): 5–32. doi: 10.1023/A:1010933404324.

Du, K.L.; Swamy, M. N. S. (2014). *Neural Networks and Statistical Learning*, Springer-Verlag New York.

Ho, T. K. (1995) Random Decision Forests, *Proceedings of the 3rd International Conference on Document Analysis and Recognition*, Montreal, QC, 14–16 August: pp. 278–282.

UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO *Metodi e analisi statistiche 2017*
DIPARTIMENTO DI ECONOMIA E FINANZA
ISBN: 978-88-6629-013-1
pp. 263-282

Una proposta per la stima della misura della disuguaglianza nella distribuzione della ricchezza nel Tardo Medioevo in Terra di Bari

Vito Ricci*

Università degli studi di Bari Aldo Moro - UO Statistiche di Ateneo

Riassunto: In questo contributo partendo dall'esame delle donazioni effettuate a favore delle cattedrali di Ruvo di Puglia e di Giovinazzo in due necrologi del Trecento, dopo aver effettuato una serie di analisi quali-quantitative propedeutiche, si propone di utilizzare i lasciti in denaro come una *proxy* della ricchezza al fine di ottenere una misura di disugaglianza nella distribuzione di tale grandezza che per il Tardo Medioevo nel Regno di Napoli, al contrario di altri contesti geografici (Toscana, Piemonto), risultano mancanti.

Keywords: Distribuzione ricchezza; disuguaglianza; funzione di Pareto; indice di Gini

1. Introduzione

Gli obituari o necrologi erano dei libri o registri nei quali venivano annotati i nomi dei defunti con il giorno della settimana e la data di morte (*obitus*) e le donazioni *pro anima*, in natura o in denaro, che venivano effettuate a favore di chiese, monasteri, confraternite e ospedali al fine di ricevere suffragi da parte delle comunità che facevano parte di tali enti. Per tale scopo nell'obituario si teneva conto solo dell'anniversario della morte e raramente veniva riportato l'anno della morte del defunto.

In questa sede si vuole esaminare l'uso che degli obituari può essere fatto come fonti per la storia economica di un determinato contesto territoriale attraverso l'analisi delle donazioni effettuate a favore degli enti religiosi. Nella fattispecie si tratta di due obituari della prima metà del XIV secolo provenienti da due centri del-

* Autore corrispondente: vito.ricci@uniba.it.

la Terra di Bari, entrambi sedi vescovili attestate dal XI secolo: Giovinazzo, città portuale sull'Adriatico, e Ruvo di Puglia, città dell'interno attraversata dalla via Traiana, arteria ancora importante nel corso del Medioevo. Il *Quaternus de fraternitate communitatis nostri episcopi* è un codice proveniente dalla Cattedrale di Giovinazzo (Garufi, 1911), mentre la *Matricula maioris ecclesie Rubensis* è un codice pergamenaceo facente parte dell'Archivio Capitolare di Ruvo di Puglia (Ficco, 2005). Dalle trascrizioni di questi due necrologi si sono costruiti dei database con le variabili necessarie (nome, forma cognominale, titolo o professione, genere del defunto, importo in denaro donato, descrizione dei beni mobili o immobili donati, località di ubicazione dei beni immobili) per poter esaminare nel dettaglio le donazioni *pro anima*. Partendo da tali database sono state introdotte delle riclassificazioni del titolo o professione del defunto, che abbiamo chiamato qualifica socio-economica, e dei beni donati, che abbiano definito tipologia del lascito, in modo da ridurre l'estrema varietà desumibile dai valori originali.

Nel secondo paragrafo si esaminerà la composizione dei benefattori secondo alcune caratteristiche socio-economiche e demografiche, nel terzo si procederà all'analisi delle donazioni, il quarto avrà per oggetto l'analisi delle elargizioni in denaro, mentre il quinto sarà dedicato alla concentrazione delle donazioni in moneta e sarà proposto l'utilizzo di tali dati per una stima della disuguaglianza nella distribuzione della ricchezza.

2. I benefattori: caratteristiche demografiche e socio-economiche

Entrambi gli obituari contengono un numero abbastanza elevato di osservazioni, nella *Matricula* di Ruvo si possono contare 654 benefattori, mentre nel *Quaternus* di Giovinazzo il loro numero ammonta a 705. Nelle Tabb. 1 e 2 si riportano le distribuzioni secondo il genere e la qualifica socio-economica.

Occorre precisare che nell'obituario di Ruvo sono presenti 15 donazioni che sono state effettuate da un nucleo familiare e che pertanto sono state escluse dall'analisi secondo il genere, mentre in quello di Giovinazzo per due benefattori non è stato possibili risalire al genere, e anche in questo caso sono stati esclusi da tale analisi. Purtroppo negli obituari mancano completamente i dati relativi all'età del defunto che sarebbe potuta essere un'altra variabile di estremo interesse.

Per quanto riguarda la composizione secondo il genere (Tab. 1), si nota una leggera prevalenza maschile: 55,9% a Ruvo e 54,2% a Giovinazzo, con incidenze percentuali che risultano essere molto simili tra i due centri.

Tabella 1. *Distribuzione dei benefattori secondo il genere*

	Ruvo		Giovinazzo	
Genere	N.	*%*	N.	*%*
M	357	*55,9*	381	*54,2*
F	282	*44,1*	322	*45,8*
Totale	639	*100,0*	703	*100,0*

Tabella 2. *Distribuzione dei benefattori secondo la qualifica*

	Ruvo			Giovinazzo		
Qualifica	N.	%	% senza mancanti	N.	%	% senza mancanti
Artigiano	35	*5,5*	*21,1*	24	*3,4*	*15,5*
Chierico	80	*12,5*	*48,2*	48	*6,8*	*31,0*
Giudice	14	*2,2*	*8,4*	20	*2,8*	*12,9*
Nobile	28	*4,4*	*16,9*	43	*6,1*	*27,7*
Notaio	9	*1,4*	*5,4*	5	*0,7*	*3,2*
Professioni del mare	0	*0,0*	*0,0*	7	*1,0*	*4,5*
Altro	0	*0,0*	*0,0*	8	*1,1*	*5,2*
Non indicata	473	*74,0*		548	*78,0*	
Totale	639	*100,0*	*100,0*	703	*100,0*	*100,0*

Dalla Tab. 2 si osserva subito come in entrambe gli obituari per circa i tre quarti dei benefattori non erano disponibili informazioni di natura socio-economica tali da poter configurare una qualche qualifica. A ragione di ciò si è ritenuto più corretto calcolare le incidenze percentuali anche sui soli valori validi: in entrambi gli obituari risultano prevalere i chierici (48,2% di Ruvo vs 31% di Giovinazzo), dato abbastanza prevedibile in quanto il clero era parte integrante delle istituzioni destinatarie delle donazioni; seguono a Ruvo gli artigiani (21,1%), mentre a Giovinazzo i nobili (27,7%).

Nel complesso le due distribuzioni secondo la qualifica risultano essere lievemente difformi, calcolando l'indice semplice di dissomiglianza (Piccolo, 1998):

$$D = \frac{1}{2}\sum_{i=1}^{k}|f_{1i} - f_{2i}|$$

dove le f_i sono le frequenze relative delle k modalità, si ottiene un valore di 0,25, molto prossimo al minimo che può assumere questo indice che varia tra 0 (quando le due distribuzioni di frequenza relative coincidono perfettamente) e 1 (quando si verifica la dissomiglianza massima, ovvero quando la distribuzione è concentrata in una sola modalità differente in entrambe). Un altro aspetto interessante delle due

distribuzioni che può essere esaminato è quello relativo all'eterogeneità, che può essere misurata tramite l'indice relativo di eterogeneità di Gini (Piccolo, 1998):

$$E = \frac{k}{k-1}\sum_{i=1}^{k} 1 - f_i^2$$

dove k è il numero delle modalità del carattere qualitativo e f_i sono le frequenze relative; esso ha un campo di variazione compreso tra 0, nel caso di massimo omogeneità ovvero quando le frequenze si concentrano su di un'unica modalità, e 1, quando le frequenze sono distribuite uniformemente tra le *k* modalità avendosi in tale circostanza la massima eterogeneità. Se si calcola l'indice E per la distribuzione dei benefattori secondo la qualifica socio-economica di Ruvo si ottiene un valore di 0,783, mentre per quella di Giovinazzo 0,892; si evince come si abbia una maggiore eterogeneità nella composizione dei benefattori a Giovinazzo rispetto a Ruvo, dovuta anche alla presenza di alcune qualifiche non rilevate nel secondo centro.

3. Analisi delle donazioni secondo la tipologia

L'oggetto delle elargizioni da parte dei defunti poteva essere estremamente vario e per questo motivo si è proceduto a classificarle in un numero abbastanza limitato e significativo di categorie tale da consentire delle analisi di tipo qualitativo. Le donazioni potevano riguardare beni fondiari (appezzamenti di terra e case), suppellettili e arredi liturgici (tovaglie, camici, calici, etc.), denaro, oppure altre categorie marginali che sono state raggruppate nella voce Altro. Nella Tab. 3 si riporta la distribuzione delle donazioni secondo la tipologia e il genere per l'obituario di Ruvo; oltre ai valori assoluti, sono stati calcolati anche i profili di colonna e quelli di riga.

Tabella 3. *Distribuzione dei lasciti secondo la tipologia e il genere (Ruvo)*

Tipologia	Valori rilevati			Profili colonna			Profili riga		
	Genere			Genere			Genere		
	M	F	Totale	M	F	Totale	M	F	Totale
Casa	28	6	34	*7,8*	*2,1*	*5,3*	*82,4*	*17,6*	*100,0*
Chiesa	11		11	*3,1*	*0,0*	*1,7*	*100,0*	*0,0*	*100,0*
Denaro	167	225	392	*46,8*	*79,8*	*61,3*	*42,6*	*57,4*	*100,0*
Suppellettili	6	5	11	*1,7*	*1,8*	*1,7*	*54,5*	*45,5*	*100,0*
Terreno	128	40	168	*35,9*	*14,2*	*26,3*	*76,2*	*23,8*	*100,0*
Altro	17	6	23	*4,8*	*2,1*	*3,6*	*73,9*	*26,1*	*100,0*
Totale	357	282	639	*100,0*	*100,0*	*100,0*	*55,9*	*44,1*	*100,0*

Dalla Tab. 3 emerge come la principale di tipologia di donazione è quella in moneta con il 61,3%, seguita a molta distanza dai terreni (26,3%); piuttosto esigue sono le donazioni di suppellettili o di chiese (1,7% ciascuna). Esaminando i profili di colonna per genere si nota che sia per i maschi che per le femmine la modalità con la maggiore frequenza è Denaro, sebbene con percentuali diverse: 46,8% per gli uomini e 79,8% per le donne; la minore percentuale delle donazioni in moneta per gli uomini è giustificata dalla maggiore incidenza delle donazioni di beni fondiari (terreni e case) rispetto alle donne: nel periodo in esame era il genere maschile a detenere la maggior parte della ricchezza immobiliare. Se si calcola l'indice semplice di dissomiglianza tra le due distribuzioni secondo il genere si ottiene un valore pari a 0,331 che mostra una diversità del comportamento nelle elargizioni da parte dei due generi, sebbene non molto accentuata. La maggiore propensione femminile per le donazioni in contanti emerge anche dai profili di riga: in tutte le tipologie si ha una prevalenza maschile, mentre per questa modalità il 57,4% delle donazioni in denaro sono state effettuate da donne. Nella Tab. 4 si riportano le donazioni secondo la tipologia e il genere per il necrologio di Giovinazzo.

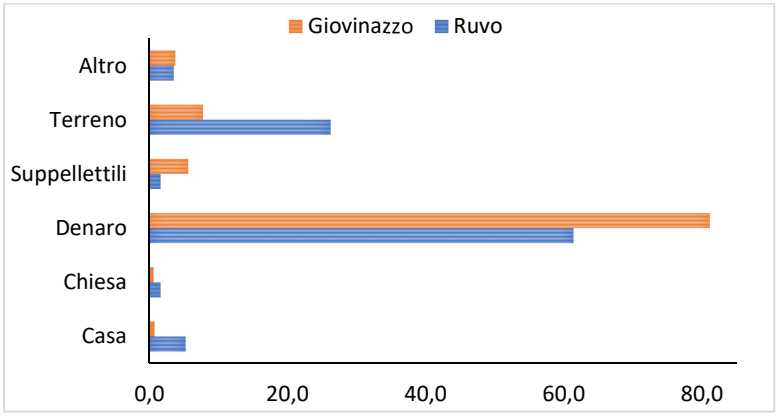
Tabella 4. *Distribuzione dei lasciti secondo la tipologia e il genere (Giovinazzo)*

Tipologia	Valori assoluti			Profili colonna			Profili riga		
	Genere			Genere			Genere		
	M	F	Totale	M	F	Totale	M	F	Totale
Casa	4	2	6	*1,0*	*0,6*	*0,9*	*66,7*	*33,3*	*100,0*
Chiesa	4	1	5	*1,0*	*0,3*	*0,7*	*80,0*	*20,0*	*100,0*
Denaro	300	270	570	*78,7*	*83,9*	*81,1*	*52,6*	*47,4*	*100,0*
Suppellettili	20	20	40	*5,2*	*6,2*	*5,7*	*50,0*	*50,0*	*100,0*
Terreno	33	22	55	*8,7*	*6,8*	*7,8*	*60,0*	*40,0*	*100,0*
Altro	20	7	27	*5,2*	*2,2*	*3,8*	*74,1*	*25,9*	*100,0*
Totale	381	322	703	*100,0*	*100,0*	*100,0*	*54,2*	*45,8*	*100,0*

Dal *Quaternus* emerge come oltre l'80% delle donazioni è stata effettuata in moneta, mentre l'incidenza dei beni fondiari, soprattutto delle case, appare molto ridimensionata rispetto alla *Matricula* di Ruvo. Non vi sono differenze significative tra i due generi, essendo i profili di colonna quasi identici con l'indice D=0,061. Se si considerando i profili di riga si nota una prevalenza del genere maschile per quasi tutte le modalità, solo per le donazioni di suppellettili e arredi sacri si registra equi-distribuzione tra i due generi. Effettuando un confronto tra i dati di Ruvo e quelli di Giovinazzo (Fig. 1) si nota subito come in entrambe gli obituari la maggior parte delle donazioni effettuate dai benefattori è in denaro, sebbene con inci-

denze percentuali diverse (61,3% a Ruvo e 81,1% a Giovinazzo). La maggiore incidenza delle donazioni in contante a Giovinazzo potrebbe trovare una spiegazione nell'economia commerciale di questo centro che, essendo posto sulla costa e possedendo un porto, vedeva la presenza di un ceto di mercanti con patrimoni mobiliari.

Figura 1. *Composizione percentuale dei lasciti secondo la tipologia*



Si può osservare come a Ruvo, centro ad economia prevalentemente legata all'agricoltura, vi fosse una maggiore propensione a donare beni immobili 33,3% (di cui il 26,3% terreni), contro appena il 9,4% di Giovinazzo, sintomo questo di una maggiore disponibilità da parte dei ruvesi di legare alla cattedrale una parte importante del proprio patrimonio fornendole dei beni che le garantivano una rendita costante nel tempo; mentre a Giovinazzo si riscontra un'incidenza più elevata rispetto a Ruvo (5,7% contro 1,7%) della donazione di suppellettili religiose, mostrando una maggiore attenzione da parte dei benefattori giovinazzesi per le immediate esigenze liturgiche della cattedrale. L'indice D tra le distribuzioni dei due centri è pari a 0,24, evidenziando una moderata dissomiglianza.

Può essere sicuramente interessante studiare l'eterogeneità nella distribuzione dei lasciti secondo la tipologia. Nella Tab. 5 è stato calcolato l'indice E per le distribuzioni di Ruvo e Giovinazzo distinguendo per genere dei benefattori.

Si può subito osservare come nell'obituario di Ruvo si riscontra una maggiore eterogeneità rispetto a quello di Giovinazzo, imputabile al fatto che in quest'ultimo la maggior parte delle frequenze sono addensante attorno ad una sola modalità, mentre a Ruvo sono distribuite più uniformemente, inoltre nel primo si assiste anche ad una divergenza anche all'interno dei generi, con la maggiore eterogeneità in quello maschile; nel secondo, sebbene si nota una maggiore mutabilità tra il genere maschile, questa non è di molto superiore a quella della distribuzione femminile.

Tabella 5. *Indice relativo di eterogeneità di Gini della distribuzione dei lasciti secondo la tipologia, per genere dei benefattori*

Genere dei benefattori	Ruvo	Giovinazzo
M	0,772	0,440
F	0,410	0,345
Totale	0,660	0,398

4. Analisi delle donazioni in denaro

Nei paragrafi precedenti si è visto come le donazioni in denaro siano quelle maggioritarie sia nel necrologio di Ruvo che in quello di Giovinazzo; trattandosi di una variabile di tipo quantitativo si presta meglio ad una più dettagliata analisi statistica. Tratteremo di tre aspetti: il calcolo di alcuni indici di posizione e variabilità che descrivono la distribuzione delle donazioni, una sommaria analisi inferenziale e la rappresentazione analitica con la stima dei parametri di un modello teorico di tipo *power law*. Gli importi delle donazioni registrate negli obituari sono espresse in once, tarì e grana (1 oncia=30 tarì=600 grana) a Ruvo, mentre a Giovinazzo sono presenti più monete: augustali, moneta emessa in epoca federiciana ma con corso anche successivo (1 augustale=7,5 tarì) e fiorini, moneta molto utilizzata negli scambi commerciali nel Tardo Medioevo (1 fiorino=6 tarì), oltre che ad once, tarì e grana. Per facilitare i calcoli ed effettuare delle comparazioni tutti gli importi sono stati trasformati in grana, tenendo conto dei rapporti di equivalenza.

4.1 Analisi esplorativa

Nelle Tabb. 6 e 7 si riportano le distribuzioni di frequenza delle donazioni in denaro per Ruvo e Giovinazzo; per le prima la classe modale è 75-100 tarì con l'85,6%, mentre per la seconda è 101-125 tarì con il 26,3%.

Tabella 6. *Distribuzione di frequenza delle donazioni in denaro(grana) per Ruvo*

Classi di donazione	N.	*%*
20-74	11	*2,7*
75-100	346	*85,6*
101-125	5	*1,2*
126-250	27	*6,7*
oltre 250	15	*3,7*
Totale	404	*100,0*

Tabella 7. *Distribuzione di frequenza delle donazioni in denaro(grana) per Giovinazzo*

Classi di donazione	N.	*%*
20-74	83	*14,5*
75-100	51	*8,9*
101-125	151	*26,3*
126-250	141	*24,6*
251-300	86	*15,0*
oltre 300	62	*10,8*
Totale	574	*100,0*

Emerge subito come la distruzione di Ruvo risulti molto più concentrata attorno alla classe modale rispetto a quella di Giovinazzo, caratterizzata quest'ultima da una certa incidenza percentuale dei valori delle classi estreme: la frequenza relativa delle donazioni superiori a 250 tarì è pari al 3,7% a Ruvo e al 25,8% a Giovinazzo.

Nella Tab. 8 sono presentati i principali indicatori descrittivi delle due distribuzioni.

Tabella 8. *Principali indicatori descrittivi dell'importo delle donazioni in grana per città e genere*

Collettivo	N	Media	Dev. St.	Min	Max	Range
Ruvo						
Tutti	404	132,11	616,45	50	12.000	11.950
M	177	186,66	913,09	50	12.000	11.950
F	227	89,58	155,55	60	2.400	2.340
Giovinazzo						
Tutti	574	384,94	753,48	20	6.000	5.980
M	302	428,99	851,39	30	6.000	5.970
F	271	336,90	626,07	20	2.400	2.380

Si nota immediatamente la divergenza dell'importo medio tra le donazioni di Ruvo e quelle di Giovinazzo, imputabile probabilmente, per lo meno in parte, anche alla diversa economia dei due centri (più rurale a Ruvo e più commerciale a Giovinazzo). All'interno dei due collettivi esistono differenze secondo il genere, tra i donatori di genere maschile si osserva il valore medio più elevato rispetto al genere femminile, così come anche il valore massimo e anche una maggiore variabilità misurata dalla deviazione standard. Nel periodo storico in esame gli uomini avevano una capacità reddituale notevolmente maggiore delle donne: questo particolare può spiegare le differenze degli importi per genere.

Nella Tab. 9 abbiamo calcolato gli stessi indicatori distinguendo in base alla categoria socio-economica. Occorre precisare come alcune qualifiche presentano un numero molto esiguo di osservazioni e ciò potrebbe influire sul calcolo degli indicatori. In entrambi gli obituari si riscontrano importi medi e quelli massimi particolarmente elevati per i chierici (a Ruvo sono i primi in assoluto, mentre a Giovinazzo sono secondi alle spalle delle professioni marittime); il dato è spiegabile sia per la maggior propensione del clero nell'elargire dei lasciti alle istituzioni religiose di cui gli stessi erano parte e sia per il fatto che, non avendo una discendenza diretta, potevano disporre dei propri beni senza il vincolo dell'eredità ai figli, circostanza questa che in qualche modo potrebbe aver invece influito sulle scelte degli altri ceti

sociali. Il dato medio delle professioni del mare, sebbene più elevato di quello dei chierici, è da prendere con cautela dato l'esiguo numero di osservazioni.

I ceti più altolocati (clero, nobili e professioni marittime - a Giovinazzo-) presentano i valori medi più alti rispetto alle altre qualifiche; tale caratteristica fa propendere per l'ipotesi che l'importo delle donazioni era funzione del livello di ricchezza del benefattore, essendo la ricchezza dei ceti più benestanti maggiore. Sicuramente altre variabili esplicative potevano influire sulla determinazione dell'importo delle elargizioni in denaro come ad esempio l'età, lo stato civile, il numero dei figli etc., dati purtroppo assenti negli obituari.

Tabella 9. *Principali indicatori descrittivi dell'importo delle donazioni in grana per città e qualifica socio-economica*

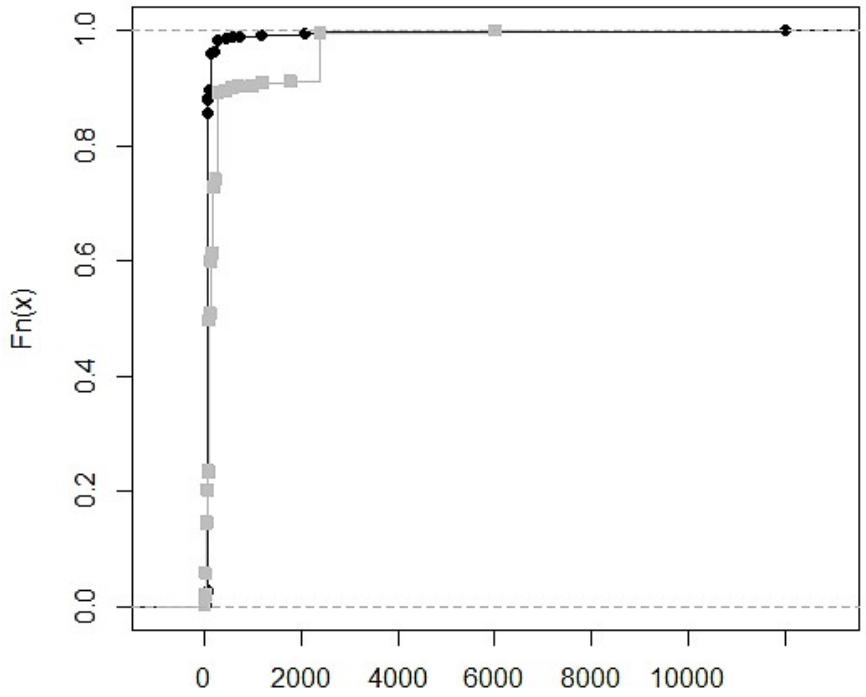
Collettivo	N	Media	Dev. St.	Min	Max	Range
Ruvo						
Artigiano	21	113,81	122,58	60	600	540
Chierico	23	805,22	2.486,68	75	12.000	11.925
Giudice	10	97,50	36,23	75	150	75
Nobile	10	142,50	89,79	75	300	225
Notaio	5	105,00	41,08	75	150	75
Non indicata	335	88,17	129,73	50	2.400	2.350
Giovinazzo						
Artigiano	19	377,37	715,39	30	2.400	2.370
Chierico	43	619,07	1.315,67	60	6.000	5.940
Giudice	12	201,67	81,00	120	300	180
Nobile	28	545,36	774,37	60	2.400	2.340
Notaio	5	150,00	90,00	60	300	240
Professioni mare	7	1.028,57	2.093,74	80	6.000	5.920
Altro	7	470,00	854,69	60	2.400	2.340
Non indicata	453	349,30	642,45	20	2.400	2.380

4.2 Inferenza

Dalle distribuzioni di frequenza (Tabb. 6 e 7) si nota abbastanza chiaramente come esse si discostino parecchio dal modello gaussiano, circostanza confermata anche dal test di normalità di Shapiro (Piccolo, 1998) risultato estremamente significativo per entrambi i centri, e lasciano propendere per una distribuzione di tipo *power law*, nella fattispecie in quella di Pareto; di tale aspetto ci occuperemo di seguito, in questo momento è importante sottolineare la non normalità dei dati che comporterà nell'analisi inferenziale l'utilizzo prevalente dei test non parametrici.

In primo luogo, effettuiamo il test di Kolmogorov-Smirnov (Piccolo, 1998) per verificare se i due campioni provengono o meno dalla medesima e ignota popolazione, ovvero se hanno la medesima funzione di ripartizione con identici valori dei parametri; si tratta di un test generalista, cioè permette di valutare la significatività complessiva dovuta a differenze di tendenza centrale, dispersione, simmetria e curtosi, e *distribution free*, in quanto prescinde dalla funzione di ripartizione della popolazione di appartenenza. Nella Fig. 2 si riportano le funzioni di ripartizione empiriche delle due distribuzioni delle donazioni in moneta, in nero quella relativa a Ruvo e in grigio quella di Giovinazzo. Il valore test di Kolmogorov-Smirnov è pari a 0,71 e risulta molto significativo; possiamo quindi affermare che le popolazioni di provenienza dei due campioni di osservazioni sono completamente diverse.

Figura 2. *Funzioni di ripartizione empiriche degli importi delle donazioni in denaro negli obituari di Ruvo (nero) e Giovinazzo (grigio*)



Per il confronto delle medie si è impiegato il test di Welch (data l'elevata numerosità del campione), mentre per la dispersione si è utilizzato quello di Levene; entrambi i test statistici sono risultati molto significativi e permettono di affermare che nel complesso le distribuzioni delle donazioni di Ruvo e Giovinazzo hanno ten–denza centrale e dispersione diverse.

Relativamente ai confronti di genere all'interno dello stesso ambito geografico abbiamo che per quanto riguarda Ruvo il test di Welch è risultato non significativo (*p-value*= 0,164), ovvero che le medie delle donazioni tra maschi e femmine non sono significativamente diverse, allo stesso modo per la dispersione, il test di Levene è non significativo (*p-value*=0,116). Per quanto concerne Giovinazzo, il quadro è simile: il test di Welch ha *p-value*=0,138 e quello di Levene *p-value*=0,192, entrambi quindi non significativi. Possiamo concludere che nell'ambito del medesimo contesto geografico al loro interno i sottogruppi di donatori maschi e femmine sono abbastanza omogenei sia per quanto concerne la tendenza centrale che per la dispersione.

È stato effettuato anche il test di Kruskal-Wallis (Soliani, 2010), l'alternativa non parametrica all'ANOVA, per verificare l'uguaglianza delle mediane trale diverse qualifiche socio-economiche; tale è risultato molto significativo sia per Ruvo che per Giovinazzo, avendosi differenze sensibili tra le diverse qualifiche all'interno della medesima distribuzione geografica. La variabile qualifica socio-economica ha una forte influenza nella determinazione dell'importo delle elargizioni in denaro da parte dei benefattori.

4.3 Rappresentazione analitica

Si è portati a ritenere che un modello di tipo *power law* possa rappresentare abbastanza bene analiticamente la distribuzione delle donazioni in moneta; in particolare si opta per la funzione di Pareto che è utilizzata sovente nella rappresentazione delle distribuzioni tronche di alcune grandezze economiche come redditi o ricchezze. La funzione di densità della frequenza della distribuzione di Pareto è la seguente:

$$f(x) = \alpha x_{min}^{\alpha} x^{-(\alpha+1)}$$

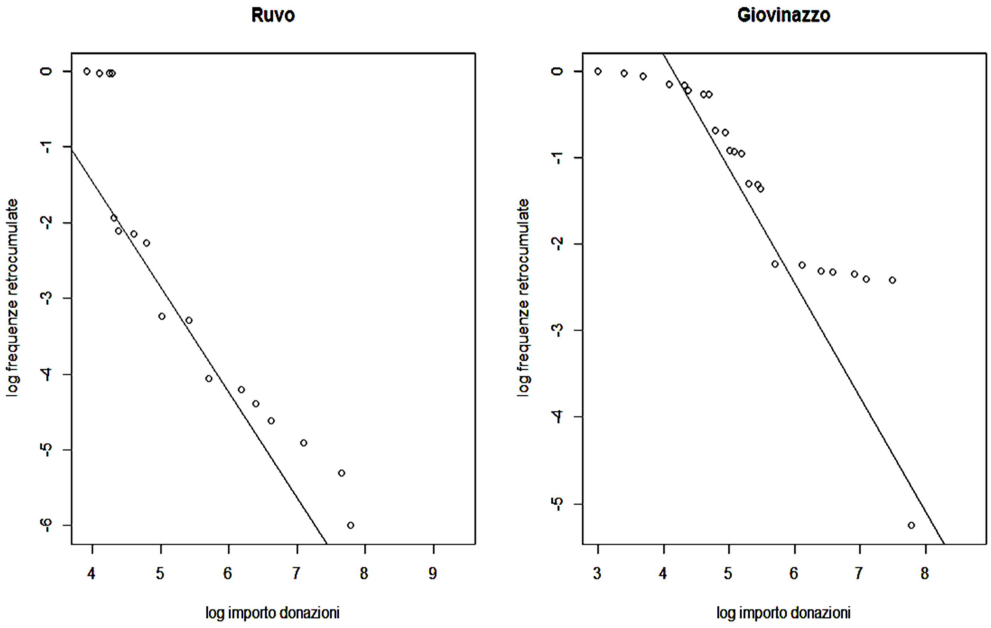
definita per $x>x_{min}$, mentre α è un parametro di scala sul quale avremo modo di tornare. Se consideriamo il complemento della funzione di ripartizione (ovvero le frequenze retrocumulate relative) abbiamo:

$$G(x) = 1 - F(x) = \left(\frac{x}{x_{min}}\right)^{-\alpha}$$

che in scala doppio logaritimica diventa lineare. Questa trasformazione permette di ottenere una stima dei parametri con il metodo dei minimi quadrati (Piccolo, 1998). L'approccio con la trasformazione logaritmica e il metodo dei minimi quadrati è quello utilizzato da Pareto nelle sue ricerche.

Dalla Fig. 3, rappresentazione grafica in scala doppio logaritmica dei dati e del–la retta stimata con i minimi quadrati, si osserva che il modello paretiano si adatta molto bene ai dati delle distribuzioni delle donazioni dei centri avendosi un indice di determinazione R^2, che può considerarsi come una misura di adattamento del modello teorico ai dati osservati, pari a 0,79 per Ruvo e 0,94 per Giovinazzo. In entrambi i casi si nota come gli importi più bassi delle donazioni sono quelli per i quali le frequenze retrocumulate si discostano maggiormente dal modello teorico; per Giovinazzo anche un parte dei valori più elevati si posiziona un po' distante dalla retta logaritmica.

Figura 3. *Importo donazioni vs frequenze relative retro-cumulate in scala log-log*



Il paramento α misura l'inclinazione della retta logaritmica ed è interpretato come un indice sommario dell'uguaglianza nella distribuzione dei valori; a maggiori valori di α corrisponde una distribuzione più egualitaria. Il suo aumento significa l'incremento dei valori più bassi e la diminuzione dei più elevati, invece la sua diminuzione ha significato opposto. I valori di α che abbiamo stimato per le due distribuzioni delle donazioni in denaro sono abbastanza simili al valore calcolato da Pareto per diverse distribuzioni del reddito in età moderna.

La stima dei parametri del modello paretiano tuttavia presenta alcune limitazioni abbastanza forti (Clauset et al., 2009) per tale morivo è preferibile utilizzare delle stime di massima verosimiglianza (Rytgaard, 1990; Quandt, 1966). Su tale me-

todo, come sugli altri esistenti in letteratura (Quandt, 1966), non ci sofferma essendo in questa sede lo scopo della rappresentazione analitica di natura meramente descrittiva.

Nella Tab. 10 si riportano le stime dei due parametri della distribuzione di Pareto con il metodo dei minimi quadrati per le donazioni in denaro di Ruvo e Giovinazzo. Il coefficiente angolare della retta in scala doppio-logaritmica α è abbastanza simile nei due centri, mentre risulta molto diversa l'intercetta dipendente dal valore minimo delle elargizioni in denaro.

Dal valore dell'indice di determinazione si comprende come per entrambi gli obituari il modello paretiano sia adatti molto bene nel rappresentare analiticamente i dati delle donazioni in denaro.

Tabella 10. *Stime dei parametri del modello paretiano con il metodo dei minimi quadrati*

Parametri	Ruvo	Giovinazzo
α	1,39	1,32
x_{min}	19,05	63,80
R^2	0,79	0,94

5. La concentrazione dei lasciti in denaro. Proposta per una stima per la misura della disuguaglianza nella distribuzione della ricchezza

Nei paragrafi precedenti sono stati esaminati i dati delle donazioni da un punto di vista qualitativo e quantitativo; tale analisi sono risultate propedeutiche per affrontare l'argomento della concentrazione delle donazioni in denaro attraverso il calcolo dell'indice di Gini (Piccolo, 1998), misura utilizzata per i caratteri trasferibili. Ordinando le donazioni in senso non decrescente, indichiamo con p_i la quota dei primi i donatori e con q_i la quota delle elargizioni in denaro donate dagli i benefattori abbiamo che l'indice di Gini ha la seguente formula:

$$R = \frac{\sum_{i=1}^{n-1}(p_i - q_i)}{\sum_{i=1}^{n-1} q_i}$$

con campo di variazione compreso tra 0 in caso di equi-distribuzione e 1 nel caso di massima concentrazione.

Nella Tab. 11 è stato calcolato l'indice di Gini nei due centri con distinzione in base al genere, mentre nelle Figg. 3 e 4 si riportano le rispettive curve di Lorenz.

Tabella 11. *Indice di Gini dei lasciti in denaro a Ruvo di Puglia e Giovinazzo per genere dei benefattori*

Genere	Ruvo	Giovinazzo
M	0,579	0,627
F	0,167	0,652
Totale	0,427	0,643

Dai dati della precedente tabella emerge come nel complesso i lasciti di Giovinazzo presentano una maggiore concentrazione rispetto a quelli di Ruvo; tale differenza è da imputare probabilmente anche al diverso tipo di economia nei due centri (agricolo a Ruvo e marittimo-commerciale a Giovinazzo). Se si esaminano i valori per genere il divario è minimo tra i due centri per il genere maschile, mentre è particolarmente sensibile per il genere femminile. A Ruvo si osserva una notevole differenza tra i due generi, con un valore nettamente maggiore per gli uomini, mentre a Giovinazzo i due valori dell'indice di Gini sono abbastanza simili, con una lieve prevalenza di quello delle donne.

Il valore basso dell'indice per il genere femminile a Ruvo potrebbe trovare una spiegazione nella maggiore omogeneità delle condizioni economiche delle donne per le quali dall'obituario mancano del tutto informazioni sulla qualifica (su 282 solo di 15 è riportata la qualifica).

Figura 3. *Curve di Lorenz dei lasciti in denaro Ruvo vs Giovinazzo*

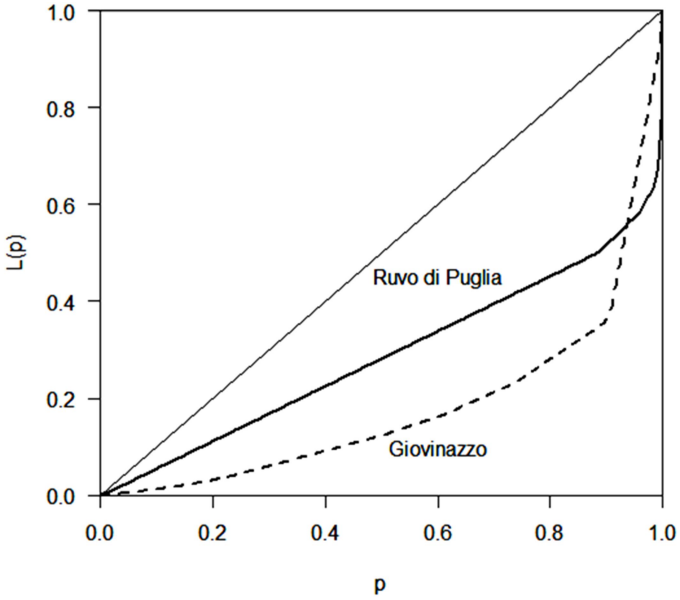
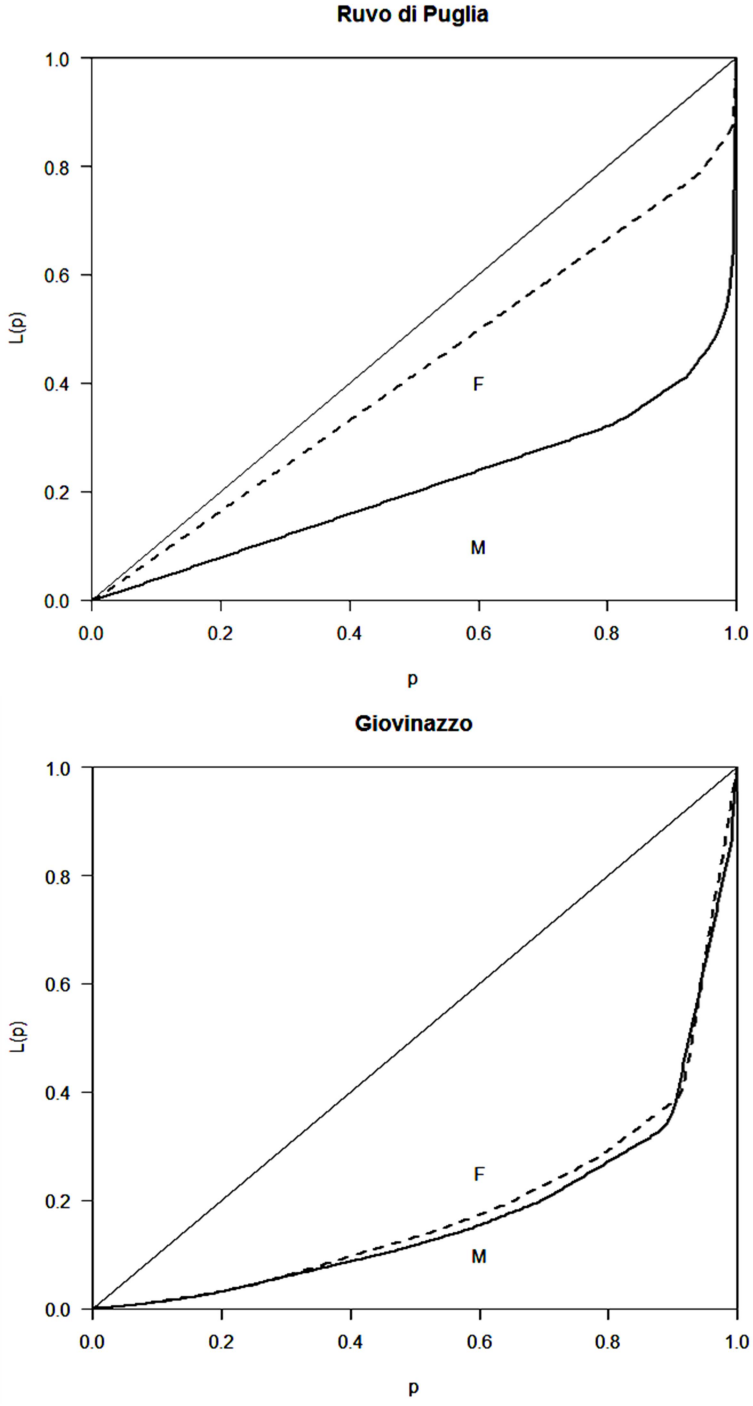


Figura 4. *Curve di Lorenz dei lasciti in denaro Ruvo vs Giovinazzo, secondo il genere*



Negli ultimi anni si è sviluppato un notevole interesse per lo studio e la misurazione della disuguaglianza nella distribuzione del reddito e della ricchezza in età pre-industriale, ne è un esempio il progetto EINITE[1] (*Economic Inequality across Italy and Europe, 1300-1800*) con il quale si è creata una base di dati relativa a diversi Paesi europei per un periodo che va dal tardo Medioevo sino agli anni precedenti la Rivoluzione industriale. Dall'esame di alcune recenti pubblicazioni (Alfani, Sardone, 2015) emerge che per il Mezzogiorno continentale italiano i primi dati disponibili risalgono solo al XVI secolo, mentre per altri contesti geografici italiani (ad esempio Toscana, Piemonte) si dispone di informazioni risalenti al XIV secolo desumibili da rilevazioni di tipo fiscale (estimi).

Purtroppo nel Regno di Napoli tali fonti sono piuttosto rare, se non proprio inesistenti. Per sopperire a tale carenza in questo paragrafo si vuole proporre un nuovo metodo di lavoro per ottenere la stima di indici di disuguaglianza nella distribuzione della ricchezza basato sull'utilizzo dei lasciti in denaro desunti dagli obituari di Ruvo e di Giovinazzo. Dall'analisi di tali dati è emerso come le donazioni siano collegate alla posizione socio-economica dei benefattori, registrando importi medi più elevati per i ceti alti (nobili e clero) e più bassi per gli altri (artigiani, notai, giudici), e pertanto è lecito supporre che le oblazioni siano direttamente funzionali alla capacità economica del donante, ovvero al suo reddito e alla sua ricchezza.

Si ritiene quindi che le donazioni possono essere una buona *proxy* della ricchezza degli abitanti dei due centri nella prima metà del Trecento; a ragione di ciò dall'analisi della distribuzione delle donazioni in denaro è possibile ottenere delle stime di indicatori della disuguaglianza della distribuzione della ricchezza (coefficiente di Gini). Si ritiene più opportuno collegare gli importi delle donazioni alla ricchezza piuttosto che al reddito in quanto in una società di tipo prettamente agrario, come quella dei due centri nel periodo in esame, si può considerare abbastanza plausibile l'esistenza di un rapporto di quasi proporzionalità tra le due grandezze, ritenendo questo particolarmente valido per i ceti più elevati che dal patrimonio essenzialmente fondiario traevano delle rendite; la ricchezza era quindi fondamentale per la produzione del reddito. Occorre considerare anche che nel periodo pre-industriale è poco probabile che vi fossero significative divergenze tra il livello di disuguaglianza del reddito e quello della ricchezza.

Un'ulteriore circostanza che spinge a considerare le elargizioni dei benefattori come *proxy* della ricchezza è che, oltre alle donazioni monetarie, vi erano anche quelle in natura che in molti casi erano costituite da beni immobili (case, terreni,

[1] http://www.dondena.unibocconi.it/wps/wcm/connect/cdr/centro_dondena/home/research/einite.

chiese), espressione quindi della ricchezza fondiaria del donante piuttosto che del reddito corrente. Le donazioni in beni fondiari sono risultate pari al 33,3% a Ruvo e al 9,4% a Giovinazzo. Un conforto viene anche da un punto di vista statistico, dato che le distribuzioni delle donazioni in denaro sembrano seguire la funzione paretiana, modello teorico che in genere descrive la distribuzione della ricchezza, con un buon grado di adattamento misurato tramite l'indice di determinazione.

Il livello di disuguaglianza ottenuto con il coefficiente di Gini nei due centri è sicuramente sottostimato in quanto negli obituari i ceti più popolari difficilmente potevano entrare in quanto era previsto il pagamento di una "quota" minima non sempre alla portata delle fasce più povere della popolazione. Tuttavia questo non costituisce un limite in quanto esso si presenta anche con l'utilizzo di fonti fiscali come gli estimi, gli apprezzi e i catasti nei quali i nullatenenti e i poveri in genere non erano registrati e risultavamo assenti nel computo della misura della disuguaglianza.

Come termine di confronto delle stime della disuguaglianza della distribuzione della ricchezza ottenute dagli obituari si è utilizzato il dato desunto da un apprezzo (*Liber appretii*) del 1417 redatto nella città di Molfetta (De Gennaro, 1963) confinante geograficamente con Giovinazzo e distante appena 15 chilometri da Ruvo di Puglia. Tale fonte presenza diversi punti di contatto con gli estimi dell'Italia Centro-settentrionale e sicuramente fornisce un quadro della misura della concentrazione della ricchezza a Molfetta agli inizi del Quattrocento. Se Ruvo di Puglia nel Medioevo, sebbene sede diocesana, è da considerare come un centro rurale, Giovinazzo e Molfetta erano di sicuro dei centri urbani che presentano diverse caratteristiche in comune: oltre ad essere contigue, erano simili per dimensione demografica, almeno nel corso del Trecento, entrambe erano sedi vescovili, ubicate sulla costa e pertanto città portuali, erano città demaniali, quindi non erano infeudate, possedevano un vasto hinterland, con piccoli insediamenti rurali (casali, villaggi), ove era praticata l'olivicoltura e la produzione di olio era destinata all'esportazione verso l'Italia centro-settentrionale.

Il computo del coefficiente di Gini con i dati degli obituari ha prodotto i valori di 0,427 per Ruvo e 0,643 per Giovinazzo che costituiscono una stima della disuguaglianza nella distribuzione della ricchezza in questi due centri; invece per Molfetta nel 1417 si è ottenuto un indice di Gini pari a 0,532. I valori sembrano confermare quanto riscontrato in altri ambiti geografici, ovvero come il coefficiente di Gini risulti più basso nei centri rurali rispetto a quelli urbani e nei centri di minori dimensioni rispetto a quelli più grandi; inoltre confrontando il valore di Giovinazzo della prima metà del Trecento con quello di Molfetta del 1417, quindi dopo la Pe-

ste Nera che imperversò tra il 1347 e il 1352 e dopo le nefaste conseguenze delle guerre dinastiche all'interno della casa d'Angiò che videro la Puglia come principale terreno di scontro – Ruvo di Puglia nel 1349 fu assediata e saccheggiata con gravi conseguenze sulla popolazione (non sappiamo se e in quale misura tale evento ebbe qualche effetto sulle donazioni dell'obituario) –, si nota una riduzione della disuguaglianza, avendo l'epidemia di peste, e probabilmente anche le conseguenze delle guerre dinastiche, un effetto perequatore nella distribuzione della ricchezza (Alfani, 2015; Ammannati, 2015).

Si consideri anche come la quota dei più ricchi (top 10%) passa dal 64,6% di Giovinazzo prima della Peste Nera al 39,7% di Molfetta dei primi del Quattrocento. Il confronto tra Giovinazzo e Molfetta appare alquanto peregrino date le caratteristiche molto simili dei due centri geograficamente confinanti; è anche opportuno il confronto tra le risultanze dei due necrologi di Ruvo e Giovinazzo e quelle del–l'apprezzo di Molfetta in quando in entrambi i casi si ha una distribuzione troncata a sinistra: per poter essere iscritti negli obituari occorreva elargire una quota minima d'ingresso che i ceti più poveri non potevano permettersi, mentre nell'apprezzo sono esclusi i nullatenenti (e anche i possessori della sola casa di abitazione che era esentata dalla tassazione) in quanto si tratta di un censimento delle proprietà fondiarie.

6. Conclusioni

In questo contributo sono state esaminate le donazioni *pro anima* effettuate dai benefattori del Trecento a favore delle cattedrali di Ruvo di Puglia e di Giovinazzo; in primo luogo sono state prese in considerazione le caratteristiche dei donatori soffermandosi su genere e qualifica socio-economica dei medesimi. I dati hanno mostrato come non vi sono differenze notevoli nella composizione tra i due centri, presentandosi una maggiore eterogeneità per i benefattori di Ruvo.

In seguito sono state analizzate le donazioni secondo la tipologia: in entrambi i centri le donazioni più frequenti sono risultate quelle in denaro, sebbene con diversa incidenza percentuale. Si è condotta un'analisi esplorativa delle elargizioni in denaro e di seguito sono stati condotti alcuni test statistici che hanno evidenziato differenze molto significative tra i due centri. Il modello paretiano si è mostrato particolarmente adatto a descrivere le distribuzioni dei lasciti in denaro sia a Ruvo che a Giovinazzo; si è proceduto alla stima dei parametri di tale funzione con il metodo dei minimi quadrati.

Nell'ultimo paragrafo ci si è soffermati sulla misura della concentrazione e della disuguaglianza dei lasciti; l'indice di Gini ha evidenziato differenze sostanziali tra Ruvo e Giovinazzo, in base al genere non vi sono significative differenze a Giovinazzo, mentre sono presenti a Ruvo. Si è proposto anche di utilizzare le donazioni in denaro come una variabile *proxy* della ricchezza dei benefattori e stimare la misura di disuguaglianza nella distribuzione di questa grandezza con quella delle donazioni, sopperendo alla mancanza di fonti fiscali (estimi, apprezzi, catasti) che vengono usualmente impiegati per analisi del genere per la prima metà del Trecento nel Regno di Napoli. I risultati ottenuti costituiscono il punto di partenza per ulteriori indagini ed analisi.

Riferimenti bibliografici

Alfani, G., Sardone, S. (2015). *Long-term trends in economic inequality in southern Italy. The Kingdoms of Naples and Sicily, 16th-18th centuries: first results*. Economic History Association, annual meeting, Nashville, September 11-13, 2015 (consultato on line al seguente indirizzo: http://eh.net/eha/wp-content/uploads/2015/05/Alfani.pdf).

Alfani, G. (2015). Economic inequality in northwestern Italy: A long-term view (fourteenth to eighteenth centuries), *Journal of Economic History*, 75 (4), 1058-1096.

Ammannati, F. (2015). La Peste Nera e la distribuzione della proprietà nella Lucchesia del tardo medioevo, *Popolazione e Storia*, 2, 1-45.

Clauset, A.; Shalizi C. R.; Newman, M. E. J. (2009). Power-Law Distributions in Empirical Data. *Society for Industrial and Applied Mathematics Review*, 51 (4): 661-703.

De Gennaro, G. (1963). *Il Liber appretii di Molfetta dei primi del Quattrocento*, Istituto di storia economica, Università di Bari, Bari.

Ficco, F. (2005). La "Matricula" della Cattedrale di Ruvo. Un obituario inedito del '300. *Odegitria*, 12: 315-379.

Garufi, C. A. (1911). L'obituario della "Confraternita dell'Episcopio" conservato nell'Archivio Capitolare di Giovinazzo (cod. n. 12). *Apulia*, 2: 5-36 e 150-158.

Piccolo, D. (1998). *Statistica*, Il Mulino, Bologna.

Quandt, R. E. (1966). Old and new methods of estimation and the Pareto distribution. *Metrika* 10 (1): 55-82.

Rytgaard, M. (1990). Estimation in the Pareto distribution. *ASTIN Bulletin*, 20: 201-216.

Soliani, L. (2010). *Manuale di statistica: statistica univariata e bivariata, parametrica e non-parametrica*, Uninova, Parma.

UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO
Metodi e analisi statistiche 2017 ISBN: 978-88-6629-013-1
DIPARTIMENTO DI ECONOMIA E FINANZA pp. 283-316

Analisi Fattoriale sui dati *INVALSI*

Leonardo Mariella[*1], **Marco Tarantino**[**2]
[1] *Università del Salento*
[2] *I.I.S.S. "Giovanni Penna", Asti (AT)*

Riassunto: L'*Analisi Fattoriale* permette di costruire un modello, noto come *modello fattoriale*, sulla base del quale ogni variabile osservata risulti parzialmente influenzata sia da fattori comuni che da fattori unici. L'eventuale relazione esistente tra un fattore sottostante ed una variabile osservata dipende dall'influenza che tale fattore esercita su alcune variabili rispetto alle altre. Il punto fondamentale di tale metodologia è rappresentato dalla matrice di covarianza (o correlazione) tra le variabili osservate, le quali, se altamente (scarsamente) correlate, risultano influenzate dagli stessi (differenti) fattori.
Nel presente lavoro, l'*Analisi Fattoriale* è stata applicata ai risultati ottenuti nella prova *INVALSI* di Italiano dalle classi II della Scuola Secondaria di II grado, con l'obiettivo di ipotizzare un modello teorico, in grado di valutare la corrispondenza o meno dei singoli quesiti all'aspetto di comprensione della lettura o all'ambito grammaticale di appartenenza.

Keywords: Analisi Fattoriale; correlazione policorica; correlazione prodotto-momento; statistica chi quadrato generalizzata; *INVALSI*.

1. L'Analisi Fattoriale

L'*Analisi Fattoriale* rappresenta un insieme di metodi utilizzati per esaminare le relazioni sottostanti un determinato numero di variabili. Più in dettaglio, tale analisi può essere *Esplorativa* o *Confermativa*.

L'*Analisi Fattoriale Esplorativa* individua le variabili latenti che influenzano un insieme di variabili osservate. L'obiettivo è quello di determinare il numero di fattori comuni che influenzano un insieme di variabili e, conseguentemente, l'entità della

[*] leonardo.mariella@unisalento.it
[**] marco.tarantino1@istruzione.it

relazione tra il singolo fattore e la variabile osservata.
Solitamente, tale approccio viene impiegato per individuare le relazioni sottostanti le risposte in una area di contenuti specifici, gli insiemi di *item* che "si sostengono l'un l'altro" in un questionario, la dimensionalità di una scala di misura, le caratteristiche più importanti nella classificazione di un gruppo di *item*, i valori di relazioni sottostanti (punteggi fattoriali) da utilizzare in ulteriori analisi.

Per eseguire una *Analisi Fattoriale Esplorativa*, è necessario:

1. selezionare le variabili dalle unità statistiche prese in esame;
2. calcolare la matrice di covarianza (o correlazione) tra le variabili;
3. individuare il minor numero di fattori che "conservino" gran parte della covarianza presente nei dati;
4. stimare l'insieme iniziale di fattori;
5. ruotare opportunamente i fattori ottenuti, al fine individuare una soluzione uguale a quella iniziale, ma più facile da interpretare;
6. costruire punteggi fattoriali, nel caso in cui si desideri eseguire ulteriori analisi, utilizzando i fattori come variabili.

D'altra parte, l'*Analisi Fattoriale Confermativa* verifica se l'insieme di variabili latenti influenzi le variabili osservate, secondo le relazioni ipotizzate. L'obiettivo principale è quello di testare l'adattabilità del modello fattoriale scelto ad un insieme di dati.

Tale approccio viene utilizzato per valutare la validità di un modello fattoriale, la comparabilità tra due differenti modelli sullo stesso insieme di dati, la significatività di un peso fattoriale specifico, la relazione tra due o più pesi fattoriali, la presenza o meno di correlazione tra un insieme di fattori, la validità discriminante e convergente di un insieme di variabili.

Per eseguire una *Analisi Fattoriale Confermativa*, è necessario:

1. definire il modello fattoriale da testare, individuando il numero di fattori ed il valore dei pesi tra i fattori e le variabili;
2. selezionare le variabili dalle unità statistiche prese in esame;
3. calcolare la matrice di covarianza (o correlazione) tra le variabili;
4. stimare i pesi fattoriali, ipotizzati liberi di variare;
5. valutare l'adeguatezza del modello ed, eventualmente, il suo confronto con altri modelli.

In base al Decreto Legislativo 286/2004, l'*Istituto Nazionale per la VAlutazione del Sistema educativo di Istruzione e di formazione* (*INVALSI*) "attua verifiche periodiche e sistematiche sulle conoscenze ed abilità degli studenti". Più in dettaglio,

l'*INVALSI* considera *competenza linguistica*, il possesso ben strutturato di una lingua, assieme alla capacità di servirsene per i vari scopi comunicativi.

Dunque, attraverso una *Analisi Fattoriale Esplorativa*, è stato possibile costruire un modello che interpreti le relazioni tra gli *item* (variabili osservate) presenti nella prova, in termini di fattori sottostanti (variabili latenti) e, successivamente, attraverso una *Analisi Fattoriale Confermativa*, valutare la bontà di adattamento del modello ipotizzato ai dati campionati. Tale modello ha premesso, poi, di verificare il grado di rappresentatività degli *item* alle macro categorie, previste dall'*INVALSI*, riguardanti particolari aspetti o ambiti da valutare.

Le tecniche di *Analisi Fattoriale* rappresentano i cardini fondamentali dell'esposizione, dal momento che permettono di esplorare la base di dati e, successivamente, proporre un modello teorico per il raggiungimento dell'obiettivo prefissato. Il caso di studio applica ai dati *INVALSI*, gli strumenti teorici trattati in precedenza, al fine di proporre una nuova chiave interpretativa sulla valutazione delle prove e, più in generale, del sistema di istruzione.

L'implementazione di tali tecniche ha richiesto l'utilizzo del software R (R Development Core Team, 2008) ed, in particolare, di alcuni suoi pacchetti applicativi, appositamente progettati per tale scopo.

2. I dati *INVALSI*

L'*INVALSI* ha il compito di "attuare verifiche periodiche e sistematiche sulle conoscenze ed abilità degli studenti"[1]. In particolare, nell'anno scolastico 2013/14, i livelli scolari coinvolti nella valutazione sono stati i seguenti:

- le classi II e V della Scuola Primaria;
- la classe III della Scuola Secondaria di I grado;
- la classe II della Scuola Secondaria di II grado.

Per tale anno scolastico, la rilevazione degli apprendimenti ha riguardato entrambi i cicli di istruzione, coinvolgendo tutte le scuole del Paese, statali e paritarie (circa 13 200) e tutti gli studenti delle classi dei quattro livelli scolari interessati.

Nel presente lavoro, le tecniche di *Analisi Fattoriale* sono state applicate ai risultati ottenuti nella prova *INVALSI* di Italiano dalle classi II campione della Scuola Secondaria di II grado. L'obiettivo è stato quello di esprimere le singole risposte (variabili osservate) in un numero inferiore di fattori (variabili ipotetiche o latenti) e, successivamente, di ipotizzare un modello per valutare la corrispondenza o meno dei singoli quesiti all'aspetto di comprensione della lettura ed ambito grammaticale di appartenenza. L'implementazione di simili tecniche alla base di dati oggetto di studio ha richiesto l'utilizzo del software R (R Development Core Team, 2008) ed, in particolare, di alcuni suoi pacchetti applicativi.

[1] Decreto Legislativo 286/2004.

3. Le modalità di svolgimento delle prove

Nell'anno scolastico 2013/14, la somministrazione delle prove *INVALSI* è iniziata il 6 maggio ed è terminata il 19 giugno con la *prova nazionale* della classe III della Scuola Secondaria di I grado: in tale livello scolare, la prova è parte dell'Esame di Stato di Licenza Media.[2] Le prove sono state somministrate nello stesso giorno, una di seguito all'altra, ad eccezione della Scuola Primaria, in cui tale rilevazione si è svolta, come negli anni precedenti, in due giornate distinte, al fine di evitare l'*effetto affaticamento*, dovuto all'età degli alunni coinvolti (AA.VV., 2014).

Sebbene la rilevazione sia censuaria, per ciascun livello scolare, sono state individuate delle classi campione, nelle quali le prove si sono svolte alla presenza di un osservatore esterno (tabella 1).

Tabella 1: *classi e studenti delle prove INVALSI 2014.*

Livello	Classi	Classi campione	Studenti
II Primaria	29 719	1 468	568 251
V Primaria	29 685	1 468	561 183
III Secondaria Primo Grado	29 462	1 418	597 639
II Secondaria Secondo Grado	26 540	2 256	560 672
			2 287 745

Nelle classi non campione, la somministrazione è stata condotta da un insegnante della stessa scuola ma, solitamente, non della classe interessata dalla rilevazione e non della materia oggetto della prova. Nelle classi campione, invece, essa è avvenuta alla presenza di un osservatore esterno, il cui compito è stato quello di

- monitorare la somministrazione per garantire il rispetto delle procedure;
- riportare le risposte fornite dagli allievi su apposite schede elettroniche.

Nell'Esame di Stato della classe III della Scuola Secondaria di I grado, tale ruolo è stato svolto dal Presidente di commissione.

Il tempo previsto (in *minuti*) per lo svolgimento di ciascuna prova è stato differenziato in base al livello scolare (tabella 2).

Tabella 2: *tempi di somministrazione delle prove INVALSI 2014.*

Prova	Primaria		Secondaria I grado	Secondaria II grado
	classe II	classe V	classe III	classe II
Preliminare di lettura	2			
Italiano	45	75	75	90
Matematica	45	75	75	90
Questionario studente		30		30

[2] Legge 176/2007.

Al termine di ciascuna giornata di somministrazione, l'*INVALSI* ha inviato a tutte le scuole, tramite *e-mail*, le griglie di correzione delle prove. Sia per le classi campione, che per le classi non campione, anche se con scadenze temporali differenziate, tutti i dati relativi alle classi sono stati poi trasmessi, per via telematica, all'*INVALSI*, mediante apposite maschere elettroniche.

4. L'attendibilità dei dati

Al fine di prevenire comportamenti scorretti da parte di studenti e/o insegnanti (*cheating*), i fascicoli delle prove sia di Italiano che di Matematica sono stati predisposti in cinque versioni differenti: per ciascuna domanda, le opzioni di risposta sono state disposte in ordine diverso e, nel caso delle prove di Matematica, sono state anche ruotate le domande relative ai vari ambiti di contenuto. Nelle classi campione, inoltre, alla tradizionale presenza di osservatori esterni, si è affiancata la presenza di controllori di secondo livello, inviati in talune scuole, scelte casualmente ed indipendentemente dall'essere o meno parte del campione, al fine di riportare informazioni sul grado di regolarità della somministrazione e della successiva correzione delle prove.

Si osservi che, nel 2013, le procedure di correzione del *cheating* sono state riviste. La metodologia seguita ha tenuto conto della differenza, che comunque permane nei risultati tra classi campione, ove la somministrazione è vigilata da un osservatore esterno, e classi non campione, ed opera iterativamente, allo scopo di prevenire il rischio che una *performance* particolarmente brillante di una classe venga erroneamente attribuita ad anomalie (*falsi positivi*). Più in dettaglio, tale procedura segue i seguenti passi, effettuati separatamente per ciascuna prova (Italiano e Matematica) e per ciascun livello scolare:

1. si esaminano i dati grezzi di ciascuna classe sulla base di 4 indicatori, ovvero

   - media dei risultati all'interno della classe,
   - variabilità dei risultati all'interno della classe,
   - grado di omogeneità del modello delle risposte,
   - risposte omesse,

   al fine di fornire una prima valutazione sulla presenza di anomalie (Quintano et al., 2009);

2. sulla base dei dati delle classi campione, si stimano modelli di regressione esplicativi della media e della variabilità interna dei risultati di ogni classe, campione e non campione (*fitting over sample*), in cui le covariate sono, in prevalenza, variabili relative alla composizione della classe medesima;

3. si stima un punteggio medio di classe corretto, combinando tra loro:

- la stima della variabilità interna alla classe (punto 2),
- due indicatori di plausibilità, costruiti a loro volta utilizzando
  - la stima della variabilità (punto 2),
  - il valore della correlazione tra risultati grezzi nelle prove *INVALSI* e voti attribuiti ai singoli alunni nel I quadrimestre;

tali risultati vengono ritenuti tanto più plausibili e, quindi, non anomali (sebbene con elevata media e bassa variabilità all'interno della classe), quanto più bassa è la variabilità "spiegata" da fattori di composizione e quanto più elevata è la correlazione tra voti e risultati;

4. si modifica l'entità della correzione apportata ai dati grezzi, mediante la procedura precedente (punto 1).

Si osservi che le anomalie sono derivanti, in parte, dal modello dei risultati grezzi (punto 1) e che quest'ultimo può risentire di caratteristiche intrinseche di ciascuna prova. Per tale ragione, si procede comunque a correggere i risultati, solo nella misura in cui la correzione stimata per ciascuna classe (punto 3) superi la mediana dei valori nella macro-area maggiormente "virtuosa", intesa come quella ove la correzione per le anomalie (punto 3) risulta complessivamente meno intensa.

5. La prova di Italiano

I principi ispiratori e le linee-guida che sottostanno alla struttura e ai contenuti delle prove sono ampiamente illustrate e discusse nel *Quadro di Riferimento* per la prova di Italiano nell'istruzione obbligatoria, coerente con l'attuale formulazione delle Indicazioni Nazionali e con le Linee-guida per i Licei, gli Istituti Tecnici e gli Istituti Professionali[3]. Più in dettaglio, la prova della classe II della Scuola Primaria è fatta precedere da una prova preliminare di velocità di lettura, comprendente 40 elementi (*item*), ciascuno dei quali richiede la corretta associazione tra una parola ed una delle quattro figure assegnate. Tale test non prevede l'assegnazione di alcun punteggio: il suo obiettivo, invece, è quello di rilevare la percentuale di alunni con un insufficiente grado di automatismo nella decodifica di parole scritte.

La prova di Italiano della classe III della Scuola Secondaria di I grado è inserita, assieme alla parallela prova di Matematica e con lo stesso peso, nella Prova nazionale dell'esame di conclusione del primo ciclo di istruzione. Tale prova, dunque, è l'unica ad avere un duplice obiettivo:

- monitorare l'efficacia del sistema di istruzione;
- contribuire alla valutazione degli studenti.

[3] Regolamento Ministeriale del 16 novembre 2012.

Per tale ragione, i punteggi dei test di Italiano e di Matematica sono trasformati, attraverso una procedura, definita di anno in anno, in un unico voto decimale, nel quale confluisce l'esito di entrambe le prove.

Le prove di Italiano di tutti i livelli scolari interessati alle rilevazioni *INVALSI* sono suddivise in sezioni, ciascuna delle quali prende in esame differenti aspetti della materia oggetto di studio (tabella 3).

Tabella 3: *caratteristiche delle prove di Italiano.*

Classe	Sezioni	Numero quesiti per formato			*Item*
		scelta multipla		risposta	
		semplice	complessa	aperta	
II primaria	testo narrativo	16	2	2	30
	esercizi linguistici		2		17
V primaria	testo narrativo	14	1	4	22
	testo espositivo	9	1	4	19
	grammatica	5	2	3	26
III secondaria I grado	testo narrativo	12	1	7	30
	testo espositivo	14	1	3	24
	grammatica	7	1	2	23
II secondaria II grado	testo regolativo misto	6	1	1	12
	testo narrativo letterario	17	2	4	39
	testo espositivo	8	3	3	25
	testo non continuo	4		1	5
	grammatica	6	1	2	20

Si osservi che, il numero dei quesiti non coincide con il numero degli *item*, in quanto uno stesso quesito può avere più di un *item*.

5.1. Analisi multivariata dei dati INVALSI

Nelle prove di Italiano somministrate nell'anno scolastico 2013/14, gli *item* (variabili osservate) che caratterizzano il collettivo di studenti sono raggruppate dall'*INVALSI*, sulla base degli aspetti di comprensione della lettura e degli ambiti grammaticali, per ognuna delle classi interessate alle rilevazioni (tabella 4).

Si osservi che, gli esercizi linguistici sono valutati sulla base dei seguenti criteri:

1. riconoscere il significato uguale o contrario di coppie di parole;

2. collegare in maniera congruente soggetto e predicato di una serie di 5 frasi.

L'*Analisi Fattoriale* permette, dunque, di ottenere un ristretto numero di nuove variabili (variabili latenti o fattori) che ricostruiscono, attraverso una combinazione lineare, le variabili originarie, solitamente correlate tra di loro, presenti nella base di dati.

Nonostante siano state proposte molte regole empiriche, il numero di soggetti necessari per ottenere una stima stabile dei fattori dipende, in realtà, dalla chiarezza della struttura in esame (MacCallum, Widaman, Preacher et al., 2001; MacCallum, Widaman, Zhang et al., 1999).

Tabella 4: *aspetti di comprensione e ambiti grammaticali.*

Classe	Sezioni	N. quesiti per aspetto/ambito							
		1	2	3	4	5a	5b	6	7
II primaria	testo narrativo	3	3	3	2	6	2	1	
	esercizi linguistici								
V primaria	testo narrativo	2	2		2	6	5	2	
	testo espositivo	2	3	1	1	3	1	3	
	grammatica	1	3	1	2	2		1	
III secondaria I grado	testo narrativo	6	2	3	1	6	1	1	
	testo espositivo	3	6	2	2	2	1	2	
	grammatica	1	2	1	2	3		1	
II secondaria II grado	testo regolativo misto	3			1	1	1	1	1
	testo narrativo letterario	6	3	2	2	3	2	4	1
	testo espositivo	3	2	2	3	3	1		
	testo non continuo		2	2		1			
	grammatica	1	3	1	1	1		2	

Il presente caso di studio analizza la base di dati riguardante le classi II della Scuola Secondaria di II grado, facenti parte del campione individuato dall' *INVALSI*, composta da $36\,618$ studenti: la presenza di un osservatore esterno, infatti, dovrebbe aver aumentato l'attendibilità dei dati raccolti ed evitato, per quanto possibile, il fenomeno del *cheating*. In particolare, ad ogni singolo studente, corrispondono una serie risposte di tipo vero/falso agli *item* presenti sul questionario, riguardanti aspetti di comprensione della lettura ed ambiti grammaticali individuati dall'*INVALSI* (tabella 5).

Tabella 5: *item valutati nella prova di Italiano.*

Aspetto								Ambito					
1	2	3	4	5a	5b	6	7	1	2	3	4	5	6
A4	B8	B18	A5	A2	A3	A1	A6_a	E6_1	E2	E5	E1_a	E9	E4
A7	B21	B20_a	B3	B5	B1	B4	A6_b	E6_2	E3		E1_b		E8
A8	B22_a	B20_b	B16	B11	B7	B13	A6_c	E6_3	E7_a		E1_c		
B2	B22_b	B20_c	C5	B14	C2	B17	A6_d	E6_4	E7_b				
B6	B22_c	B20_d	C6	C4		B23	A6_e	E6_5	E7_c				
B9	B22_d	B20_e	C11	C12_a			B12		E7_d				
B10	B22_e	C1		C12_b					E7_e				
B15	B22_f	D1		C12_c					E7_f				
B19	B22_g	D4		C12_d									
C8	C3			C12_e									
C9	C7_a			C14									
C10	C7_b			D3									
	C7_c												
	C7_d												
	C7_e												
	C13												
	D2												
	D5												

Dal momento che, inoltre, una *AFC* richiede una dimensione campionaria più grande rispetto ad una *AFE*, in quanto genera statistiche inferenziali, si è scelto di

estrarre, senza ripetizione, una serie di sottocampioni; in particolare:

1. un sottocampione di $1\,000$ studenti, per individuare, attraverso un'*AFE*, i fattori che rappresentano la struttura portante dei dati in una forma sintetica;
2. un sottocampione di $2\,000$ studenti, per testare, attraverso una *AFC*, la generalità dei fattori estratti;
3. una successione di sottocampioni, per testare la validità del modello ottenuto.

Le diverse componenti della competenza di lettura sono raggruppate in 6 aspetti relativi alla prima parte della prova.

Definizione 5.1 (Aspetto 1)**.** *Comprendere il significato, letterale e figurato, di parole ed espressioni e riconoscere le relazioni tra parole.*

Le domande relative all'aspetto 1 mirano a

- individuare o spiegare il significato di un termine o di una espressione usati nel testo; saper distinguere tra significato letterale e figurato di una parola, di un'espressione o di una frase; saper riconoscere le relazioni tra parole del testo;
- trovare nel testo il termine che corrisponde a una spiegazione in esso fornita o a una definizione data nella formulazione del quesito.

Definizione 5.2 (Aspetto 2)**.** *Individuare informazioni date esplicitamente nel testo.*

Nell'aspetto 2, sono comprese le domande in cui si chiede di ritrovare, nel testo, una o più informazioni date in maniera esplicita, o anche tramite una parafrasi di quanto detto.

Definizione 5.3 (Aspetto 3)**.** *Fare un'inferenza diretta, ricavando un'informazione implicita da una o più informazioni date nel testo e/o tratte dall'enciclopedia personale del lettore.*

Le domande relative all'aspetto 3 valutano la capacità di inferire una singola informazione puntuale, non data in maniera esplicita nel testo, da una o più informazioni in esso presenti, attingendo anche all'enciclopedia personale.

Rientrano in tale aspetto anche le domande che richiedono l'operazione inversa: data una certa informazione, rintracciare nel testo la frase o le frasi da cui essa può essere inferita.

Definizione 5.4 (Aspetto 4)**.** *Cogliere le relazioni di coesione e di coerenza testuale (organizzazione logica entro e oltre la frase).*

Le domande che rientrano nell'aspetto 4 riguardano

- la coesione, al fine di individuare il riferimento di anafore e catafore, comprendere il significato dei connettivi, dei segni di interpunzione e, in generale, dei legami grammaticali e testuali fra elementi o parti del testo;

- la coerenza, al fine di saper cogliere i rapporti logico-semantici fra parti del testo.

Definizione 5.5 (Aspetto 5a)**.** *Ricostruire il significato di una parte più o meno estesa del testo, integrando più informazioni e concetti, anche formulando inferenze complesse.*

In riferimento all'aspetto 5a, è necessario rielaborare il testo, collegando e integrando più informazioni e concetti, espressi sia in maniera esplicita che implicita in un punto o anche in punti diversi, anche basandosi sull'enciclopedia personale. In particolare, tali domande sono focalizzate su singoli punti, passaggi o parti del testo.

Definizione 5.6 (Aspetto 5b)**.** *Ricostruire il significato globale del testo, integrando più informazioni e concetti, anche formulando inferenze complesse.*

In riferimento all'aspetto 5b,

- le domande suppongono un punto di vista globale sul testo e sul suo significato;
- le domande, pur formulate su un argomento specifico, richiedono che si tenga presente e si consideri l'insieme del testo e ciò che esso vuol complessivamente comunicare.

Definizione 5.7 (Aspetto 6)**.** *Sviluppare un'interpretazione del testo, a partire dal suo contenuto e/o dalla sua forma, andando al di là di una comprensione letterale.*

Nell'aspetto 6, sono comprese le domande che presuppongono una "presa di distanza" dal testo, un'osservazione "esterna" del suo contenuto o delle sue caratteristiche formali, per identificarne il messaggio, lo scopo, l'intenzione comunicativa, o per riconoscerne il genere, il registro, il tono, lo stile.

Definizione 5.8 (Aspetto 7)**.** *Riflettere sul testo e valutarne il contenuto e/o la forma alla luce delle conoscenze ed esperienze personali.*

Nell'aspetto 7, sono comprese le domande che chiedono di riflettere sul testo e di valutare contenuto o forma. Tali domande si distinguono da quelle incluse nell'aspetto precedente, poiché sollecitano l'espressione di un giudizio o di una presa di posizione da parte del lettore.

I quesiti della sezione grammaticale della prova sono classificati in ambiti di contenuto, a seconda dell'argomento su cui vertono.

Ambito 1. *Ortografia*. Uso di accenti e apostrofi, maiuscole e minuscole, segmentazione delle parole, uso delle doppie, casi di non corrispondenza tra fonemi e grafemi.

Ambito 2. *Morfologia*. Flessione; categorie lessicali e sottocategorie e loro funzione nella frase.

Ambito 3. *Formazione delle parole*. Parola base e parole derivate; parole alterate; parole composte; polirematiche.

Ambito 4. *Lessico e semantica*. Relazioni di significato tra parole; polisemia; campi semantici e famiglie lessicali; usi figurati e principali figure retoriche; espressioni idiomatiche; struttura e uso del dizionario.

Ambito 5. *Sintassi*. Accordo; sintagma; frase; frase dichiarativa, interrogativa, ecc.; elementi della frase semplice; gerarchia della frase complessa; uso di tempi e modi nella frase.

Ambito 6. *Testualità*. Segnali di organizzazione del testo e fenomeni di coesione; aspetti pragmatici del linguaggio.

La sintesi delle informazioni relative alle variabili prese in esame può essere effettuata, inizialmente, attraverso i principali momenti campionari (Piccolo, 1998).

Solo successivamente, si può ricorrere all'analisi della correlazione statistica, al fine di valutare il grado di interdipendenza esistente fra due variabili. Dunque, la relazione tra due *item* può essere specificata in termini di:

- *concordanza*, nel caso in cui, al crescere (decrescere) dei punteggi di un *item*, crescono (decrescono), prevalentemente, i punteggi dell'altro *item*;
- *discordanza*, nel caso in cui, al crescere (decrescere) dei punteggi di un *item*, decrescono (crescono), prevalentemente, i punteggi dell'altro *item*.

Nel caso in cui gli *item* mutino indipendentemente l'uno dall'altro, non si avrà alcuna correlazione.

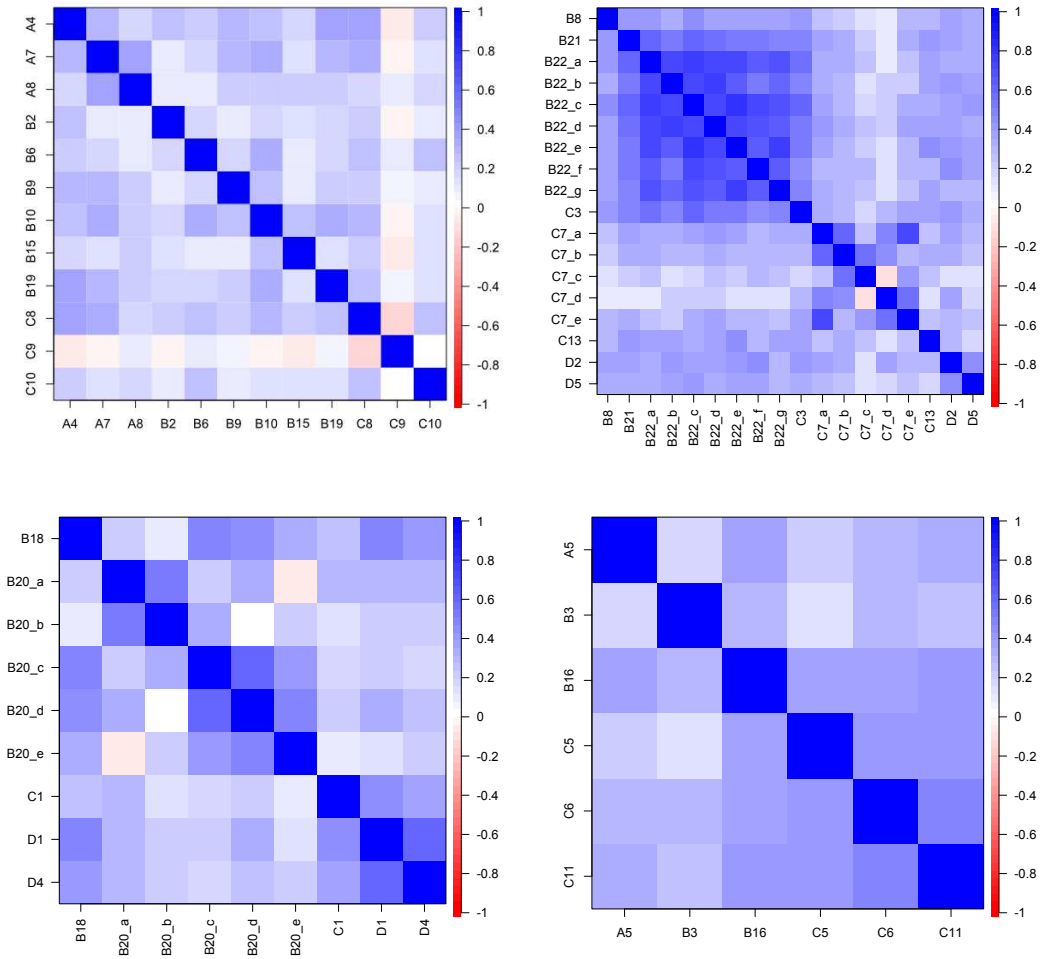
Gli indici statistici e le matrici di correlazione sono state ottenute attraverso il pacchetto applicativo `psych`. Più in dettaglio, l'analisi di alcuni indici di posizione e, conseguentemente, di variabilità sui primi quattro aspetti sulla comprensione della lettura (tabella B.1) rivelano come alcune domande si siano rivelati particolarmente ostiche per gli studenti, quali gli *item* B6, B15, C9 e C10 dell'aspetto 1, B8 e C13 dell'aspetto 2 e C1 dell'aspetto 3.

Inoltre, premesso che le variabili analizzate sono dicotomiche, si è ritenuto opportuno utilizzare il *coefficiente di correlazione policorica* (Olsson, 1979, cfr. def. A.1 in Appendice).

In particolare, le matrici di correlazione dei primi quattro aspetti (figura 1) hanno evidenziato alcune caratteristiche importanti:

- nel caso dell'aspetto 1, la discordanza dell'*item* C9 rispetto agli *item* rimanenti;
- nel caso dell'aspetto 2, la prevedibile forte concordanza tra gli *item* del gruppo B22;
- nel caso dell'aspetto 3, oltre alla prevedibile concordanza tra gli *item* del gruppo B20, anche la buona concordanza tra C1, D1 e D4.

Figura 1: *matrici di correlazione degli aspetti 1, 2, 3 e 4.*

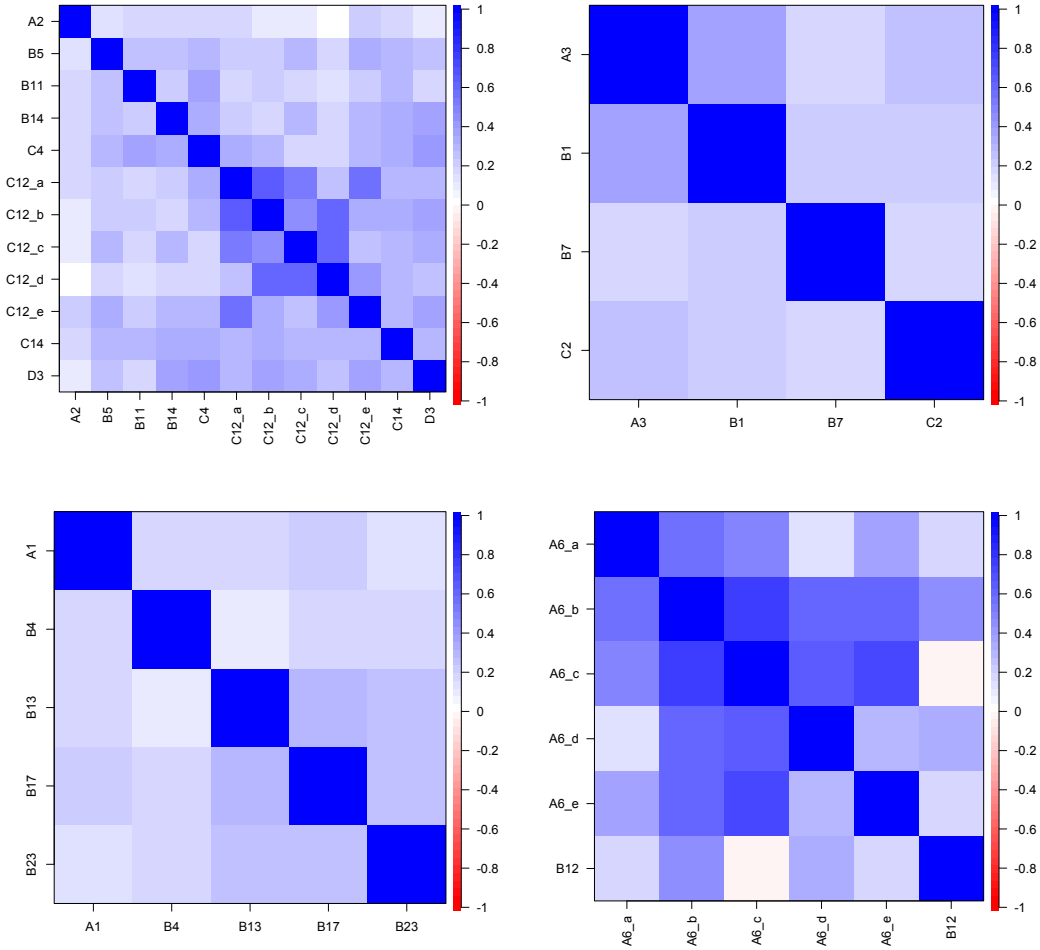


D'altra parte, l'analisi degli indici di posizione e di variabilità sugli ulteriori quattro aspetti (tabella B.2) rivelano risultati "migliori", dal momento che solo la domanda B12 è risultata essere, per il campione dei 1 000 studenti, particolarmente difficile.

L'analisi dell'interdipendenza tra le variabili, invece, ha confermato la forte concordanza tra gli *item* facenti parte di un unico gruppo (figura 2); in particolare,

- nel caso dell'aspetto 5a, la forte concordanza tra gli *item* del gruppo C12;
- nel caso dell'aspetto 7, la forte concordanza tra la maggior parte degli *item* del gruppo A6.

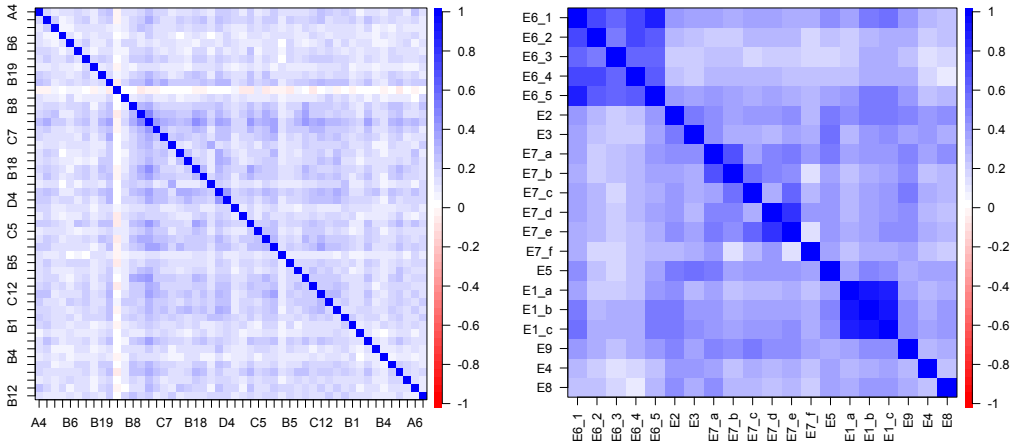
Figura 2: *matrici di correlazione degli aspetti 5a, 5b, 6 e 7.*



Dal momento che le matrici di correlazione dei diversi aspetti hanno rivelato forte concordanza tra gli *item* dello stesso gruppo di domande, si è ritenuto opportuno sintetizzare tali gruppi con un unico valore, corrispondente alla media aritmetica troncata del 5% dei valori più bassi e del 5% dei valori più alti. Dunque, la nuova base di dati (tabella B.4) ha perso la caratteristica iniziale di dicotomicità e, conseguentemente, nell'analisi dell'interdipendenza, ha richiesto l'applicazione del *coefficiente di correlazione prodotto-momento*, proposto da K. Pearson nel 1895 (cfr. def. A.2 in Appendice).

D'altra parte, la base di dati degli ambiti grammaticali (tabella B.3) ha richiesto la costruzione di una matrice di correlazione, attraverso il *coefficiente di correlazione policorica*.

Figura 3: *matrici di correlazione degli aspetti e degli ambiti.*



In riferimento sia agli aspetti sulla comprensione della lettura (figura 3a) che agli ambiti grammaticali (figura 3b), si osservi, ancora una volta, la forte concordanza presente tra gli *item* appartenenti ad uno stesso gruppo di domande.

Nel caso dei diversi aspetti, poi, si osservi la discordanza dell'*item* C9 rispetto agli altri *item* presenti nella matrice.

5.2. Analisi fattoriale esplorativa

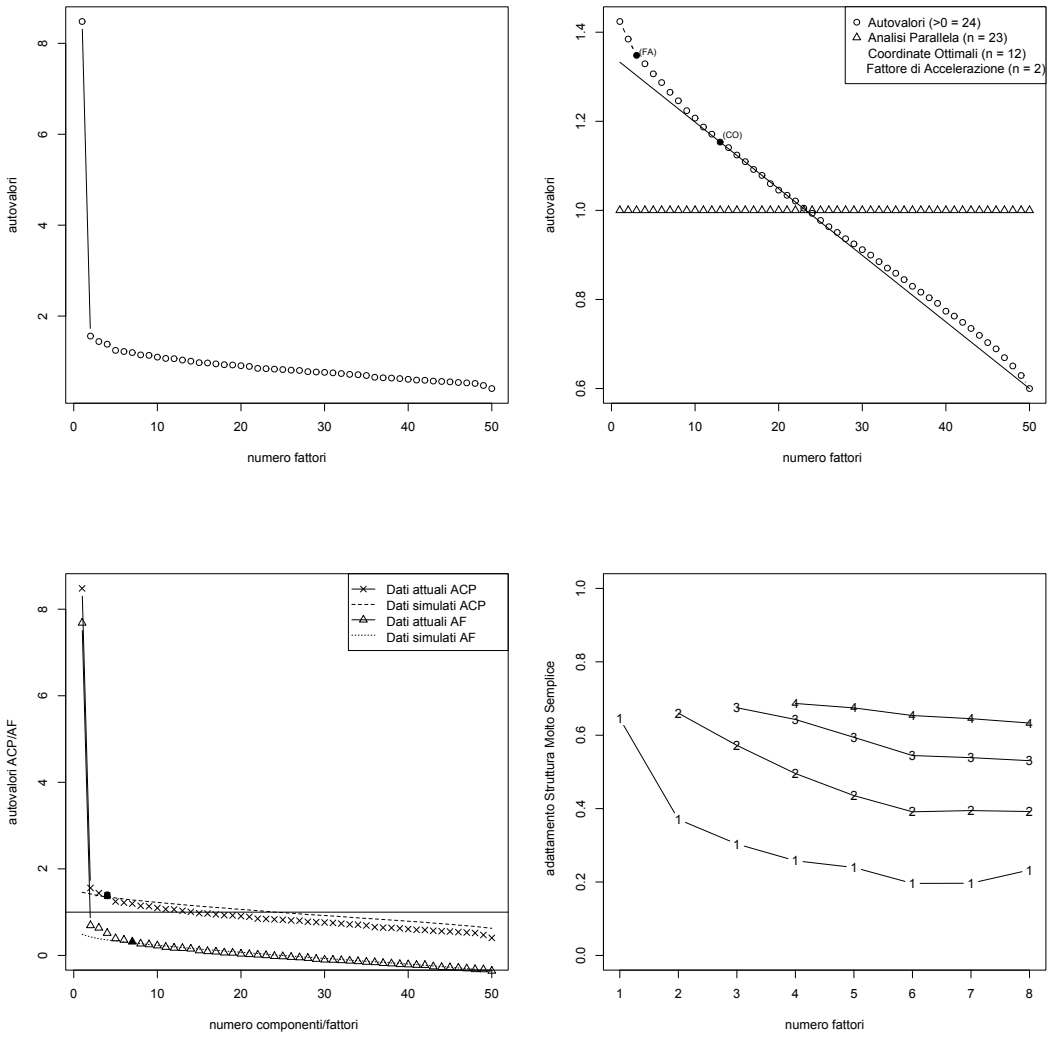
L'*AFE* individua il numero di fattori comuni che influenzano le variabili osservate e la relazione esistente tra ciascun fattore comune e la variabile corrispondente (DeCoster, 1998). In particolare, tale analisi permette di identificare il modello teorico alla base delle risposte date in un questionario, determina gli insiemi di *item* interconnessi tra loro, dimostra la profondità e la larghezza delle scale di misurazione, classifica le caratteristiche più importanti di un gruppo di *item* e, se necessario, genera i punteggi fattoriali del modello sottostante (DeCoster, 1998).

Dunque, l'approccio esplorativo consente di ridurre il numero di variabili (Williams et al., 2010), attraverso opportune combinazioni lineari (fattori).

Più in dettaglio, estratto un sottocampione di 1 000 studenti, per quanto concerne l'analisi degli aspetti di comprensione della lettura, è stato necessario, innanzitutto, individuare il numero di fattori che porrebbero sintetizzare, al meglio, le variabili osservate. Il pacchetto `nFactors` ha permesso di implementare sia metodi euristico-empirici che test di ipotesi.

Come era prevedibile, a causa dell'elevato numero di variabili osservate, il semplice *grafico degli autovalori* non risulta essere sufficientemente esplicativo (figura 4a).

Figura 4: *criteri di selezione.*



Di conseguenza, si è pensato di ricorrere a criteri di selezione alternativi (figure 4b, 4c e 4d), quali

- la *regola degli autovalori di Kaiser* (Kaiser, 1960, 1970);
- l'*analisi parallela* (Hayton et al., 2004; J. Horn, 1965);
- il criterio di *struttura molto semplice* (Revelle & Rocklin, 1979).

Ulteriori indicazioni provengono dal *test di massima verosimiglianza sugli autovalori* (cfr. def. A.3 in Appendice).

In particolare,

- nel caso della *statistica di Anderson* (Anderson, 1963), si ha

$$G = n - 1;$$

- nel caso della *statistica di Bartlett* (Bartlett, 1950, 1951; Bentler & Yuan, 1996; J. L. Horn & Engstrom, 1979), si ha

$$G = n - q - 1 - \frac{2\tilde{p}^2 + \tilde{p} + 2}{6\tilde{p}};$$

- nel caso della *statistica di Lawley e James* (James, 1969; Lawley, 1956), si ha

$$G = n - q - 1 - \frac{2\tilde{p}^2 + \tilde{p} + 2}{6\tilde{p}} + \sum_{h=1}^{q} \frac{\bar{\gamma}^{(\tilde{p})}}{(\gamma_h - \bar{\gamma}^{(\tilde{p})})^2}.$$

Dalle elaborazioni prodotte (tabella 6), si è ipotizzato, almeno inizialmente, un modello fattoriale composto da 23 variabili latenti.

Tabella 6: *numero di fattori per il modello.*

metodi euristico-empirici			test di ipotesi		
Kaiser (figura 4b)	Analisi parallela (figure 4b e 4c)	Struttura Molto Semplice (figura 4d)	Anderson	Bartlett	Lawley
24	23 (≥ 7)	≥ 4	29	23	33

A tal proposito (figura 4b), assegnata una qualunque statistica di posizione LS_h, si osservi che

- il *Fattore di Accelerazione* (*FA*) rappresenta la soluzione numerica corrispondente al "gomito" del grafico degli autovalori, ovvero

$$n_{FA} \equiv \text{Se}\left[(\gamma_h > LS_h) \wedge \max(AF_h)\right];$$

- le *Coordinate Ottimali* (*CO*) corrispondono ad una estrapolazione del precedente autovalore da una linea di regressione che passa tra le coordinate dell'autovalore in esame e quelle dell'ultimo autovalore, ovvero

$$n_{OC} = \sum_{h=1}^{p} \left[(\gamma_h > LS_h) \cap (\gamma_h > \gamma_h^{(pred)})\right].$$

Per stimare i parametri incogniti del modello, sono stati comparati i risultati ottenuti ricorrendo a differenti metodi di stima; in particolare, si è scelto di utilizzare

- metodi che non richiedono alcuna assunzione distributiva, quali
  - *metodo delle componenti principali* (figura B.1a);

  - *metodo dei fattori principali* (figura B.1b);

- metodi di ricerca iterativa per minimizzare un particolare criterio, quali
  - *metodo dei minimi quadrati ponderati* (figura B.2a);
  - *metodo di stima della massima verosimiglianza* (figura B.2b).

La soluzione fattoriale ottenuta con i differenti metodi di stima è stata ruotata con il *criterio varimax* di rotazione ortogonale (Harman, 1976; Kaiser, 1958). Il confronto tra le soluzioni ha permesso di individuare, non soltanto il numero di fattori (pari a 2), tra loro non correlati, da includere nel modello fattoriale finale, ma anche gli *item* corrispondenti al fattore selezionato (tabella 7).

Tabella 7: *item comuni ai fattori per gli aspetti di comprensione.*

	PC1	PC3	PA1	PA15	WLS1	WLS17	ML5	ML10
C8	0,32							
B22	0,48		0,49			0,46		0,45
C3	0,57			0,40	0,41		0,42	
C7	0,40			0,32	0,34		0,35	
D2		0,56			0,30		0,31	
D5		0,61						
B18			0,58			0,56		0,56
B20			0,38			0,38		0,37
D1		0,57						
D4		0,49						
B16			0,44			0,44		0,44
C5	0,51							
C6	0,43			0,32	0,32		0,33	
C11	0,67			0,56	0,57		0,58	
C12	0,58			0,48	0,50		0,50	
D3		0,62						
C2	0,43							
B17			0,36			0,36		0,37
A6		0,37						

Per le su descritte ragioni, il modello fattoriale corrispondente agli aspetti sulla comprensione della lettura risulta essere il seguente:

$$\underbrace{\begin{bmatrix} \text{C3} \\ \text{C7} \\ \text{C6} \\ \text{C11} \\ \text{C12} \\ \text{B22} \\ \text{B18} \\ \text{B20} \\ \text{B16} \\ \text{B17} \end{bmatrix}}_{\boldsymbol{X}} = \underbrace{\begin{bmatrix} \lambda_{11} & 0 \\ \lambda_{21} & 0 \\ \lambda_{31} & 0 \\ \lambda_{41} & 0 \\ \lambda_{51} & 0 \\ 0 & \lambda_{62} \\ 0 & \lambda_{72} \\ 0 & \lambda_{82} \\ 0 & \lambda_{92} \\ 0 & \lambda_{102} \end{bmatrix}}_{\boldsymbol{\Lambda}} \underbrace{\begin{bmatrix} \xi_1 \\ \xi_2 \end{bmatrix}}_{\boldsymbol{\xi}} + \underbrace{\begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \\ \epsilon_5 \\ \epsilon_6 \\ \epsilon_7 \\ \epsilon_8 \\ \epsilon_9 \\ \epsilon_{10} \end{bmatrix}}_{\boldsymbol{\epsilon}} . \tag{1}$$

Dunque, l'*AFE* ha permesso di individuare 2 variabili latenti che sintetizzano, in maniera significativa, gruppi di *item* presenti nella sezione della prova *INVALSI* riguardante gli aspetti di comprensione della lettura.

D'altra parte, per quanto concerne l'analisi degli ambiti grammaticali, si è ipotizzato, innanzitutto, un numero di fattori pari al numero degli ambiti a cui appartengono le variabili osservate. Di conseguenza, si è ipotizzato, almeno inizialmente, un modello fattoriale composto da 6 variabili latenti.

Per stimare i parametri incogniti del modello, sono stati comparati i risultati ottenuti ricorrendo ai precedenti metodi di stima. In questo caso, però, soltanto il *metodo delle componenti principali* (figura B.3a) ed il *metodo dei minimi quadrati ponderati* (figura B.3b) hanno fornito "buone" soluzioni fattoriali.

Il confronto tra le soluzioni ottenute, ruotate ancora una volta con il *criterio varimax* (Harman, 1976; Kaiser, 1958), ha permesso di individuare 3 fattori, tra loro non correlati, da includere nel modello fattoriale finale.

Tale confronto, inoltre, ha contribuito a selezionare gli *item* corrispondenti al fattore selezionato (tabella 8).

Tabella 8: *item comuni ai fattori per gli ambiti grammaticali.*

	PC2	PC3	PC1	PC4	PC6	WLS2	WLS3	WLS1
E6_1	0,83					0,85		
E6_2	0,81					0,75		
E6_3	0,76					0,67		
E6_4	0,84					0,78		
E6_5	0,79					0,80		
E2				0,56				0,44
E3				0,72				0,46
E7_a					0,62			0,86
E7_b					0,59			0,75
E7_c			0,59					
E7_d			0,79					
E7_e			0,85					
E7_f								
E5				0,62				0,47
E1_a		0,93					0,90	
E1_b		0,88					0,88	
E1_c		0,85					0,85	
E9			0,44					0,38
E4				0,75				0,30
E8					0,71			0,42

Di conseguenza, il modello fattoriale corrispondente agli ambiti grammaticali risulta essere il seguente:

$$\underbrace{\begin{bmatrix} \text{E6\_1} \\ \text{E6\_2} \\ \text{E6\_3} \\ \text{E6\_4} \\ \text{E6\_5} \\ \text{E1\_a} \\ \text{E1\_b} \\ \text{E1\_c} \\ \text{E2} \\ \text{E3} \\ \text{E5} \\ \text{E4} \end{bmatrix}}_{\boldsymbol{X}} = \underbrace{\begin{bmatrix} \lambda_{11} & 0 & 0 \\ \lambda_{21} & 0 & 0 \\ \lambda_{31} & 0 & 0 \\ \lambda_{41} & 0 & 0 \\ \lambda_{51} & 0 & 0 \\ 0 & \lambda_{62} & 0 \\ 0 & \lambda_{72} & 0 \\ 0 & \lambda_{82} & 0 \\ 0 & 0 & \lambda_{93} \\ 0 & 0 & \lambda_{103} \\ 0 & 0 & \lambda_{113} \\ 0 & 0 & \lambda_{123} \end{bmatrix}}_{\boldsymbol{\Lambda}} \underbrace{\begin{bmatrix} \xi_1 \\ \xi_2 \\ \xi_3 \end{bmatrix}}_{\boldsymbol{\xi}} + \underbrace{\begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \\ \epsilon_5 \\ \epsilon_6 \\ \epsilon_7 \\ \epsilon_8 \\ \epsilon_9 \\ \epsilon_{10} \\ \epsilon_{11} \\ \epsilon_{12} \end{bmatrix}}_{\boldsymbol{\epsilon}} . \tag{2}$$

L'*AFE* ha permesso di individuare 3 variabili latenti che sintetizzano, in maniera significativa, gruppi di *item* presenti nella sezione della prova *INVALSI* riguardante gli ambiti grammaticali.

Si osservi come, nella costruzione dei modelli (1) e (2), si sia preferito considerare la matrice dei pesi fattoriali $\mathbf{\Lambda}$ ancora incognita. Tale matrice, infatti, verrà stimata, in modo più attendibile, nella successiva *AFC*.

5.3. Analisi fattoriale confermativa

L'*AFC* permette di valutare se il modello fattoriale ipotizzato può predire un insieme di dati osservati (DeCoster, 1998). Tale analisi, infatti, verifica la veridicità delle ipotesi sostenute (Ruscio & Roche, 2012; Schmitt, 2011), stabilisce la validità del modello fattoriale, confronta due modelli utilizzando gli stessi dati, testa la significatività del peso fattoriale, valuta la relazione tra i pesi fattoriali, misura l'eventuale correlazione tra i fattori e valuta la validità convergente e discriminante di misure (DeCoster, 1998).

Più in dettaglio, il pacchetto applicativo `lavaan` ha permesso di testare i modelli (1) e (2) ipotizzati, utilizzando un altro sottocampione di $2\,000$ unità, estratto dai $35\,618$ studenti rimanenti.

Dunque, i modelli fattoriali ottenuti dall'*AFC* risultano essere i seguenti:

- in riferimento agli aspetti sulla comprensione della lettura,

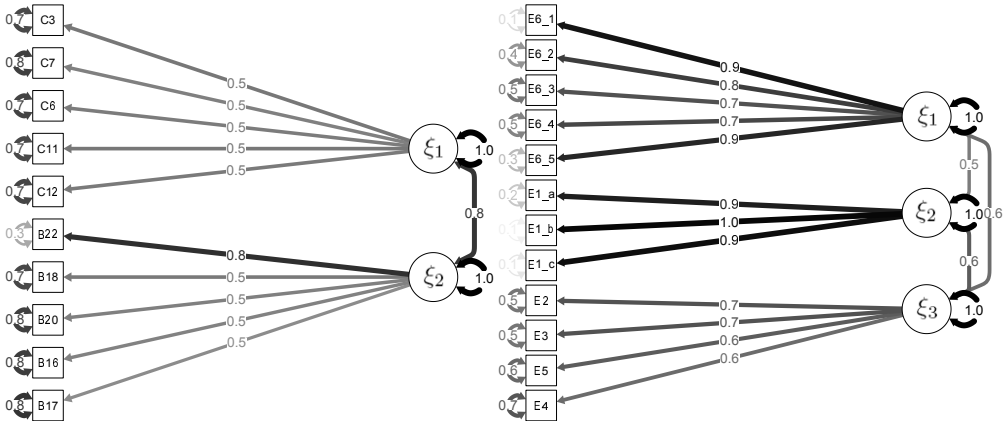
$$\underbrace{\begin{bmatrix} \text{C3} \\ \text{C7} \\ \text{C6} \\ \text{C11} \\ \text{C12} \\ \text{B22} \\ \text{B18} \\ \text{B20} \\ \text{B16} \\ \text{B17} \end{bmatrix}}_{\boldsymbol{X}} = \underbrace{\begin{bmatrix} 0{,}5 & 0 \\ 0{,}5 & 0 \\ 0{,}5 & 0 \\ 0{,}5 & 0 \\ 0{,}5 & 0 \\ 0 & 0{,}8 \\ 0 & 0{,}5 \\ 0 & 0{,}5 \\ 0 & 0{,}5 \\ 0 & 0{,}5 \end{bmatrix}}_{\mathbf{\Lambda}} \underbrace{\begin{bmatrix} \xi_1 \\ \xi_2 \end{bmatrix}}_{\boldsymbol{\xi}} + \underbrace{\begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \\ \epsilon_5 \\ \epsilon_6 \\ \epsilon_7 \\ \epsilon_8 \\ \epsilon_9 \\ \epsilon_{10} \end{bmatrix}}_{\boldsymbol{\epsilon}}, \quad \text{con} \quad \underbrace{\begin{bmatrix} 1{,}0 & 0{,}8 \\ 0{,}8 & 1{,}0 \end{bmatrix}}_{\mathbf{\Phi}}; \tag{3}$$

- in riferimento agli ambiti grammaticali,

$$\underbrace{\begin{bmatrix} \text{E6\_1} \\ \text{E6\_2} \\ \text{E6\_3} \\ \text{E6\_4} \\ \text{E6\_5} \\ \text{E1\_a} \\ \text{E1\_b} \\ \text{E1\_c} \\ \text{E2} \\ \text{E3} \\ \text{E5} \\ \text{E4} \end{bmatrix}}_{\boldsymbol{X}} = \underbrace{\begin{bmatrix} 0{,}9 & 0 & 0 \\ 0{,}8 & 0 & 0 \\ 0{,}7 & 0 & 0 \\ 0{,}7 & 0 & 0 \\ 0{,}9 & 0 & 0 \\ 0 & 0{,}9 & 0 \\ 0 & 1{,}0 & 0 \\ 0 & 0{,}9 & 0 \\ 0 & 0 & 0{,}7 \\ 0 & 0 & 0{,}7 \\ 0 & 0 & 0{,}6 \\ 0 & 0 & 0{,}6 \end{bmatrix}}_{\mathbf{\Lambda}} \underbrace{\begin{bmatrix} \xi_1 \\ \xi_2 \\ \xi_3 \end{bmatrix}}_{\boldsymbol{\xi}} + \underbrace{\begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \\ \epsilon_5 \\ \epsilon_6 \\ \epsilon_7 \\ \epsilon_8 \\ \epsilon_9 \\ \epsilon_{10} \\ \epsilon_{11} \\ \epsilon_{12} \end{bmatrix}}_{\boldsymbol{\epsilon}}, \quad \text{con} \quad \underbrace{\begin{bmatrix} 1{,}0 & 0{,}5 & 0{,}6 \\ 0{,}5 & 1{,}0 & 0{,}6 \\ 0{,}6 & 0{,}6 & 1{,}0 \end{bmatrix}}_{\mathbf{\Phi}}. \tag{4}$$

I diagrammi di percorso dei modelli (3) e (4) corrispondenti, rispettivamente, agli aspetti di comprensione della lettura (figura 5a) e agli ambiti grammaticali (figura 5b) sono stati realizzati, attraverso il pacchetto `semPlot`.

Figura 5: *modelli fattoriali.*
(a) *aspetti di comprensione della lettura.* **(b)** *ambiti grammaticali.*

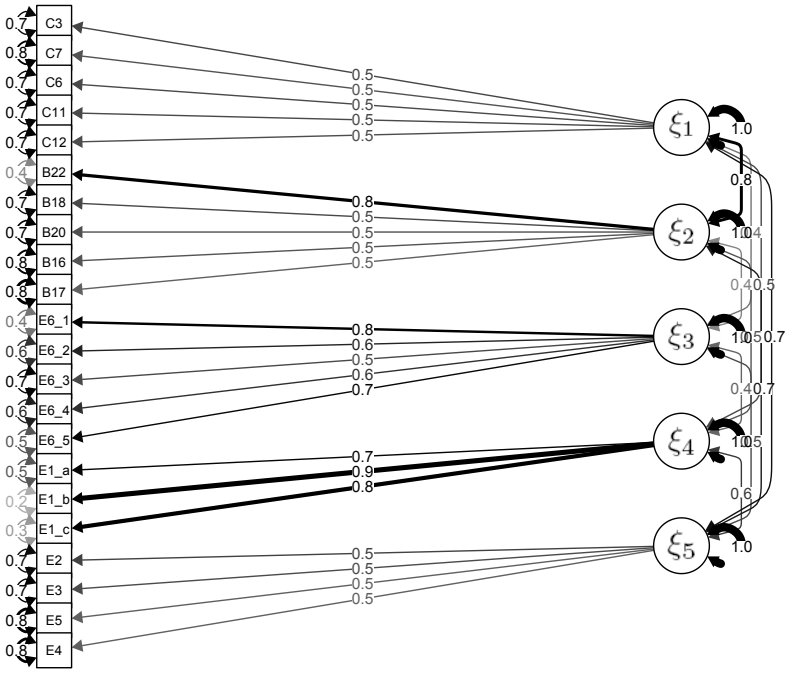


A questo punto, è apparso quasi naturale valutare una eventuale correlazione esistente tra i fattori del modello (3) e quelli del modello (4). Il modello fattoriale complessivamente ottenuto è risultato essere il seguente:

$$\underbrace{\begin{bmatrix} \text{C3} \\ \text{C7} \\ \text{C6} \\ \text{C11} \\ \text{C12} \\ \text{B22} \\ \text{B18} \\ \text{B20} \\ \text{B16} \\ \text{B17} \\ \text{E6\_1} \\ \text{E6\_2} \\ \text{E6\_3} \\ \text{E6\_4} \\ \text{E6\_5} \\ \text{E1\_a} \\ \text{E1\_b} \\ \text{E1\_c} \\ \text{E2} \\ \text{E3} \\ \text{E5} \\ \text{E4} \end{bmatrix}}_{\boldsymbol{X}} = \underbrace{\begin{bmatrix} 0{,}5 & 0 & 0 & 0 & 0 \\ 0{,}5 & 0 & 0 & 0 & 0 \\ 0{,}5 & 0 & 0 & 0 & 0 \\ 0{,}5 & 0 & 0 & 0 & 0 \\ 0{,}5 & 0 & 0 & 0 & 0 \\ 0 & 0{,}8 & 0 & 0 & 0 \\ 0 & 0{,}5 & 0 & 0 & 0 \\ 0 & 0{,}5 & 0 & 0 & 0 \\ 0 & 0{,}5 & 0 & 0 & 0 \\ 0 & 0{,}5 & 0 & 0 & 0 \\ 0 & 0 & 0{,}8 & 0 & 0 \\ 0 & 0 & 0{,}6 & 0 & 0 \\ 0 & 0 & 0{,}5 & 0 & 0 \\ 0 & 0 & 0{,}6 & 0 & 0 \\ 0 & 0 & 0{,}7 & 0 & 0 \\ 0 & 0 & 0 & 0{,}6 & 0 \\ 0 & 0 & 0 & 0{,}9 & 0 \\ 0 & 0 & 0 & 0{,}8 & 0 \\ 0 & 0 & 0 & 0 & 0{,}5 \\ 0 & 0 & 0 & 0 & 0{,}5 \\ 0 & 0 & 0 & 0 & 0{,}5 \\ 0 & 0 & 0 & 0 & 0{,}5 \end{bmatrix}}_{\boldsymbol{\Lambda}} \underbrace{\begin{bmatrix} \xi_1 \\ \xi_2 \\ \xi_3 \\ \xi_4 \\ \xi_5 \end{bmatrix}}_{\boldsymbol{\xi}} + \underbrace{\begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \\ \epsilon_5 \\ \epsilon_6 \\ \epsilon_7 \\ \epsilon_8 \\ \epsilon_9 \\ \epsilon_{10} \\ \epsilon_{11} \\ \epsilon_{12} \\ \epsilon_{13} \\ \epsilon_{14} \\ \epsilon_{15} \\ \epsilon_{16} \\ \epsilon_{17} \\ \epsilon_{18} \\ \epsilon_{19} \\ \epsilon_{20} \\ \epsilon_{21} \\ \epsilon_{22} \end{bmatrix}}_{\boldsymbol{\epsilon}}, \quad \text{con} \quad \underbrace{\begin{bmatrix} 1{,}0 & 0{,}8 & 0{,}4 & 0{,}5 & 0{,}7 \\ 0{,}8 & 1{,}0 & 0{,}4 & 0{,}5 & 0{,}7 \\ 0{,}4 & 0{,}4 & 1{,}0 & 0{,}4 & 0{,}5 \\ 0{,}5 & 0{,}5 & 0{,}4 & 1{,}0 & 0{,}6 \\ 0{,}7 & 0{,}7 & 0{,}5 & 0{,}6 & 1{,}0 \end{bmatrix}}_{\boldsymbol{\Phi}}. \tag{5}$$

Il corrispondente diagramma di percorso (figura 6) visualizza graficamente il modello (5).

Figura 6: *modello fattoriale complessivo.*



Per valutare la bontà di adattamento del modello, è stata scelta una combinazione di indici, rispettivamente, *assoluti*, *incrementali* e *parsimoniosi*, quali

- la *Radice dell'Errore Quadratico Medio di Approssimazione*, o in termini anglosassoni *Root Mean Square Error of Approximation* (*RMSEA*) (Brown, 2006; Steiger, 1990);
- la *Radice del Residuo quadratico Medio Standardizzato*, o *Standardised Root Mean square Residual (SRMR)* (Kline, 2005);
- l'*Indice di Adattamento Non Normato*, o *Non-Normed Fit Index* (*NNFI*), noto in letteratura anche come *indice di Tucker-Lewis* (Brown, 2006);
- l'*Indice di Adattamento Comparativo*, o *Comparative Fit Index* (*CFI*) (Brown, 2006);
- il *Criterio di Informazione di Akaike*, o in termini anglosassoni, *Akaike Information Criterion* (*AIC*) (Akaike, 1974).

Tale combinazione ha rivelato una "buona" corrispondenza dei modelli ottenuti ai dati osservati, ad eccezione dell'indice *AIC* (tabella 9).

Tabella 9: *indici di adattamento.*

	aspetti	ambiti	totale
RMSEA	0,03	0,09	0,02
SRMR	0,03	0,05	0,03
NNFI	0,97	0,93	0,98
CFI	0,98	0,95	0,98
AIC	12 201,71	17 008,71	37 163,65

Per comprendere la ragione di tale anomalia, si è deciso di approfondire l'analisi e di testare il modello (5), utilizzando una successione di campioni, estratti senza ripetizione, di numerosità crescente.

Tabella 10: *successione di indici di adattamento.*

n. studenti	50	100	200	500	1 000	2 000
RMSEA	0,09	0,04	0,03	0,03	0,03	0,02
SRMR	0,11	0,07	0,06	0,04	0,03	0,03
NNFI	0,78	0,92	0,95	0,97	0,97	0,98
CFI	0,81	0,93	0,96	0,97	0,98	0,98
AIC	919,19	1 967,39	3 867,70	9 089,96	18 485,84	37 163,65

Il risultato ottenuto (tabella 10) ha evidenziato che, all'aumentare della dimensione campionaria, il valore del *criterio di informazione di Akaike* aumenta in modo quasi esponenziale. Gli altri indici, però, convergono verso un valore ritenuto di "buon adattamento" del modello ai dati campionati.

6. Aspetti, ambiti e fattori

L'*Analisi Fattoriale* applicata ai dati *INVALSI* ha reso possibile l'interpretazione delle relazioni esistenti tra alcune variabili osservate (*item*), in termini di un più limitato insieme di variabili latenti (fattori).

Gli *item* proposti nella prova di Italiano appartenevano a macro categorie riguardanti particolari aspetti di comprensione della lettura o ambiti grammaticali, precedentemente individuati dall'*INVALSI*. Dunque, era ragionevole ipotizzare che ogni singolo aspetto o ambito corrispondesse ad uno specifico fattore individuato attraverso tale analisi.

Più in dettaglio (tabella 11), i fattori 3 e 4 corrispondono, rispettivamente, agli ambiti grammaticali 1 e 4. Tale corrispondenza permette di stabilire come gli *item* del gruppo E6 e quelli del gruppo E1 risultino scarsamente correlati tra loro (con un coefficiente pari a 0,4) e rispetto ai rimanenti *item* della prova, ma ben calibrati nel valutare il rispettivo ambito di appartenenza.

Tabella 11: *aspetti/ambiti e fattori.*

item	aspetti/ambiti	fattori	ξ_1	ξ_2	ξ_3	ξ_4	ξ_5
C3	2	ξ_1	1,0				
C7	2	ξ_1	1,0				
C6	4	ξ_1	1,0				
C11	4	ξ_1	1,0				
C12	5a	ξ_1	1,0				
B22	2	ξ_2	0,8	1,0			
B18	3	ξ_2	0,8	1,0			
B20	3	ξ_2	0,8	1,0			
B16	4	ξ_2	0,8	1,0			
B17	6	ξ_2	0,8	1,0			
E6_1	1	ξ_3	0,4	0,4	1,0		
E6_2	1	ξ_3	0,4	0,4	1,0		
E6_3	1	ξ_3	0,4	0,4	1,0		
E6_4	1	ξ_3	0,4	0,4	1,0		
E6_5	1	ξ_3	0,4	0,4	1,0		
E1_a	4	ξ_4	0,5	0,5	0,4	1,0	
E1_b	4	ξ_4	0,5	0,5	0,4	1,0	
E1_c	4	ξ_4	0,5	0,5	0,4	1,0	
E2	2	ξ_5	0,7	0,7	0,5	0,6	1,0
E3	2	ξ_5	0,7	0,7	0,5	0,6	1,0
E5	3	ξ_5	0,7	0,7	0,5	0,6	1,0
E4	6	ξ_5	0,7	0,7	0,5	0,6	1,0

D'altra parte, i fattori 1, 2 e 5 risultano essere altamente correlati tra loro (con un coefficiente pari, rispettivamente, a 0,8 e 0,7): non è un caso, infatti, che *item* riguardanti l'aspetto 2 sono presenti sia nel fattore 1 che nel fattore 2.

Si può ritenere, dunque, che il modello fattoriale (5) su aspetti sulla comprensione della lettura e ambiti grammaticali si dimostra sufficientemente adatto a valutare la prova *INVALSI* di Italiano delle classi II di una Scuola Secondaria di II grado.

7. Sintesi conclusiva

L'*Analisi Fattoriale* permette di individuare un ridotto insieme di variabili latenti (fattori), in grado di spiegare la covarianza (o correlazione) tra un più ampio insieme di variabili manifeste (variabili osservate). In altri termini, le variabili osservate e raccolte in una base di dati risultano essere combinazioni lineari di variabili non osservabili, note come fattori. In particolare, assegnata una matrice di covarianza (o correlazione) delle variabili osservate, l'obiettivo è quello di spiegare l'interdipendenza fra tali variabili, attraverso l'esistenza di fattori sottostanti.

Più in dettaglio, tale analisi può essere

1. *Analisi Fattoriale Esplorativa*, nel caso in cui non si abbia alcuna informazione sul numero di fattori, sulle loro caratteristiche, sui legami tra fattori e variabili;

2. *Analisi Fattoriale Confermativa*, nel caso in cui sia possibile tracciare un pri-

mo modello fattoriale e sottoporlo alla verifica dei dati, dal momento che si è ipotizzato conoscere:

- il numero di fattori sottostante;
- le relazioni tra i fattori;
- le relazioni tra fattori e variabili.

In letteratura, l'*Analisi Fattoriale* è utilizzata per valutare il grado di intelligenza (Spearman, 1904), la qualità della democrazia di un Paese (Bollen, 1980), l'ideologia di fondo di un partito politico (Gabel & Huber, 2000), l'affidabilità e la validità delle scale di misura (Carmines & Zeller, 1979).

Nel presente lavoro, tale analisi è stata applicata ai risultati ottenuti nella prova *INVALSI* di Italiano dalle classi II della Scuola Secondaria di II grado, con l'obiettivo di valutare la corrispondenza o meno dei singoli quesiti all'aspetto sulla comprensione della lettura o ambito grammaticale di appartenenza.

L'*Istituto Nazionale per la VAlutazione del Sistema educativo di Istruzione e di formazione* (*INVALSI*) ha, come finalità generale, la valutazione dell'efficacia e dell'efficienza del sistema scolastico, globalmente inteso, a livello nazionale e per singoli settori o singole istituzioni scolastiche. Il *Quadro di Riferimento* presenta le idee chiave che guidano la progettazione delle prove, per quanto riguarda:

- gli ambiti della valutazione, ovvero gli aspetti e la scelta degli argomenti;
- i modi della valutazione, ovvero le caratteristiche degli strumenti e i criteri seguiti nella costruzione delle prove.

Dunque, le prove *INVALSI* non hanno lo scopo di certificare il livello di competenza di uno studente, ma di compiere un'indagine statistica rispetto al complesso della popolazione scolastica.

Più in dettaglio, nell'analisi dei risultati *INVALSI* di Italiano delle classi II di una Scuola Secondaria di II grado, si è scelto di estrarre, senza ripetizione, una serie di sottocampioni, al fine di individuare, attraverso un'*Analisi Fattoriale Esplorativa*, i fattori che rappresentano la struttura portante dei dati in una forma sintetica, testare, attraverso una *Analisi Fattoriale Confermativa*, la generalità dei fattori estratti e valutare la validità del modello ottenuto.

Dunque, supposto che ogni singolo aspetto o ambito corrisponda ad uno specifico fattore, il modello fattoriale è risultato essere sufficientemente adatto ad interpretare gli esiti della prova di Italiano.

Riferimenti bibliografici

AA.VV. (2014). *Rilevazioni nazionali degli apprendimenti 2013-14*. Rapp. tecn. INVALSI.

Akaike, H. (1974). "A New Look at the Statistical Model Identification". In: *IEE Transactions on Automatic Control* **19**.6, pp. 716–723.

Anderson, T. W. (1963). "Asymptotic theory for principal component analysis". In: *Annals of Mathematical Statistics* **34**, pp. 122–148.

Bartlett, M. S. (1950). "Tests of significance in factor analysis". In: *British Journal of Psychology* **3**, pp. 77–85.

— (1951). "A further note on tests of significance". In: *British Journal of Psychology* **4**, pp. 1–2.

Bentler, P. M. & Yuan, K. H. (1996). "Test of linear trend in eigenvalues of a covariance matrix with application to data analysis". In: *British Journal of Mathematical and Statistical Psychology* **49**, pp. 299–312.

Bollen, K. A. (1980). "Issues in the comparative measurement of political democracy". In: *American Sociological Review* **45**, pp. 370–390.

Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. Guilford. New York.

Carmines, E. G. & Zeller, R. A. (1979). *Reliability and Validity Assessment*. Sage. Beverly Hills, California.

DeCoster, J. (1998). *Overview of Factor Analysis*.

Gabel, M. J. & Huber, J. D. (2000). "Putting parties in their place: Inferring party left-right ideological positions from party manifestos data". In: *American Journal of Political Science* **44**, pp. 94–103.

Harman, H. H. (1976). *Modern factor analysis*. Third edition. University of Chicago Press. Chicago.

Hayton, J. C., Allen, D. G. & Scarpello, V. (2004). "Factor Retention Decisions in Exploratory Factor Analysis: A Tutorial on Parallel Analysis". In: *Organizational Research Methods* **7**.2, pp. 191–205.

Horn, J. (1965). "A rationale and test for the number of factors in factor analysis". In: *Psychometrika* **30**.2, pp. 179–185.

Horn, J. L. & Engstrom, R. (1979). "Cattell's scree test in relation to Bartlett's chi-square test and other observations on the number of factors problem". In: *Multivariate Behavioral Reasearch* **14**.3, pp. 283–300.

James, A. T. (1969). "Test of equality of the latent roots of the covariance matrix". In: Krishna, P.K. *Multivariate analysis*. A cura di Academic Press. **2**. New York.

Kaiser, H. F. (1958). "The varimax criterion for analytic rotation in factor analysis". In: *Psychometrika* **23**, pp. 187–200.

— (1960). "The Application of Electronic Computers to Factor Analysis". In: *Educational and Psychological Measurement* **20**.1, pp. 141–151.

— (1970). "A second generation little jiffy". In: *Psychometrika* **35**.4, pp. 401–415.

Kline, R. B. (2005). *Principles and Practice of Structural Equation Modeling*. Second edition. The Guilford Press. New York.

Lawley, D. N. (1956). "Tests of significance for the latent roots of covariance and correlation matrix". In: *Biometrika* **43**.1/2, pp. 128–136.

MacCallum, R. C., Widaman, K. F., Preacher, K. J. & Hong, S. (2001). "Sample size in factor analysis: The role of model error". In: *Multivariate Behavioral Research* **36**.4, pp. 611–637.

MacCallum, R. C., Widaman, K. F., Zhang, S. & Hong, S. (1999). "Sample size in factor analysis". In: *Psychological Methods* **4**.1, pp. 84–99.

Olsson, U. (1979). "Maximum likelihood estimation of the polychoric correlation coefficient". In: *Psychometrika* **44**.4, pp. 443–460.

Piccolo, D. (1998). *Statistica*. Seconda edizione. Il Mulino. Bologna.

Quintano, C., Castellano, R. & Longobardi, S. (2009). "A fuzzy clustering approach to improve the accuracy of Italian student data. An experimental procedure to correct the impact of outliers on assessment test scores". In: *Statistica & Applicazioni* **7**.2, pp. 149–171.

R Development Core Team (2008). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria.

Revelle, W. & Rocklin, T. (1979). "Very simple structure - alternative procedure for estimating the optimal number of interpretable factors". In: *Multivariate Behavioral Research* **14**.4, pp. 403–414.

Ruscio, J. & Roche, B. (2012). "Determining the number of factors to retain in exploratory factor analysis using comparison data of known factorial structures". In: *Psychologocial Assessment* **24**.2, pp. 282–292.

Schmitt, T. A. (2011). "Current methodological considerations in exploratory and confirmatory factor analysis". In: *Journal of Psychoeducational Assessment* **29**.4, pp. 304–321.

Spearman, C. (1904). "General intelligence, objectively determined and measured". In: *American Journal of Psychology* **15**.2, pp. 201–293.

Steiger, J.H. (1990). "Structural model evaluation and modification". In: *Multivariate Behavioral Research* **25**.2, pp. 173–180.

Williams, M. T., Cahill, S. P. & B., Foa E. (2010). "Psychotherapy posttraumatic stress disorder". In: *Textbook of anxiety disorders*. A cura di D. J. Stein, E. Hollander & B. O. Rothbaum. Second edition. Washington, DC: American Psychiatric Publishing, Inc, pp. 603–626.

Appendice A

Definizione A.1 (Coefficiente di correlazione policorica)**.** *Si considerino due vettori ordinali* $\mathbf{o}_h$ *e* $\mathbf{o}_k$*, con* $h, k = 1, 2, \ldots, p$*, rispettivamente, con* m_h *e* m_k *categorie, la cui distribuzione marginale campionaria è rappresentata dalla seguente tabella di contingenza*

	$\mathbf{o}_k$				
$\mathbf{o}_h$	1	2	$\cdots$	m_k	
1	$n_{11}^{(hk)}$	$n_{12}^{(hk)}$	$\cdots$	$n_{1m_k}^{(hk)}$	$n_{1.}^{(hk)}$
2	$n_{21}^{(hk)}$	$n_{21}^{(hk)}$	$\cdots$	$n_{2m_k}^{(hk)}$	$n_{2.}^{(hk)}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
m_h	$n_{m_h1}^{(hk)}$	$n_{m_h2}^{(hk)}$	$\cdots$	$n_{m_hm_k}^{(hk)}$	$n_{m_h.}^{(hk)}$
	$n_{.1}^{(hk)}$	$n_{.2}^{(hk)}$	$\cdots$	$n_{.m_k}^{(hk)}$	$n^{(hk)}$

in cui $n_{cd}^{(hk)}$ è il numero di casi nelle categorie c e d, rispettivamente, sulle variabili $\mathbf{o}_h$ e $\mathbf{o}_k$.
Si ipotizzi che le variabili sottostanti X_h e X_k seguano una normale bivariata standard, con $r_{\mathbf{o}_h\mathbf{o}_k}^{(pol)}$ coefficiente di correlazione.
*Si denomina **coefficiente di correlazione policorica**, il coefficiente $r_{\mathbf{o}_h\mathbf{o}_k}^{(pol)}$ ottenuto massimizzando la funzione di log-verosimiglianza, indicata con $l(\cdot)$, della distribuzione multinomiale seguente:*

$$l\left[\pi_{cd}^{(hk)}\right] = \ln C + \sum_{c=1}^{m_h}\sum_{d=1}^{m_k} n_{cd}^{(hk)} \ln \pi_{cd}^{(hk)}, \tag{6}$$

in cui

- *C è una costante;*
- *$\pi_{cd}^{(hk)}$ è la probabilità che un'osservazione sia in (c, d), ovvero*

$$\pi_{cd}^{(hk)} = P(X_h = c,\ X_k = d) = \int_{\tau_{c-1}^{(h)}}^{\tau_c^{(h)}} \int_{\tau_{d-1}^{(k)}}^{\tau_d^{(k)}} f(y_1, y_2; r_{\mathbf{o}_h\mathbf{o}_k}^{(pol)}) dy_1 dy_2,$$

con $f(\cdot,\cdot; r_{\mathbf{o}_h\mathbf{o}_k}^{(pol)})$ densità normale bivariata standard seguente

$$f(y_1, y_2; r_{\mathbf{o}_h\mathbf{o}_k}^{(pol)}) = \frac{1}{2\pi\sqrt{1 - r_{\mathbf{o}_h\mathbf{o}_k}^{(pol)2}}} \exp\left[-\frac{y_1^2 - 2ry_1y_2 + y_2^2}{2(1 - r_{\mathbf{o}_h\mathbf{o}_k}^{(pol)2})}\right].$$

Definizione A.2 (Coefficiente di correlazione prodotto-momento)**.** *Assegnato un vettore $\mathbf{x}_h$, con $h = 1, 2, \ldots, p$, ovvero*

$\mathbf{x}_h$	x_{1h}	x_{2h}	$\cdots$	x_{nh}

*si denomina **coefficiente di correlazione prodotto-momento**, e si indica generalmente con $r_{\mathbf{x}_h\mathbf{x}_k}$, il seguente indice*

$$r_{\mathbf{x}_h\mathbf{x}_k} = \frac{s_{\mathbf{x}_h\mathbf{x}_k}}{s_{\mathbf{x}_h} s_{\mathbf{x}_k}}, \qquad h, k = 1, 2 \ldots, p, \tag{7}$$

posto

$$s_{\mathbf{x}_h\mathbf{x}_k} = \frac{1}{n-1}(\mathbf{x}_h - \mathbf{1}\bar{x}_h)^\top(\mathbf{x}_k - \mathbf{1}\bar{x}_k), \qquad s_{\mathbf{x}_h\mathbf{x}_h} = s_{\mathbf{x}_h}^2.$$

Definizione A.3 (Statistica chi quadrato generalizzata)**.** *Assegnati gli autovalori γ_h, con $h = 1, 2, \ldots, p$, si denomina **statistica chi quadrato generalizzata**, la seguente statistica*

$$-G \ln \left(\prod_{h=q+1}^{p} \frac{\gamma_h}{\bar{\gamma}^{(\tilde{p})}} \right),$$

posto

$$\bar{\gamma}^{(\tilde{p})} = \frac{1}{\tilde{p}} \sum_{h=q+1}^{p} \gamma_h, \qquad \text{con} \quad \tilde{p} = p - q,$$

in cui G è un parametro che dipende da

- n *dimensione campionaria;*
- p *numero di autovalori;*
- q *numero di autovalori da testare.*

Sotto l'ipotesi nulla H_0, tale statistica segue una χ^2, con $(\tilde{p}-1)(\tilde{p}+2)/2$ gradi di libertà, ovvero

$$-G \ln \left(\prod_{h=q+1}^{p} \frac{\gamma_h}{\bar{\gamma}^{(\tilde{p})}} \right) \underset{H_0}{\sim} \chi^2_{(\tilde{p}-1)(\tilde{p}+2)/2}.$$

Appendice B

Tabella B.1: *momenti campionari degli aspetti 1, 2, 3 e 4.*

	media	deviazione standard	mediana	deviazione assoluta mediana	asimmetria	curtosi	errore standard
A4	0,74	0,44	1	0	−1,08	−0,85	0,01
A7	0,77	0,42	1	0	−1,3	−0,31	0,01
A8	0,77	0,42	1	0	−1,3	−0,31	0,01
B2	0,72	0,45	1	0	−0,95	−1,1	0,01
B6	0,43	0,5	0	0	0,29	−1,92	0,02
B9	0,71	0,45	1	0	−0,92	−1,15	0,01
B10	0,6	0,49	1	0	−0,39	−1,85	0,02
B15	0,47	0,5	0	0	0,1	−1,99	0,02
B19	0,6	0,49	1	0	−0,41	−1,83	0,02
C8	0,57	0,5	1	0	−0,29	−1,92	0,02
C9	0,33	0,47	0	0	0,72	−1,49	0,01
C10	0,28	0,45	0	0	0,95	−1,1	0,01
B8	0,4	0,49	0	0	0,4	−1,84	0,02
B21	0,53	0,5	1	0	−0,13	−1,99	0,02
B22_a	0,58	0,49	1	0	−0,33	−1,89	0,02
B22_b	0,78	0,42	1	0	−1,32	−0,25	0,01
B22_c	0,75	0,43	1	0	−1,18	−0,61	0,01
B22_d	0,69	0,46	1	0	−0,81	−1,35	0,01
B22_e	0,63	0,48	1	0	−0,52	−1,73	0,02
B22_f	0,64	0,48	1	0	−0,61	−1,64	0,02
B22_g	0,69	0,46	1	0	−0,8	−1,36	0,01
C3	0,82	0,38	1	0	−1,68	0,83	0,01
C7_a	0,88	0,32	1	0	−2,39	3,74	0,01
C7_b	0,68	0,47	1	0	−0,75	−1,44	0,01
C7_c	0,58	0,49	1	0	−0,33	−1,89	0,02
C7_d	0,77	0,42	1	0	−1,3	−0,31	0,01
C7_e	0,75	0,43	1	0	−1,14	−0,7	0,01
C13	0,24	0,43	0	0	1,2	−0,55	0,01
D2	0,88	0,33	1	0	−2,26	3,13	0,01
D5	0,77	0,42	1	0	−1,29	−0,34	0,01
B18	0,84	0,37	1	0	−1,81	1,29	0,01
B20_a	0,75	0,43	1	0	−1,15	−0,67	0,01
B20_b	0,6	0,49	1	0	−0,42	−1,82	0,02
B20_c	0,89	0,32	1	0	−2,44	3,96	0,01
B20_d	0,77	0,42	1	0	−1,25	−0,43	0,01
B20_e	0,85	0,36	1	0	−1,94	1,75	0,01
C1	0,17	0,38	0	0	1,73	0,98	0,01
D1	0,75	0,43	1	0	−1,17	−0,63	0,01
D4	0,67	0,47	1	0	−0,72	−1,48	0,01
A5	0,93	0,26	1	0	−3,25	8,57	0,01
B3	0,58	0,49	1	0	−0,32	−1,9	0,02
B16	0,81	0,4	1	0	−1,55	0,39	0,01
C5	0,78	0,41	1	0	−1,35	−0,18	0,01
C6	0,55	0,5	1	0	−0,18	−1,97	0,02
C11	0,8	0,4	1	0	−1,54	0,36	0,01

Tabella B.2: *momenti campionari degli aspetti 5a, 5b, 6 e 7.*

	media	deviazione standard	mediana	deviazione assoluta mediana	asimmetria	curtosi	errore standard
A2	0,82	0,39	1	0	−1,62	0,63	0,01
B5	0,71	0,45	1	0	−0,94	−1,12	0,01
B11	0,59	0,49	1	0	−0,36	−1,87	0,02
B14	0,59	0,49	1	0	−0,37	−1,87	0,02
C4	0,63	0,48	1	0	−0,56	−1,69	0,02
C12_a	0,68	0,47	1	0	−0,76	−1,42	0,01
C12_b	0,75	0,44	1	0	−1,13	−0,73	0,01
C12_c	0,71	0,46	1	0	−0,9	−1,19	0,01
C12_d	0,56	0,5	1	0	−0,25	−1,94	0,02
C12_e	0,82	0,39	1	0	−1,64	0,68	0,01
C14	0,44	0,5	0	0	0,22	−1,95	0,02
D3	0,75	0,43	1	0	−1,15	−0,67	0,01
A3	0,49	0,5	0	0	0,04	−2	0,02
B1	0,85	0,36	1	0	−1,98	1,92	0,01
B7	0,5	0,5	1	0	0	−2	0,02
C2	0,58	0,49	1	0	−0,32	−1,9	0,02
A1	0,84	0,37	1	0	−1,8	1,25	0,01
B4	0,83	0,37	1	0	−1,78	1,18	0,01
B13	0,66	0,48	1	0	−0,66	−1,57	0,02
B17	0,7	0,46	1	0	−0,86	−1,27	0,01
B23	0,6	0,49	1	0	−0,43	−1,82	0,02
A6_a	0,94	0,24	1	0	−3,63	11,17	0,01
A6_b	0,98	0,15	1	0	−6,21	36,61	0
A6_c	0,96	0,18	1	0	−5,05	23,55	0,01
A6_d	0,91	0,28	1	0	−2,95	6,7	0,01
A6_e	0,96	0,19	1	0	−4,83	21,31	0,01
B12	0,18	0,39	0	0	1,64	0,68	0,01

Tabella B.3: *momenti campionari degli ambiti grammaticali.*

	media	deviazione standard	mediana	deviazione assoluta mediana	asimmetria	curtosi	errore standard
E6_1	0,66	0,47	1	0	−0,7	−1,51	0,01
E6_2	0,5	0,5	0,5	0,74	0	−2	0,02
E6_3	0,45	0,5	0	0	0,19	−1,97	0,02
E6_4	0,55	0,5	1	0	−0,2	−1,96	0,02
E6_5	0,58	0,49	1	0	−0,32	−1,9	0,02
E2	0,69	0,46	1	0	−0,81	−1,35	0,01
E3	0,79	0,41	1	0	−1,44	0,06	0,01
E7_a	0,73	0,44	1	0	−1,03	−0,94	0,01
E7_b	0,67	0,47	1	0	−0,72	−1,49	0,01
E7_c	0,62	0,49	1	0	−0,48	−1,77	0,02
E7_d	0,66	0,47	1	0	−0,69	−1,52	0,01
E7_e	0,77	0,42	1	0	−1,27	−0,38	0,01
E7_f	0,6	0,49	1	0	−0,42	−1,83	0,02
E5	0,82	0,39	1	0	−1,64	0,68	0,01
E1_a	0,45	0,5	0	0	0,22	−1,95	0,02
E1_b	0,55	0,5	1	0	−0,18	−1,97	0,02
E1_c	0,57	0,5	1	0	−0,27	−1,93	0,02
E9	0,27	0,45	0	0	1,01	−0,98	0,01
E4	0,71	0,46	1	0	−0,9	−1,19	0,01
E8	0,55	0,5	1	0	−0,18	−1,97	0,02

Tabella B.4: *momenti campionari degli aspetti di comprensione della lettura.*

	media	deviazione standard	mediana	deviazione assoluta mediana	asimmetria	curtosi	errore standard
A4	0,74	0,44	1	0	−1,08	−0,85	0,01
A7	0,77	0,42	1	0	−1,3	−0,31	0,01
A8	0,77	0,42	1	0	−1,3	−0,31	0,01
B2	0,72	0,45	1	0	−0,95	−1,1	0,01
B6	0,43	0,5	0	0	0,29	−1,92	0,02
B9	0,71	0,45	1	0	−0,92	−1,15	0,01
B10	0,6	0,49	1	0	−0,39	−1,85	0,02
B15	0,47	0,5	0	0	0,1	−1,99	0,02
B19	0,6	0,49	1	0	−0,41	−1,83	0,02
C8	0,57	0,5	1	0	−0,29	−1,92	0,02
C9	0,33	0,47	0	0	0,72	−1,49	0,01
C10	0,28	0,45	0	0	0,95	−1,1	0,01
B8	0,4	0,49	0	0	0,4	−1,84	0,02
B21	0,53	0,5	1	0	−0,13	−1,99	0,02
B22	0,68	0,34	0,71	0,42	−0,74	−0,79	0,01
C3	0,82	0,38	1	0	−1,68	0,83	0,01
C7	0,73	0,27	0,8	0,3	−0,98	0,43	0,01
C13	0,24	0,43	0	0	1,2	−0,55	0,01
D2	0,88	0,33	1	0	−2,26	3,13	0,01
D5	0,77	0,42	1	0	−1,29	−0,34	0,01
B18	0,84	0,37	1	0	−1,81	1,29	0,01
B20	0,77	0,23	0,8	0,3	−0,97	0,64	0,01
C1	0,17	0,38	0	0	1,73	0,98	0,01
D1	0,75	0,43	1	0	−1,17	−0,63	0,01
D4	0,67	0,47	1	0	−0,72	−1,48	0,01
A5	0,93	0,26	1	0	−3,25	8,57	0,01
B3	0,58	0,49	1	0	−0,32	−1,9	0,02
B16	0,81	0,4	1	0	−1,55	0,39	0,01
C5	0,78	0,41	1	0	−1,35	−0,18	0,01
C6	0,55	0,5	1	0	−0,18	−1,97	0,02
C11	0,8	0,4	1	0	−1,54	0,36	0,01
A2	0,82	0,39	1	0	−1,62	0,63	0,01
B5	0,71	0,45	1	0	−0,94	−1,12	0,01
B11	0,59	0,49	1	0	−0,36	−1,87	0,02
B14	0,59	0,49	1	0	−0,37	−1,87	0,02
C4	0,63	0,48	1	0	−0,56	−1,69	0,02
C12	0,7	0,29	0,8	0,3	−0,86	−0,13	0,01
C14	0,44	0,5	0	0	0,22	−1,95	0,02
D3	0,75	0,43	1	0	−1,15	−0,67	0,01
A3	0,49	0,5	0	0	0,04	−2	0,02
B1	0,85	0,36	1	0	−1,98	1,92	0,01
B7	0,5	0,5	1	0	0	−2	0,02
C2	0,58	0,49	1	0	−0,32	−1,9	0,02
A1	0,84	0,37	1	0	−1,8	1,25	0,01
B4	0,83	0,37	1	0	−1,78	1,18	0,01
B13	0,66	0,48	1	0	−0,66	−1,57	0,02
B17	0,7	0,46	1	0	−0,86	−1,27	0,01
B23	0,6	0,49	1	0	−0,43	−1,82	0,02
A6	0,95	0,13	1	0	−3,85	19,65	0
B12	0,18	0,39	0	0	1,64	0,68	0,01

Figura B.1: *stima dei parametri degli aspetti.*
(a) *componenti principali* **(b)** *fattori principali*

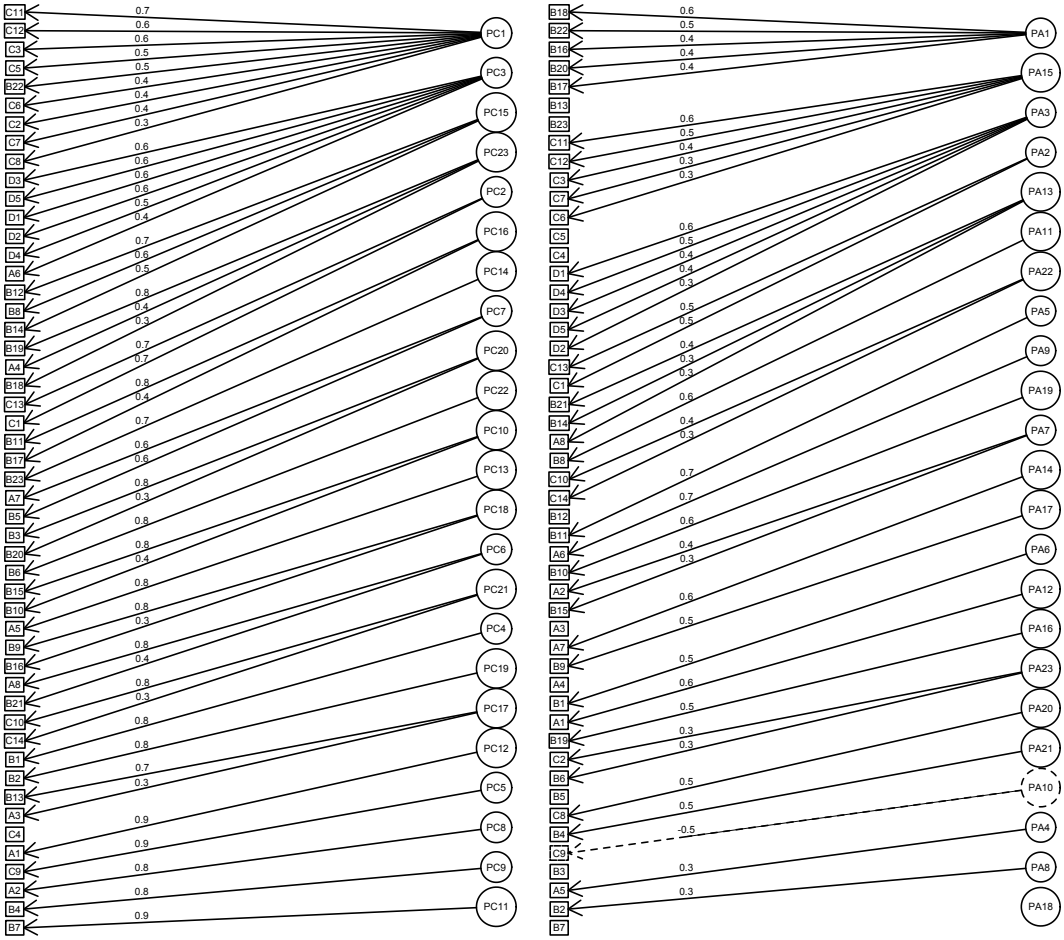


Figura B.2: *stima dei parametri degli aspetti.*
(a) *minimi quadrati ponderati* **(b)** *massima verosimiglianza*

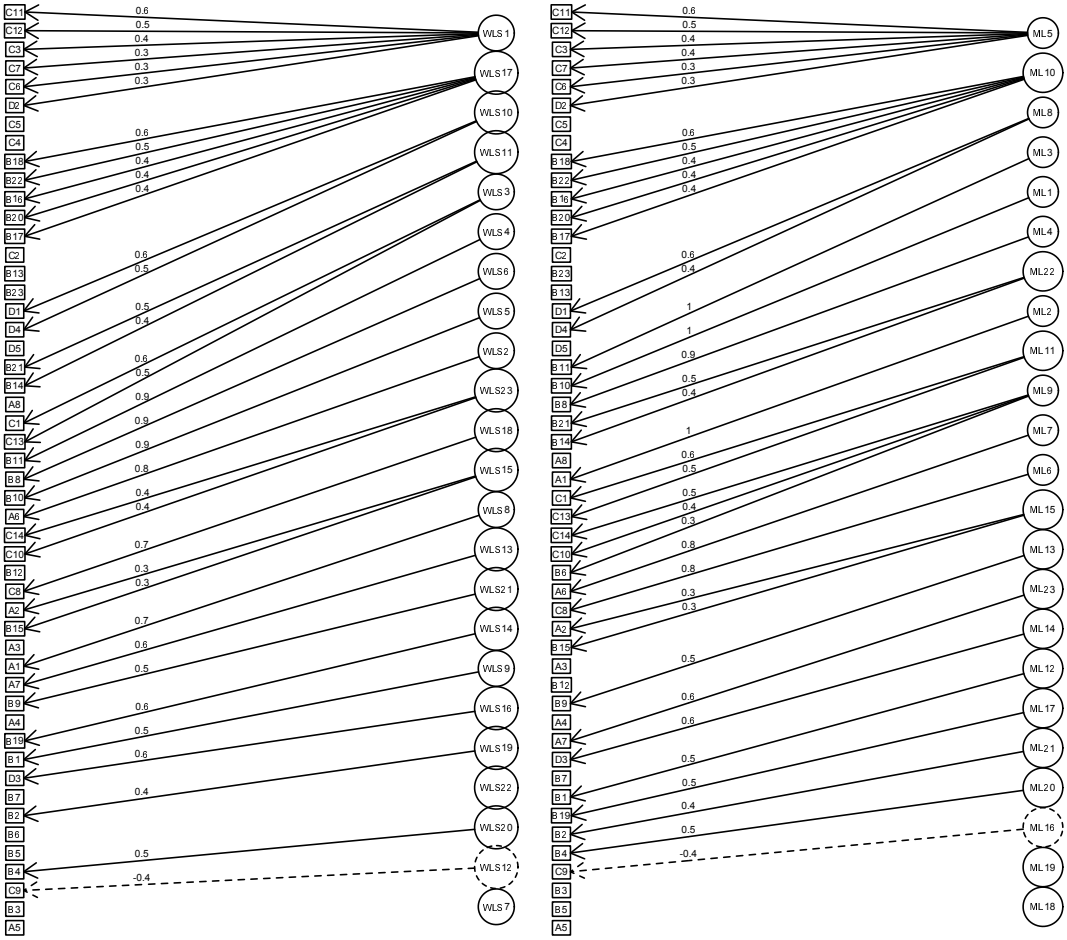
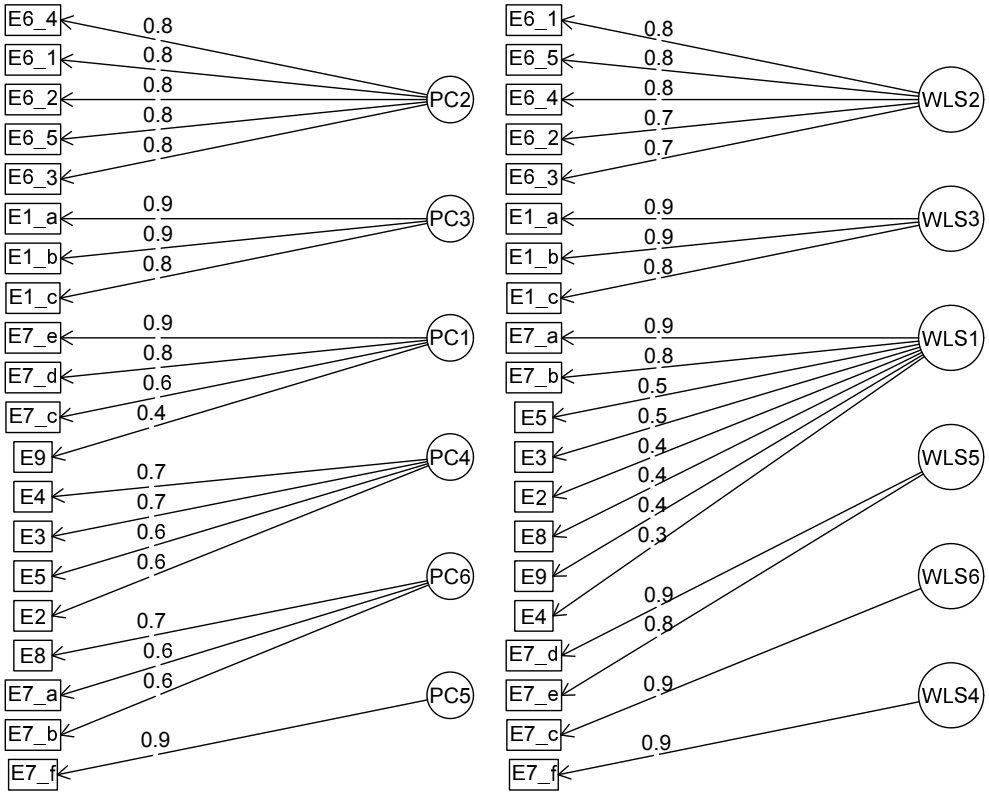


Figura B.3: *stima dei parametri degli ambiti.*

(a) *componenti principali* **(b)** *minimi quadrati ponderati*



UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO *Metodi e analisi statistiche 2017* ISBN: 978-88-6629-013-1
DIPARTIMENTO DI ECONOMIA E FINANZA pp. 317-336

Osservazioni e spunti sulla farmacoeconomia italiana

Francesco D. d'Ovidio*, Domenico Viola

Università degli studi di Bari Aldo Moro (Italy)

Riassunto: Presentazione fino a 20 righe di testo (meglio 18), senza capoversi né formule complesse né note a pie' di pagina; fino a 20 righe di testo (meglio 18), senza capoversi né formule complesse né note a pie' di pagina. fino a 20 righe di testo (meglio 18), senza capoversi né formule complesse né note a pie' di pagina. fino a 20 righe di testo (meglio 18), senza capoversi né formule complesse né note a pie' di pagina. Bla bla bla bla bla bla.

Keywords: da 3 a 6 elementi separati da punto e virgola; esempio; altro esempio.

1. Introduzione

L'Agenzia Italiana del Farmaco (AIFA) da alcuni anni ha intrapreso una fondamentale azione di raccolta ed elaborazione di dati, finalizzata alla loro diffusione e pubblicazione, per condividere anche le scelte necessarie affinché l'uso dei farmaci sia sempre più una cura efficace per la salute dei cittadini italiani.

Non è superfluo ricordare che una corretta pianificazione e programmazione degli interventi in ambito farmaceutico parte dai dati effettivi, rilevati con cura e precisione.

Il Rapporto Nazionale 2016 sull'uso del farmaci in Italia, edito quest'anno a cura dell'Osservatorio Nazionale sull'impiego dei Medicinali (Agenzia Italiana del Farmaco), ha appunto il merito di fornire questi dati e informazioni.

* Autore corrispondente: francescodomenico.dovidio@uniba.it

Il lavoro qui presentato è frutto di un progetto comune, ma D. Viola ha curato la redazione dei paragrafi 1-3, mentre F. D. d'Ovidio ha redatto i paragrafi 4-6.

Obiettivo del Rapporto è una dettagliata descrizione dell'utilizzazione dei medicinali a livello nazionale e regionale, nel corso del 2016. Lo scenario offerto deriva dalla lettura delle informazioni raccolte attraverso i diversi flussi informativi, consentendo la ricomposizione dei consumi e dell'assistenza farmaceutica in Italia. In particolare, il Rapporto analizza i dati relativi ai farmaci erogati in regime di assistenza convenzionata, presenta i dati relativi ai medicinali utilizzati dai pazienti a fronte della loro dispensazione in distribuzione diretta e per conto, o nel contesto specifico dell'assistenza ospedaliera.

Le analisi contenute nel Rapporto e approfondite in questo intervento sono da considerarsi in fase evolutiva e non ancora definitiva perché i dati sono "fermi" al 19 maggio 2017 (non tengono conto di tutte le revisioni che le ditte e le Regioni hanno chiesto di inviare al sistema NSIS per l'anno 2016).

Un approfondimento specifico è dedicato all'analisi dell'acquisto dei medicinali da parte delle strutture sanitarie pubbliche (ASL, Ospedali, Penitenziari, ecc.).

2. Quadro attuale e innovazioni

Sempre maggiore è il valore dei registri di monitoraggio AIFA, uno strumento riconosciuto e apprezzato a livello internazionale in quanto efficace nella generazione di evidenze nella fase post-marketing e di promozione dell'appropriatezza prescrittiva.

I registri e i piani terapeutici web-based hanno consentito di raccogliere i dati relativi a 1,2 milioni di trattamenti e a circa 1 milione di pazienti. Gli interventi riguardano maggiormente la popolazione anziana, generalmente meno rappresentata negli studi clinici..

Proviamo a capire meglio il quadro attuale dell'economia farmaceutica in Italia e le novità introdotte nel 2016.

- Nel 2016 la spesa farmaceutica ha fatto registrare un incremento per l'utilizzo dei nuovi farmaci ad azione antivirale diretta di seconda generazione (DAAs) per la cura dell'epatite C cronica e dei farmaci oncologici. Gli antineoplastici e gli antimicrobici si confermano infatti le prime due categorie in termini di spesa pubblica, (circa il 40% della spesa totale Servizio Sanitario Nazionale-SSN).
- Il 2016 è stato anche l'anno di grandi sforzi messi in atto per garantire l'accesso ai farmaci innovativi, per i quali in seguito (2017) sono stati istituiti due fondi, uno per i farmaci innovativi oncologici e l'altro per i farmaci inno-

vativi non oncologici, con uno stanziamento economico, ciascuno, di 500 milioni di euro. La Commissione Tecnico Scientifica dell'AIFA ha attribuito il carattere dell'innovatività a quattro farmaci, tre dei quali oncologici.

- Nel 2016 vi sono state numerose autorizzazioni di farmaci orfani (medicinali potenzialmente utili per trattare malattie rare con bassa frequenza nella popolazione meno di 1 abitante su 2000 casi): l'Agenzia Europea dei Medicinali (EMA) ha autorizzato ben 14 molecole con la qualifica di farmaco orfano; in Italia la spesa per i farmaci orfani, in crescita, rappresenta oltre il 6% della spesa SSN, e tra questi sono i farmaci antineoplastici quelli a maggior impatto.

3. La spesa farmaceutica in Italia

Nel 2016 la spesa farmaceutica nazionale totale (pubblica e privata) in Italia ha quasi raggiunto i 29,400 miliardi di euro, con un incremento di 1,6% rispetto all'anno precedente. Il 77,4% di questa somma è stata a carico del SSN.

Tabella 1. *Composizione della spesa farmaceutica nel 2016.*

	Spesa (milioni di €)	%	Δ% 16/15
Spesa convenzionata lorda (fascia A/SSN)	10.652,0	36,24	-1,9%
Distribuzione Diretta e per conto di fascia A/SSN	5.605,5	19,07	13,9%
Acquisto privato di fascia A/SSN	1.309,0	4,45	-11,9%
Classe C/SSN con ricetta	3.014,0	10,26	0,6%
Automedicazione (SOP e OTC)	2.322,0	7,90	-2,3%
ASL, A.O., RIA, penitenziari*	6.487,5	22,07	3,3%
Totale spesa farmaceutica	**29.390,0**	**100,00**	**1,6%**

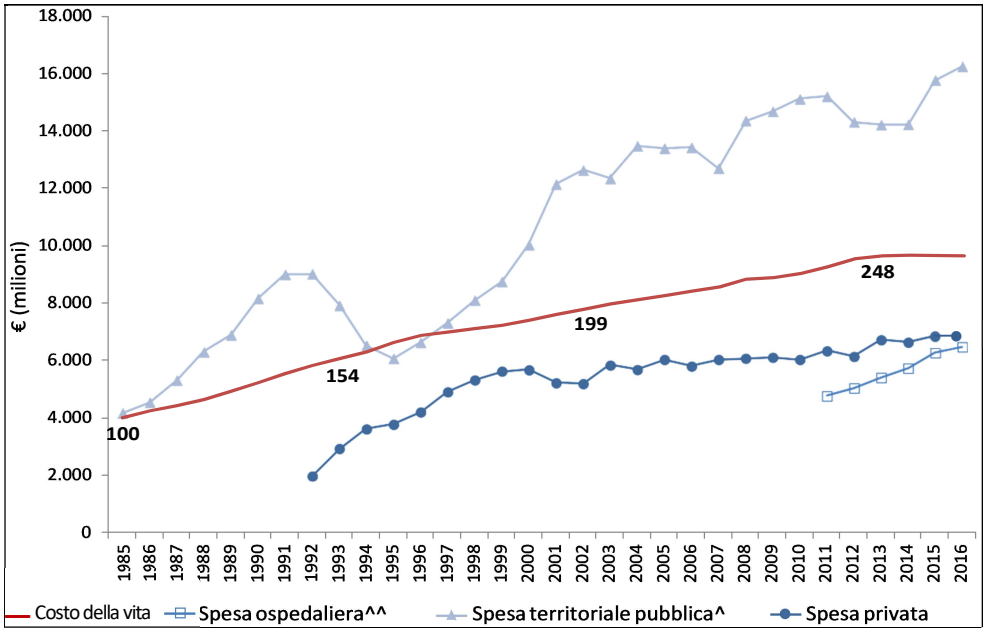
** al netto della spesa per i farmaci erogati in distribuzione diretta e per conto di fascia A*
Fonte: Osservatorio Nazionale sull'impiego dei Medicinali (2017)

La spesa complessiva, come dettagliato nella Tab. 1, è cosi distribuita:

- la voce più rappresentativa (36% del totale) è quella riguardante la spesa convenzionata lorda di fascia A/SSN, che ammonta a 10.652 milioni di euro e, rispetto all'anno precedente, ha subito una lieve riduzione, pari all'1,9%;
- la voce relativa alla distribuzione diretta e per conto di Fascia A/SSN è pari a 5.605 milioni di euro (19% del totale), con un consistente incremento rispetto all'anno precedente, pari al 13,9%;

- la voce Classe A/SSN privato, con 1.309 milioni di euro di spesa, pari al 4,5% della spesa complessiva, ha subito una riduzione di 11,9% rispetto all'anno precedente;
- passando ad esaminare la Classe C/SSN con ricetta, la spesa è di 3.014 milioni di euro (10,3% dell'intera spesa generata nel 2016), e appare quasi stabile nel confronto con il 2015 (appena 0,6% di incremento);
- La spesa per automedicazione (SOP, OTC presso Farmacie pubbliche e private) ammonta a 2.322 milioni di euro, quasi 8% della spesa totale, con una piccola riduzione (2,3%) rispetto al 2015;
- nelle strutture sanitarie pubbliche (ASL, Aziende Ospedaliere, Penitenziari ecc.) si spendono in farmaci quasi 6.500 milioni di euro (22% del totale), con un lieve incremento del 3,3% rispetto al 2015.

Figura 1. *Spesa farmaceutica e costo della vita (1985=100) in Italia, 1985–2016.*



Fonti: http://www.istat.it/it/archivio/30440; Osservatorio Nazionale sull'impiego dei Medicinali (2017)

Come si osserva in Fig. 1, la spesa farmaceutica nazionale nel trentennio passato è andata ciclicamente crescendo ben oltre il livello di crescita del costo della vita al lordo dell'inflazione, seguendo dunque l'andamento demografico (poiché l'Italia sta invecchiando, e gli anziani sono quasi sempre a rischio malattia e quindi più frequentemente dei giovani sotto terapia farmacologica). Fa parzialmente eccezione la spesa privata, che dal 1998 cresce con ritmo molto meno sostenuto.

Della spesa ospedaliera, nel breve periodo osservato (gli ultimi 6 anni), si nota la tendenza lineare nettamente crescente. Se tale spesa non fosse scorporata dalla spesa convenzionata, la crescita di questa sarebbe ancora maggiore di quella registrata, superando già nel 2012 il tetto dei 20 miliardi di euro.

Si può dunque arguire che, mentre negli anni 2000 la spesa farmaceutica a carico dei privati cittadini sia rimasta quasi stabile (a dispetto del sempre maggiore invecchiamento della popolazione e della migrazione di un consistente numero di farmaci dalla fascia A/SSN, convenzionata o gratuita, alla fascia C/SSN, del tutto a carico privato), la spesa farmaceutica a carico della sanità pubblica sia andata crescendo in misura insostenibile per il bilancio nazionale e regionale, a cui compete, come è noto, la maggior parte di tale spesa.

Sulle motivazioni di tale dinamica non è dato di sapere granché, pur se dalla politica vengono fornite molte e varie interpretazioni (non sempre rispondenti alla realtà dei fatti), a seconda se si tratti di partiti responsabili dell'amministrazione oppure partiti di opposizione. In questo studio, ancora sulla base dei dati raccolti e pubblicati dall'Osservatorio Nazionale sull'impiego dei Medicinali, si cercherà di definire meglio il problema.

Poiché la maggior parte della spesa farmaceutica pubblica attiene ai bilanci regionali, è sulle regioni che occorre focalizzare l'attenzione. Un dato fondamentale da conoscere è dunque quello relativo alla tendenza al consumo dei farmaci da parte dei residenti delle regioni italiane, ma per evitare l'effetto della dimensione demografica è opportuno analizzare non la spesa e il consumo grezzo, ma la spesa e il consumo pro-capite.

A tale scopo, però, è necessario tener conto della differente composizione per classe di età delle diverse regioni, e del differente peso di tali classi di età in termini di ricorso ai farmaci, e dunque standardizzare le popolazioni regionali in base a tali caratteristiche, in modo che una Regione con una popolazione più anziana della media nazionale (e dunque maggiormente consumatrice di farmaci) avrà una popolazione pesata maggiore di quella residente, e viceversa.

Il procedimento seguito per la standardizzazione delle popolazioni regionali, così come descritto in Appendice del rapporto dell'Osservatorio Nazionale sull'impiego dei Medicinali, ha seguito tre passi successivi:

a) innanzitutto, sulla base della distribuzione per età e sesso della spesa farmaceutica convenzionata, è stato calcolato un sistema di pesi differenziati appunto per fasce di età e per sesso (Tab. 2);
b) dal database http://demo.istat.it/ è stata determinata la numerosità della popolazione delle medesime fasce di età in ciascuna regione, distintamente per

maschi e femmine, che è stata poi moltiplicata per il corrispondente peso di "spesa farmaceutica";

c) la sommatoria dei valori così ottenuti a livello regionale, infine, è stata riproporzionata rispetto alla popolazione nazionale effettivamente rilevata (Tab. 3).

Tabella 2. *Pesi per fascia d'età e sesso in base alla spesa farmaceutica nel 2016.*

	Pesi		
Fascia d'età	**M**	**F**	**MF**
0	0,133	0,099	0,116
1 – 4	0,210	0,166	0,188
5 – 14	0,163	0,121	0,142
15 – 44	0,266	0,291	0,279
45 – 64	1,094	0,991	1,039
65 – 74	2,720	2,318	2,501
> 75	3,578	2,862	3,146

Fonte: Osservatorio Nazionale sull'impiego dei Medicinali (2017)

Tabella 3. *Popolazione residente ISTAT e popolazione standardizzata e pesata, 2016.*

Regione	**Popolazione residente 01/01/2016**	**Popolazione pesata 2016**
Piemonte	4.404.246	*4.733.549*
Valle d'Aosta	127.329	*130.958*
Lombardia	10.008.349	9.971.836
Prov. Auton. Bolzano	520.891	478.770
Prov. Auton. Trento	538.223	524.683
Veneto	4.915.123	*4.934.340*
Friuli Venezia Giulia	1.221.218	*1.331.907*
Liguria	1.571.053	*1.833.141*
Emilia Romagna	4.448.146	*4.640.456*
Toscana	3.744.398	*4.029.041*
Umbria	891.181	*953.924*
Marche	1.543.752	*1.627.837*
Lazio	5.888.472	5.746.528
Abruzzo	1.326.513	*1.363.694*
Molise	312.027	*327.245*
Campania	5.850.850	5.166.599
Puglia	4.077.166	3.930.846
Basilicata	573.694	*577.298*
Calabria	1.970.521	1.890.264
Sicilia	5.074.261	4.790.702
Sardegna	1.658.138	*1.681.930*
Italia	**60.665.551**	**60.665.551**

Fonti: http://demo.istat.it; elaborazioni Osservatorio Nazionale sull'impiego dei Medicinali (2017)

4. Distribuzione territoriale della spesa farmaceutica pro-capite

Le tabelle successive (strutturalmente abbastanza complesse) riportano per ogni regione o provincia autonoma alcuni fondamentali indicatori:

- la spesa farmaceutica annua pro-capite, calcolata sulla popolazione standardizzata per classe di età ed espressa in euro;
- il numero medio di dosi di farmaco (Defined Daily Dose[1]) consumate annualmente, in media giornaliera, da 1000 abitanti (anche questo calcolato su popolazione standardizzata), indicato come «DDD/*ab.die* × 1000».

I calcoli, rivenienti dall'Osservatorio Nazionale sull'impiego dei Medicinali (2017), sono basati anche sui dati NSIS per la Tracciabilità del Farmaco (DM 15 luglio 2004) per quanto riguarda le strutture sanitarie pubbliche, e sulle liste di trasparenza mensili dell'AIFA per quanto riguarda i farmaci a brevetto scaduto.

Come si osserva in Tab. 4, nel 2016 le regioni con maggiore spesa farmaceutica pro-capite di classe A/SSN erogata in regime di assistenza convenzionale (escludendo ossigeno e altri presidi non farmaceutici) sono risultate la Campania, con poco più di 219 euro e la Puglia con circa 214. Mostrano invece una spesa pro-capite ben più ridotta della media nazionale (175,3 euro pro-capite) la provincia autonoma di Bolzano, con 128,7 euro, e l'Emilia Romagna (132,2 euro).

Appare evidente, tra l'altro, che quasi tutte le regioni del Mezzogiorno, ad eccezione del Molise, presentano una spesa farmaceutica convenzionata pro-capite superiore alla media italiana, mentre quasi tutte le regioni poste fuori da tale area (fuorché Marche e Lazio) abbiano una spesa inferiore alla soglia.

Una divisione Nord-Sud altrettanto netta non si pone, invece, per la spesa farmaceutica di classe A/SSN delle strutture sanitarie, sempre pro-capite: qui, infatti, la Sicilia risulta avere una spesa inferiore alla media nazionale (che è poco sotto i 196 €), e ancora minori sono gli importi relativi ad Abruzzo e Molise, mentre sono maggiori della soglia le spese farmaceutiche pro-capite di Marche, Umbria, Emilia-Romagna e Toscana; quest'ultima regione, anzi, si trova al terzo posto in graduatoria con i suoi 231 euro pro-capite, subito dopo la Sardegna (233,2 €) e la capolista Campania (240,6 €). I territori ove la spesa farmaceutica annuale pro-capite nelle strutture sanitarie è minima sono la Valle d'Aosta (145,3 €) e la Provincia Autonoma di Trento (151,2 €).

[1] La *dose definita giornaliera* (DDD) esprime la dose media di mantenimento giornaliera di ciascun farmaco, tenendo conto della sua indicazione terapeutica principale nell'adulto. La DDD è generalmente assegnata ad un principio attivo già classificato dalla World Health Organization con uno specifico codice ATC, e rappresenta un'utile unità di misura nella parametrazione dei consumi in funzione delle diverse esigenze di monitoraggio (va sottolineato, comunque, che essa non riflette la dose media giornaliera prescritta).

Tabella 4. *Confronto regionale della spesa farmaceutica di classe A/SSN e privata (fascia C e automedicazione), in € pro-capite.*

Regioni / Prov. autonome	Spesa classe A/SSN	Spesa A/SSN strutture	Spesa classe C con ricetta	Spesa Automedicazione (SOP, OTC)	Spesa totale
Piemonte	151,71	172,2	49,1	37,8	410,8
Valle d'Aosta	142,02	145,3	*51,4*	***45,7***	**384,5**
Lombardia	173,33	173,5	*51,0*	*42,2*	440,0
P.A. Bolzano	**128,77**	177,1	**38,2**	***47,7***	391,7
P.A. Trento	145,88	**151,2**	42,9	*44,2*	**384,1**
Veneto	147,88	172,8	46,8	*40,6*	408,1
Friuli Ven. G.*	157,89	**164,5**	40,9	33,5	396,8
Liguria	148,89	183,5	***58,6***	*44,5*	435,5
Emilia Romagna	**132,40**	*201,6*	*51,7*	*39,8*	425,5
Toscana	142,69	***231,0***	***53,7***	*42,8*	*470,2*
Umbria	165,72	*199,0*	49,4	35,0	449,0
Marche*	*177,39*	*196,4*	*51,3*	35,8	*460,9*
Lazio	*198,70*	189,9	***53,2***	*42,2*	*483,9*
Abruzzo	*204,46*	181,3	43,6	31,9	*461,2*
Molise	167,02	175,2	**35,1**	**26,4**	403,7
Campania	***219,18***	***240,6***	*52,4*	37,4	***549,6***
Puglia	***214,10***	*230,0*	43,6	31,4	***519,1***
Basilicata	*179,19*	*213,1*	**37,9**	**26,0**	456,1
Calabria	*205,24*	*210,4*	47,8	31,8	*495,2*
Sicilia	*192,18*	190,4	49,4	32,5	*464,5*
Sardegna*	*188,87*	***233,2***	47,4	31,3	*500,8*
Italia	175,25	195,8	49,7	38,3	459,1

** Regioni senza ticket per ricetta nel 2016*
Fonte: Osservatorio Nazionale sull'impiego dei Medicinali (2017)

Ben differente (e molto più contenuta rispetto alla classe A/SSN) è la distribuzione regionale della spesa farmaceutica annua pro-capite a diretto carico dei cittadini, sia per i farmaci di classe C, con ricetta, e sia, ancor di più, per i farmaci di "automedicazione" senza ricetta. Per i primi, gli importi maggiori competono all'anziana Liguria (58,6 €), seguita a distanza da Toscana e Lazio (con 53,7 e 53,2 euro pro-capite rispettivamente) e quelli minori a Molise (circa 35 €), Basilicata e Provincia di Bolzano (entrambi con più o meno 38 €); la Campania è l'unica regione meridionale i cui abitanti spendano in media pro-capite più della media nazionale (pari a 49,7 €). Per l'automedicazione, invece, si nota una chiara inversione territoriale rispetto alla farmaceutica convenzionata classe A/SSN: addirittura, nessuna delle regioni del Mezzogiorno presenta una spesa pro-capite superiore alla media nazionale di 38,3 €, con i record negativi ancora di Basilicata (26 €) e Molise (26,4 €); il massimo livello di spesa pro-capite per auto-medicazione si nota, invece, nella Provincia Autonoma di Bolzano (47,7 €) e in Valle d'Aosta (45,7 €).

Nel complesso, la maggior spesa farmaceutica pro-capite annuale si registra in Campania (548 €), Puglia (518 €) e Sardegna (500,6 €), mentre quella minore (circa 384 € annui) compete alla Provincia Autonoma di Trento e alla Valle d'Aosta.

In base all'analisi descrittiva fin qui dettagliata, può facilmente ipotizzarsi una serie di relazioni tra le distribuzioni regionali delle diverse spese farmaceutiche, ma, non avendo motivo di ritenere che tali relazioni siano lineari, esse non saranno misurate con il noto coefficiente di correlazione lineare di Pearson, bensì col meno noto (ma più robusto) coefficiente di cograduazione ρ di Spearman, che può essere raffigurato come un indice di correlazione tra *ranghi*, ossia tra i posti che le varie regioni assumono nelle graduatorie relative alle varie fonti di spesa farmaceutica[2], e come il primo varia tra −1 (massima correlazione inversa dei dati) e 1 (massima correlazione diretta).

Le relazioni di rango maggiormente rilevanti sono quelle tra spesa convenzionata e spesa nelle strutture ($\rho = 0{,}48$) e tra spesa convenzionata e spesa per automedicazione ($\rho = -0{,}67$, dunque una forte correlazione inversa), mentre la spesa per farmaci di classe C risulta incorrelata con quella convenzionata ($\rho = -0{,}04$) ma direttamente connessa con la spesa per automedicazione ($\rho = 0{,}48$).

Ma la spesa farmaceutica pro-capite è solo *uno* degli aspetti che risultano molto differenziati a livello interregionale; in realtà, essa scaturisce (o dovrebbe scaturire) dal consumo di farmaci da parte delle diverse popolazioni, espresso nella Tab. 5 in termini di DDD/*ab.die*×1000. In detta tabella mancano, non essendo purtroppo ancora disponibili, i dati regionali relativi a farmaci non classificati in fascia A; di tali categorie si riporta solo, per comparazione di massima, la stima su base nazionale.

Per quanto riguarda i dati regionali sul consumo di farmaci di classe A/SSN, espresso in DDD/*ab.die*×1000, appare evidente in tabella che i dati relativi alle strutture sanitarie (fisiologicamente molto minori di quelli in convenzione, dato che il denominatore dell'indice è relativo all'intera popolazione e non a quella frazione di essa che è stata ospitata in strutture sanitarie nel 2016) risultano mediamente maggiori al Centro-Nord rispetto al Mezzogiorno[3], evidenziando in ciò un comportamento praticamente opposto a quello del consumo di farmaci in regime convenzionato ($\rho = -0{,}38$), ma anche una totale incorrelazione di rango con la spesa pro-capite nelle strutture medesime ($\rho = -0{,}09$). Il consumo regionale pro-dose annuale di farmaci convenzionati è invece coerente con la spesa pro-capite ($\rho = 0{,}78$), con valori massi-

[2] Cfr., ad es., Delvecchio 2015

[3] I consumi massimi si rilevano infatti in Emilia Romagna, con oltre 329 DDD/*ab.die*×1000, ma anche i valori di Veneto e Toscana sono molto più alti della media nazionale (166,2 DDD/*ab.die*×1000); d'altro canto, i consumi più ridotti si rilevano in Molise, Lombardia ed Abruzzo (102,4, 105,6 e 109,7 DDD/*ab.die*×1000).

mi in Puglia e Lazio (ambedue oltre 1263 DDD/*ab.die*×1000) e valori minimi nella provincia autonoma di Bolzano e in Valle d'Aosta (936 e 939 DDD/*ab.die*×1000), ben inferiori a quelli medi nazionali (1134,2 DDD/*ab.die*×1000).

Tabella 5. *Confronto regionale del consumo pro-capire di farmaci di classe A/SSN, in DDD/*ab.die*×1000. Stime differenziali di consumo privato (DDD/*ab.die*×1000) di farmaci di classe C e di automedicazione in Italia.*

Regioni / Prov. autonome	Farmaci A/SSN in convenzione	Farmaci A/SSN strutture	Totale farmaci erogati in classe A/SSN
Piemonte	1.042,2	*171,3*	1.213,4
Valle d'Aosta	**939,1**	*173,0*	1.112,0
Lombardia	1.072,3	**105,6**	1.177,8
Prov.Auton. Bolzano	**935,8**	*239,8*	1.175,6
Prov.Auton. Trento	1.102,0	161,7	1.263,6
Veneto	1.101,7	***263,9***	*1.365,6*
Friuli Venezia Giulia*	1.087,0	*178,8*	1.265,8
Liguria	950,6	*185,1*	**1.135,7**
Emilia Romagna	1.065,1	***329,3***	***1.394,4***
Toscana	1.086,2	***260,1***	*1.346,4*
Umbria	*1.191,2*	*187,6*	*1.378,8*
Marche*	1.131,3	158,6	1.289,8
Lazio	***1.263,1***	113,8	*1.376,8*
Abruzzo	*1.182,5*	**109,7**	1.292,1
Molise	1.061,2	**102,4**	**1.163,5**
Campania	*1.239,2*	115,9	*1.355,1*
Puglia	***1.263,5***	134,6	***1.398,1***
Basilicata	*1.077,8*	127,7	1.205,6
Calabria	*1.216,9*	121,3	*1.338,2*
Sicilia	*1.157,4*	119,1	1.276,5
Sardegna*	*1.219,6*	157,3	*1.376,8*
Italia	1.134,2	166,2	1.300,3
	Farmaci classe C con ricetta	*Farmaci per Automedicazione*	*Totale farmaci classe C e automedicazione*
Italia (stime)	*154,3*	*164,1*	*318,4*

** Regioni senza ticket per ricetta nel 2016*
Fonte: Osservatorio Nazionale sull'impiego dei Medicinali (2017)

Dalle osservazioni precedenti, si deduce facilmente che vi è un serio problema di standardizzazione della spesa farmaceutica di Classe A/SSN, ma non solo. Infatti, risulta obiettivamente preoccupante il fatto che le regioni ove, in termini pro-capite, si spende maggiormente in farmaci nelle strutture sanitarie siano quasi le medesime in cui si registra il minor consumo.

Tabella 6. *Confronto regionale del costo pro-DDD/*ab.die *dei farmaci di classe A/SSN e stime di costo pro-DDD/*ab.die *dei farmaci di classe C e di automedicazione in Italia.*

Regioni / Prov. autonome	Farmaci A/SSN in convenzione	Farmaci A/SSN strutture	Farmaci erogati in classe A/SSN
Piemonte	145,6	1.005,5	267,0
Valle d'Aosta	151,2	840,2	258,4
Lombardia	*161,6*	*1.643,9*	*294,5*
Prov.Auton. Bolzano	137,6	738,3	260,1
Prov.Auton. Trento	132,4	935,3	**235,1**
Veneto	134,2	**655,0**	**234,8**
Friuli Venezia Giulia*	145,3	919,9	254,7
Liguria	*156,6*	991,2	*292,7*
Emilia Romagna	**124,3**	**612,1**	**239,5**
Toscana	**131,4**	888,1	277,6
Umbria	139,1	1.060,7	264,5
Marche*	*156,8*	*1.238,8*	*289,8*
Lazio	*157,3*	*1.669,4*	282,2
Abruzzo	***172,9***	*1.652,8*	*298,5*
Molise	*157,4*	*1.711,4*	*294,1*
Campania	***176,9***	***2.076,1***	*339,3*
Puglia	*169,4*	*1.709,5*	*317,7*
Basilicata	*166,2*	*1.668,2*	*325,4*
Calabria	*168,7*	***1.734,4***	*310,6*
Sicilia	*166,1*	*1.598,7*	*299,7*
Sardegna*	*154,9*	*1.483,0*	*306,6*
Italia	154,5	1.178,6	285,4
	Farmaci classe C con ricetta	*Farmaci per Automedicazione*	*Farmaci classe C e automedicazione*
Italia (stime)	*322,0*	*233,2*	*276,2*

** Regioni senza ticket per ricetta nel 2016*
Fonte: Elaborazioni proprie su dati Osservatorio Nazionale sull'impiego dei Medicinali (2017)

La Tab. 6 rende esplicita l'incongruenza sopra sottolineata, dettagliando il costo annuale pro dose media a persona, calcolato come rapporto (opportunamente diviso per mille) tra i due indicatori *Spesa farmaceutica pro-capite* e *Consumo di farmaci in DDD/ab.die×1000*. La tabella riporta per confronto anche il costo medio nazionale pro-dose di farmaci in classe C/SSN e di automedicazione.

Detto che la spesa nazionale media si aggira intorno ai 280 € per DDD/*ab.die* (con poca differenza tra i farmaci di classe A/SSN e gli altri, ma con un divario ben più sostanzioso tra i farmaci di classe C con ricetta medica e i farmaci per autome-

dicazione), si osservi nel sottoinsieme dei primi l'enorme scompenso a livello nazionale: a fronte di appena 153,8€ pro-DDD/*ab.die* nel caso dei farmaci di classe A/SSN in convenzione (peraltro con una variabilità molto bassa, rilevandosi un $C.V. = 100 \cdot \hat{\sigma}/\mu = 9{,}5\%$), la media nazionale dei farmaci A/SSN erogati nelle strutture sanitarie balza a 1178,6 € pro-dose annuale, con *C.V.* = 33,6%.

Nel primo caso, risulta dunque poco rilevante la differenza tra le regioni con maggiore spesa pro-dose annuale (Campania ed Abruzzo, rispettivamente 175,6 € e 170,2 €) e quelle che spendono meno (Emilia Romagna e Toscana, con 124,1 € e 130,4 €); nel secondo, invece, il divario appare pressoché incolmabile: a fronte di un costo medio nazionale di 1.178,6 € pro-DDD, le strutture sanitarie in Campania fanno registrare 2.076 euro di costo per dose farmaceutica annuale (seguite a molta distanza da quelle calabresi, con 1.734 euro), e in tutto in Mezzogiorno, ma anche in Lazio e nelle Marche, i costi risultano sensibilmente maggiori della media; invece, nelle regioni centro-settentrionali (con l'eccezione della spendacciona Lombardia) la maggioranza dei costi pro-DDD è inferiore ai 1.000 € annuali, con il minimo assoluto in Emilia Romagna (612 euro) e in Veneto (655 euro).

Circa i motivi di questi divari, nulla trapela dai dati fin qui descritti, per cui possono essere avanzate ipotesi di qualsiasi tipo: dalla mancanza di un sistema di "costi standard ospedalieri", che potrebbe ricadere sui prezzi pagati dalle differenti ASL per i medesimi farmaci, a una improbabile (ma possibile) concentrazione nel Mezzogiorno di patologie richiedenti costosi farmaci di ultima generazione; oppure carenze organizzative o logistiche che possono indurre le farmacie interne delle strutture sanitarie ad effettuare approvvigionamenti enormemente sovradimensionati (e conseguentemente a doversi disfare, in seguito, di grandi quantità di farmaci inutilizzati a scadenza, facendo quindi registrare grandi spese e ridotti utilizzi); oppure, la carenza di registri di carico e scarico per il controllo accurato dei flussi farmaceutici nelle strutture sanitarie, con conseguente dispersione interna e forse (ipotesi che a volte trova qualche corrispondenza nella cronaca nera) distrazione di farmaci da parte di personale infedele o addirittura di persone "di passaggio" che possono approfittare di una sorveglianza insufficiente (e talora del tutto nulla) nei reparti sanitari e nelle medicherie.

Sull'ultima possibile fonte di costi inutili, purtroppo, può essere fatto poco, stante la tendenza al risparmio sulle "spese superflue" che le Pubbliche Amministrazioni tendono a interpretare riducendo i servizi di sorveglianza (come testimoniano molti casi di violenza, anche letale, a carico di personale medico in alcuni contesti sociali), mentre sulle altre sarebbe senz'altro utile una riorganizzazione dei processi, possibilmente più reale e meno burocratica di altre già effettuate.

5. I farmaci a brevetto scaduto e i loro "equivalenti generici"

Un altro aspetto che viene affrontato nel Rapporto Nazionale sull'uso dei farmaci in Italia è la diffusione dei farmaci non protetti da brevetto (o meglio, a brevetto scaduto), sia prodotti ancora dalle aziende che erano state titolari del relativo brevetto, sia prodotti da altre aziende con diverso nome commerciale ma identica funzione terapeutica, e in genere la medesima composizione chimica: i cosiddetti farmaci equivalenti (inizialmente detti "generici", termine che però è a sua volta generico)

Tabella 7. *Confronto regionale della spesa farmaceutica di classe A/SSN, per titolarità di brevetto e sua vigenza, in € pro-capite.*

Regioni / Prov. autonome	Farmaci classe A/SNN a brevetto scaduto			Farmaci classe A/SSN con brevetto vigente
	Farmaci titolari a brevetto scaduto	Farmaci equivalenti (generici) °	Totale farmaci a brevetto scaduto	
Piemonte	84,6	*6,6*	91,1	319,5
Valle d'Aosta	**75,2**	6,3	**81,5**	302,7
Lombardia	87,7	***8,4***	96,1	343,7
Prov. Auton. Bolzano	**72,4**	5,5	**77,9**	313,7
Prov. Auton. Trento	81,2	***9,2***	90,4	**293,3**
Veneto	82,8	*6,8*	89,6	318,1
Friuli Venezia Giulia*	85,6	*7,4*	93,1	303,4
Liguria	84,7	6,0	90,7	344,6
Emilia Romagna	86,0	6,5	92,5	332,7
Toscana	88,7	5,8	94,4	*374,7*
Umbria	*101,8*	*6,6*	*108,4*	340,6
Marche*	*103,5*	5,7	*109,2*	351,7
Lazio	*111,9*	6,0	*117,9*	*365,4*
Abruzzo	*105,0*	*6,6*	*111,6*	346,5
Molise	97,8	5,0	102,9	**299,7**
Campania	***123,5***	5,5	***129,0***	***418,9***
Puglia	***119,1***	6,0	***125,1***	***393,0***
Basilicata	*105,5*	**4,3**	*109,8*	345,8
Calabria	*116,5*	**4,7**	*121,2*	*373,2*
Sicilia	*108,9*	5,5	*114,4*	346,8
Sardegna*	*104,8*	5,7	*110,5*	*390,1*
Italia	98,3	6,5	104,8	353,4

° Dato stimato dalla % sul totale del dato per i farmaci a brevetto scaduto. Si intendono farmaci equivalenti i medicinali a base di principi attivi con brevetto scaduto, ad esclusione di quelli che hanno goduto di copertura brevettuale, ai sensi dell'art.1bis, del DL 27 maggio 2005, n. 87, convertito, con modificazioni, dalla Legge 26 luglio 2005, n. 149.

** Regioni senza ticket per ricetta nel 2016*

Fonte: Elaborazioni proprie su dati dell'Osservatorio Nazionale sull'impiego dei Medicinali (2017)

Concentrando l'attenzione sui soli farmaci di classe A/SSN, che comunque rappresentano la maggior parte del consumo e della spesa farmaceutica, la Tab. 7 pone chiaramente in luce che in Italia, a fronte di una spesa pro-capite per farmaci con brevetto vigente pari a oltre 353 euro, la spesa di farmaci con brevetto scaduto è pari a meno di 105 € pro-capite. La variabilità di entrambe le serie territoriali non è alta, in termini di coefficiente di variazione: C.V. 9,6% per la prima, C.V. 13,4% per la seconda; anche i rispettivi massimi e minimi di spesa sono poco rilevanti, ed i primi localizzati in Campania e in Puglia, mentre i minimi livelli sono meno coerenti (Provincia Autonoma di Trento e Molise per quanto riguarda i farmaci da brevetto, e invece Provincia Autonoma di Bolzano e Valle d'Aosta per quelli a brevetto scaduto). Più interessante, nonostante i quasi risibili livelli di spesa pro-capite (6,5 € pro-capite), appare la distribuzione dei farmaci equivalenti, i cui massimi competono alla Provincia Autonoma di Trento e alla Lombardia (rispettivamente 9,2 e 8,4 € pro-capite), mentre i minimi a due regioni del profondo Sud: Basilicata e Calabria, ciascuna con circa 4,5 € pro-capite. Interessante è anche il fatto che tale inversione rispetto ai farmaci "originali" riguarda la maggior parte delle regioni: il Mezzogiorno, che spende di più del Nord per i farmaci di marca, spende invece di meno per gli equivalenti; la relazione territoriale inversa è ben confermata dal coefficiente di correlazione di Spearman: $\rho = -0{,}41$.

Ora, si potrebbe ritenere che in Italia si spende di meno per i farmaci a brevetto scaduto in quanto sono pochi o poco noti e dunque i loro consumi sono inferiori.

Questa ipotesi, come dimostra la Tab. 8, è del tutto priva di fondamento: infatti, in Italia i consumi di farmaci con brevetto vigente (di cui molti sono farmaci innovativi o salvavita, talora di costo tale da renderne impensabile l'acquisto privato) ammontano ad appena 377,4 DDD/*ab.die*×1000, mentre il consumo di farmaci a brevetto scaduto è poco meno di tre volte tanto: circa 924 DDD/*ab.die*×1000; addirittura, i farmaci equivalenti, la cui spesa annua pro-capite è di un cinquantesimo (1/50) di quella dei farmaci a brevetto vigente, fanno registrare un consumo pari al 55% del consumo di questi ultimi.

La distribuzione territoriale del consumo di farmaci, naturalmente, è molto articolato, pur se la sua variabilità relativa non è eccessiva: il massimo (C.V. 24,7%) viene rilevato proprio per i farmaci equivalenti. Per quanto riguarda i farmaci con brevetto vigente, colpisce che i valori più cospicui si osservino per il Veneto e la Sardegna, mentre i consumi standardizzati minori competano alla Sicilia e, con valori quasi identici, alla Valle d'Aosta e alla Provincia Autonoma di Trento. Questo territorio è anche quello che consuma più dosi annuali di farmaci equivalenti, seguito dall'Emilia Romagna, mentre i minori consumatori risiedono in Basilicata e Molise.

Tabella 8. *Confronto regionale dei consumi farmaceutici di classe A/SSN, per titolarità di brevetto e sua vigenza, in DDD/ab.die×1000.*

Regioni / Prov. autonome	Farmaci classe A/SNN a brevetto scaduto			Farmaci classe A/SSN con brevetto vigente
	Farmaci titolari a brevetto scaduto	Farmaci equivalenti (generici) °	Totale farmaci a brevetto scaduto	
Piemonte	629,7	*238,9*	868,6	345,1
Valle d'Aosta	**566,7**	*218,2*	**784,9**	**327,2**
Lombardia	**589,1**	*252,5*	841,5	339,1
Prov. Auton. Bolzano	560,0	186,7	746,6	*429,1*
Prov. Auton. Trento	620,3	***316,7***	*937,0*	**327,3**
Veneto	657,6	*212,2*	869,8	***496,5***
Friuli Venezia Giulia*	652,5	*246,2*	898,7	367,3
Liguria	595,7	198,6	**794,3**	341,9
Emilia Romagna	*734,4*	***268,9***	*1003,3*	*391,3*
Toscana	672,4	*239,9*	912,3	*434,5*
Umbria	*802,3*	*223,6*	***1025,9***	354,1
Marche*	*746,4*	170,5	916,9	375,3
Lazio	***844,5***	187,9	***1032,4***	346,1
Abruzzo	*746,8*	177,5	*924,3*	368,1
Molise	703,6	**128,1**	831,7	332,4
Campania	*830,5*	154,7	*985,2*	370,6
Puglia	*819,3*	156,1	*975,4*	*423,1*
Basilicata	*733,0*	**118,3**	851,3	355,0
Calabria	***835,2***	132,6	*967,8*	371,2
Sicilia	*788,1*	162,6	*950,7*	**326,0**
Sardegna*	*747,4*	193,9	*941,3*	***435,7***
Italia	716,0	207,9	923,9	377,4

° *Dato stimato dalla % sul totale del dato per i farmaci a brevetto scaduto.*
* *Regioni senza ticket per ricetta nel 2016*
Fonte: Elaborazioni proprie su dati dell'Osservatorio Nazionale sull'impiego dei Medicinali (2017)

Va detto che non si rilevano correlazioni degne di nota tra le serie territoriali descritte in Tab. 8, fuorché una correlazione (evidentemente inversa) tra il consumo di farmaci equivalenti e farmaci a brevetto scaduto prodotti dagli originari titolari del brevetto: $\rho = -0{,}51$.

A questo punto, sorge spontanea la curiosità di confrontare i costi annui per dose giornaliera dei farmaci a brevetto scaduto con quelli dei farmaci con brevetto vigente. La Tab. 9 ha appunto questa finalità, ed evidenzia che la dose definita giornaliera di questi ultimi costa ogni anno 936,5 € a testa, mentre per i farmaci a brevetto scaduto il costo annuale scende a 113,4 € pro-capite.

Tabella 9. *Confronto regionale del costo pro-DDD/ab.die dei farmaci di classe A/SSN, per titolarità di brevetto e sua vigenza.*

Regioni / Prov. autonome	Farmaci classe A/SNN a brevetto scaduto			Farmaci classe A/SSN con brevetto vigente
	Farmaci titolari a brevetto scaduto	Farmaci equivalenti (generici) °	Totale farmaci a brevetto scaduto	
Piemonte	134,3	27,5	104,9	925,7
Valle d'Aosta	132,7	28,8	103,8	925,1
Lombardia	***148,9***	*33,1*	*114,2*	*1013,6*
Prov. Auton. Bolzano	129,2	29,6	104,3	**731,1**
Prov. Auton. Trento	130,8	29,1	**96,5**	896,1
Veneto	**125,9**	*32,1*	103,0	**640,7**
Friuli Venezia Giulia*	131,3	30,2	103,6	826,1
Liguria	*142,2*	30,1	114,2	*1007,8*
Emilia Romagna	**117,1**	**24,1**	**92,2**	850,4
Toscana	131,9	**24,0**	103,5	862,3
Umbria	**126,9**	29,6	105,7	*961,9*
Marche*	*138,7*	*33,3*	*119,1*	*937,2*
Lazio	132,5	*32,0*	*114,2*	***1055,7***
Abruzzo	*140,6*	*37,1*	*120,7*	*941,3*
Molise	*139,0*	***39,3***	*123,7*	901,8
Campania	***148,7***	*35,9*	***131,0***	***1130,4***
Puglia	*145,4*	***38,5***	*128,3*	928,9
Basilicata	*143,9*	*36,2*	***128,9***	*974,2*
Calabria	*139,5*	*35,7*	*125,2*	*1005,5*
Sicilia	*138,2*	*33,8*	*120,4*	***1063,8***
Sardegna*	*140,2*	29,6	*117,4*	895,3
Italia	137,3	31,3	113,4	936,5

° *Dato stimato dalla % sul totale del dato per i farmaci a brevetto scaduto*
* *Regioni senza ticket per ricetta nel 2016*
Fonte: Elaborazioni proprie su dati dell'Osservatorio Nazionale sull'impiego dei Medicinali (2017)

Dunque, il motivo dell'eccessiva spesa farmaceutica italiana sembra essere il costo (spesso giustificato dalla necessità di coprire le spese di R&S) dei farmaci innovativi e salvavita. Peraltro, il massimo costo annuale, ben sopra i 1.000 € a testa, compete alle regioni Campania e Sicilia, tallonate dal Lazio, mentre il Veneto e la Provincia Autonoma di Bolzano spendono rispettivamente 640,7 € e 731 € per abitante. E certo… ma di questi farmaci se ne consumano relativamente pochi!

Invece i farmaci a brevetto scaduto sono molto utilizzati, e ancora una volta il loro costo massimo pro-dose annua compete alla Campania (131 € per abitante) seguita a breve distanza da Basilicata e Puglia. Il minimo costo per dose annua pro-

capite è invece osservabile in riferimento all'Emilia Romagna (92,2 €) e alla Provincia Autonoma di Trento (96,5 €).

Restando nell'ambito dei farmaci a brevetto scaduto, va sottolineato che il costo per dose definita giornaliera dei farmaci equivalenti è sempre molto più basso rispetto alla concorrenza «qualificata» (31,3 € a persona contro 137,3 €); tuttavia, anche questo costo nel Mezzogiorno risulta sensibilmente maggiore che nella maggioranza dei territori del Centro-Nord: massimo in Molise e Puglia, con circa 39 €, minimo in Toscana ed Emilia Romagna (circa 24 €). Oggettivamente, questo incremento sistematico di costi per dose nei territori di per sé meno ricchi appare un chiaro ostacolo alla riduzione della spesa farmaceutica.

Si pone, tuttavia, uno spunto di riflessione: secondo i dati elaborati dall'Osser–vatorio Nazionale sull'impiego dei Medicinali (2017) la percentuale di farmaci equivalenti consumati, sul totale di quelli a brevetto scaduto, varia da meno del 14% (Calabria e Basilicata) al 30-34% (Lombardia e Provincia Autonoma di Bolzano). In Puglia, tale quota si attesta intorno al 16%.

Quali sarebbero i risparmi del Sistema Sanitario Nazionale se il consumo di farmaci equivalenti di fascia A fosse uniformemente portato ai livelli di Lombardia e provincia di Bolzano? O magari, dato che questi livelli massimi sono tali persino in assenza di una consistente promozione da parte del SSN, anche qualche punto percentuale in più?

Ipotizzando che, con attive campagne di informazione (anche sui media più seguiti, con formula «Pubblicità Progresso»), tale percentuale possa arrivare al livello, non utopistico, del 40% sul totale dei farmaci a brevetto scaduto consumati annualmente, è facilmente stimabile il risparmio per il SSN e i cittadini, senza ripercussioni problematiche sulle quote di mercato delle aziende farmaceutiche titolari, sempre meritevoli di stima per le proprie attività di R&S. Detta stima, dettagliata per territorio di competenza della spesa, è dettagliata in Tab. 10.

Ebbene, anche in assenza di complessi interventi sull'organizza–zione interna delle strutture sanitarie (che si è appurato essere la massima fonte di spesa per DDD) o sul sistema distributivo, che fa sì che in alcuni territori (ad es., nella maggior parte delle regioni del Mezzogiorno) persino gli economici farmaci equivalenti costino, in termini di DDD/*ab.die*, dal 10% al 40% in più rispetto ad altri territori consimili, risulta evidente che su scala nazionale si avrebbe un risparmio di circa 17 € annui per abitante (3,7%): oltre un miliardo di euro all'anno.

Su scala locale, ovviamente, il risparmio sarebbe molto variabile, a seconda della situazione attuale: si va dal 5,3% di Basilicata e Calabria all'1,5% della provincia di Trento, che parte però dalla situazione migliore.

Tabella 10. *Confronto regionale del risparmio stimato sulla spesa totale per farmaci di classe A/SSN, in € pro-capite (ipotesi di quota uniforme del 40% dei farmaci equivalenti sul totale dei farmaci a brevetto scaduto consumati in un anno)*

Regioni / Province autonome	Spesa attuale classe A/SSN	Stima spesa classe A/SSN	Risparmio stimato	% risparmio
Piemonte	410,6	399,0	11,6	2,8%
Valle d'Aosta	384,2	374,2	10,0	2,6%
Lombardia	439,8	430,1	9,7	2,2%
Provincia Auton. Bolzano	391,6	380,5	11,2	2,8%
Provincia Auton. Trento	383,7	377,8	5,9	1,5%
Veneto	407,8	395,0	12,7	3,1%
Friuli Venezia Giulia	396,5	385,1	11,4	2,9%
Liguria	435,2	421,9	13,3	3,1%
Emilia Romagna	425,2	412,9	12,3	2,9%
Toscana	469,1	455,6	13,5	2,9%
Umbria	449,0	430,8	18,2	4,0%
Marche	460,9	440,2	20,7	4,5%
Lazio	483,3	460,6	22,6	4,7%
Abruzzo	458,1	438,2	19,9	4,3%
Molise	402,6	382,2	20,4	5,1%
Campania	548,0	521,0	27,0	4,9%
Puglia	518,1	493,1	25,0	4,8%
Basilicata	455,6	431,7	23,9	5,3%
Calabria	494,5	468,0	26,4	5,3%
Sicilia	461,2	438,5	22,7	4,9%
Sardegna	500,6	480,4	20,2	4,0%
Italia	458,2	441,1	17,1	3,7%

Fonte: Elaborazioni proprie su dati dell'Osservatorio Nazionale sull'impiego dei Medicinali (2017)

Forse il Ministero della Sanità, che spende annualmente migliaia di miliardi di euro, potrebbe considerare questo risparmio un elemento risibile: tuttavia si tratta di un risparmio ottenibile con un investimento minimo, e che verosimilmente non creerebbe problemi di mercato per le grandi aziende farmaceutiche, ossia quelle che possono investire denaro nelle attività di ricerca e che dunque hanno diritto ai ritorni garantiti dalla protezione brevettuale, finché essa agisce. Un risparmio che però implicherebbe un rilancio delle piccole aziende che ora sopravvivono con le briciole del mercato farmaceutico. E implicherebbe una boccata di ossigeno per le casse esauste di varie ASL territoriali (in Campania, quasi 140 milioni di euro; 130 milioni in Lazio, 109 in Sicilia e quasi 100 in Puglia).

Se poi la quota di farmaci equivalenti arrivasse al 60% del totale dei farmaci a brevetto scaduto, si otterrebbero dei risultati ancor più interessanti, come dettaglia-

to in Tab. 11: 8% di risparmi in Italia, quasi 37 € all'anno per abitante, cioè oltre 2 miliardi e 200 milioni di euro. Anche nelle realtà locali, ovviamente, il risparmio sarebbe molto cospicuo, persino nelle aree già ora in situazione migliore: anche la Provincia di Trento spenderebbe 13 milioni di euro in meno ogni anno, e per le regioni meno "fortunate" si sfiorerebbe o si supererebbe i 200 milioni di euro (250 milioni in Campania e Lazio).

Tabella 11. *Confronto regionale del risparmio stimato sulla spesa totale per farmaci di classe A/SSN, in € pro-capite (ipotesi di quota uniforme del 40% dei farmaci equivalenti sul totale dei farmaci a brevetto scaduto consumati in un anno)*

Regioni / Province autonome	Spesa attuale classe A/SSN	Stima spesa classe A/SSN	Risparmio	% risparmio
Piemonte	410,6	380,4	30,2	7,3%
Valle d'Aosta	384,2	357,9	26,3	6,8%
Lombardia	439,8	410,6	29,2	6,6%
Provincia Auton. Bolzano	391,6	365,6	26,0	6,6%
Provincia Auton. Trento	383,7	358,7	25,0	6,5%
Veneto	407,8	378,7	29,1	7,1%
Friuli Venezia Giulia	396,5	366,9	29,6	7,5%
Liguria	435,2	404,1	31,1	7,2%
Emilia Romagna	425,2	394,2	31,0	7,3%
Toscana	469,1	435,9	33,2	7,1%
Umbria	449,0	410,9	38,1	8,5%
Marche	460,9	420,9	40,0	8,7%
Lazio	483,3	439,9	43,4	9,0%
Abruzzo	458,1	419,0	39,0	8,5%
Molise	402,6	365,6	37,0	9,2%
Campania	548,0	498,7	49,2	9,0%
Puglia	518,1	472,3	45,9	8,9%
Basilicata	455,6	413,3	42,3	9,3%
Calabria	494,5	447,9	46,5	9,4%
Sicilia	461,2	418,6	42,6	9,2%
Sardegna	500,6	459,6	41,0	8,2%
Italia	458,2	421,5	36,7	8,0%

Fonte: Elaborazioni proprie su dati dell'Osservatorio Nazionale sull'impiego dei Medicinali (2017)

Non occorre fare altre ipotesi più ottimistiche: invero, una quota del 100% del totale, che porterebbe a un risparmio teorico di oltre il 16% su scala nazionale, risulta un tetto oggettivamente irraggiungibile, e peraltro potrebbe causare turbative di mercato, in quanto monopolio di una parte degli operatori, seppur frammentata come l'insieme delle piccole aziende farmaceutiche che operano in Italia e nel mondo.

6. Osservazioni conclusive

Le osservazioni condotte nel presente lavoro hanno condotto a determinare una serie di situazioni regionali anomale, soprattutto in termini di spesa farmaceutica e di costo per dose annua.

Gli interventi necessari per la soluzione dei problemi economico-sanitari delle regioni italiane sono di diversa natura, e di differente impatto e problematicità:

- innanzitutto, vanno risolti i problemi di mercato o distributivi che determinano il maggior costo unitario di qualsiasi farmaco, compresi i farmaci equivalenti;
- inoltre, occorre investigare sull'appropriatezza organizzativa delle strutture sanitarie pubbliche, non solo del Mezzogiorno, allo scopo di uniformare il costo unitario dei farmaci ivi somministrati e/o distribuiti (si tenga conto che in Emilia Romagna e in Veneto il costo per DDD/ab. die è compreso tra 600 e 700 euro, ossia da metà a 1/3 del costo nelle similari strutture del Mezzogiorno);
- ma anche (intervento molto meno complesso dei precedenti) occorre intervenire per incrementare l'informazione dei cittadini e la responsabilità etica dei medici di base, superando alcune ovvie resistenze del sistema distributivo, al fine di aumentare il ricorso ai farmaci equivalenti a tutt'oggi disponibili.

A tal proposito, oltretutto, si sottolinea che molte delle officine farmaceutiche che producono "farmaci equivalenti" hanno sede principale (o unica) in Italia, a non poche ma proprio nel Mezzogiorno, su cui si impernia non solo la loro attività produttiva, ma anche il loro carico fiscale. Ogni azione mirata ad aumentare il peso di tali farmaci nel sistema sanitario migliorerebbe certo il bilancio del SSN e delle Regioni, ma potrebbe migliorare persino la bilancia dei pagamenti nazionale.

L'argomento risulta dunque di interesse non solo del SSN e delle amministrazioni regionali, ma anche dei Ministeri economici e produttivi.

Riferimenti bibliografici

Osservatorio Nazionale sull'impiego dei Medicinali (2017*). L'uso del farmaci in Italia - Rapporto Nazionale, anno 2016*, Agenzia Italiana del Farmaco, Roma.

Delvecchio, F. (2015). *Statistica per l'analisi dei fenomeni sociali*. Cleup, Padova.

Sitografia

http://www.istat.it/it/archivio/30440

UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO
DIPARTIMENTO DI ECONOMIA E FINANZA

PDF finito di comporre
il 13 dicembre 2017

ISBN 978-88-6629-013-1