

DIPARTIMENTO DI  
ECONOMIA E FINANZA

**METODI E ANALISI  
STATISTICHE  
2021**



UNIVERSITÀ  
DEGLI STUDI DI BARI  
ALDO MORO



**DIPARTIMENTO DI  
ECONOMIA E FINANZA**

**METODI E ANALISI  
STATISTICHE  
2021**



Tutti i diritti di traduzione, riproduzione e adattamento, totale o parziale, con qualsiasi mezzo (comprese le copie fotostatiche e i microfilm) sono riservati

Toma E., (a cura di) edizione 2021. *Metodi e Analisi Statistiche*, Dipartimento di Economia e Finanza, Università degli studi di Bari *Aldo Moro*.

© Copyright 2021 by Università degli Studi di Bari Aldo Moro  
[www.uniba.it](http://www.uniba.it)

Prima edizione: dicembre 2021

ISBN 978-88-6629-066-7

Gli articoli qui presentati sono stati tutti oggetto di valutazione interna.

*Editing finale:* F.D. d'Ovidio, E. Toma

# Sommario

|                                                                                                                                                                                                |        |
|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------|
| Ernesto Toma<br><b><i>Presentazione</i></b> .....                                                                                                                                              | pag. 5 |
| Massimo Bilancia, Fabio Manca, Gianvito Pio<br><b><i>A Brief Review on Compositional Inference of Microbiome Data</i></b> .....                                                                | « 7    |
| Claudia Marin, Fabio Manca<br><b><i>Application of text mining in the biomedical field</i></b> .....                                                                                           | « 21   |
| Ilaria Pepe, Ernesto Toma<br><b><i>Alcune considerazioni sulla presenza in Italia dei Minori Stranieri Non Accompagnati (MSNA) in relazione alla pandemia Covid-19</i></b> .....               | « 35   |
| Aurélien P. Scarafino, Antonella Massari, Claudia Marin<br><b><i>Big Data, Intelligenza Artificiale e turismo sostenibile</i></b> .....                                                        | « 59   |
| Nunziata Ribecco, Angela Maria D’Uggento, Arianna Ruggiero<br><b><i>L’esperienza della didattica a distanza in alcune scuole pugliesi</i></b> .....                                            | « 71   |
| Carlo Cusatelli, Crescenzo Gallo, Alessandro Rinaldi<br><b><i>Distributive adaption of the Lorenz curve to a Squirecle</i></b> .....                                                           | « 87   |
| Angela Maria D’Uggento, Marina Tarantini<br><b><i>L’approccio modellistico per una programmazione più efficace dei processi amministrativi: il caso della mortalità da Covid-19</i></b> .....  | « 101  |
| Andrea Tanga, Angela Maria D’Uggento, Danilo Di Bona<br><b><i>Efficacia del Benraluzimab nel trattamento dell’asma severo eosinofilo. Identificazione di subfenotipi di risposta</i></b> ..... | « 125  |
| Agata Maria Madia Carucci, Giovanni Vannella<br><b><i>La valutazione delle azioni di sostenibilità ambientale delle imprese attraverso i censimenti permanenti</i></b> .....                   | « 135  |

|                                                                                   |       |
|-----------------------------------------------------------------------------------|-------|
| Sabrina Diomede, Giovanni Taglialatela                                            |       |
| <i>A generalization of the logistic function</i> .....                            | « 155 |
| Anna Lucia Muserra, Francesco Grimaldi                                            |       |
| <i>Crisi finanziaria e politiche di bilancio. Alcune evidenze empiriche</i> ..... | « 167 |

# Presentazione

Il volume *Metodi e analisi statistiche*, nel 2021, giunge alla sua sesta edizione, confermandosi un'importante occasione per una elicitazione del lavoro di ricerca svolto da ricercatori di ambito prevalentemente statistico, spesso in sinergia con colleghi, universitari e non, di altri ambiti scientifici.

Essendo gli argomenti trattati nel presente volume estremamente eterogenei, potrebbe non essere semplice trovare un filo conduttore che possa permettere di raccontare la realizzazione finale dello stesso. Nonostante la presenza di taluni contributi di carattere più metodologico, ciò che accomuna la stragrande maggioranza degli articoli è sicuramente l'utilizzo del dato statistico, che risulta al centro di analisi sociali, sanitarie, economiche, aziendali.

Il primo contributo, di Bilancia & al., pone il focus su due modelli di studio del microbiota batterico dell'intestino umano, che notoriamente svolge un ruolo fondamentale nella patogenesi di molte malattie, promuovendone l'infiammazione. I modelli di Holmes et al., del 2012, e di Anderlucci & Viroli, del 2020, sono qui messi a confronto e diventano una base per possibili future ricerche.

Marin e Manca si occupano invece dell'importanza del text mining nelle sue applicazioni in ambito clinico. Un'analisi "intelligente" delle varie informazioni biomediche, consente infatti di scoprire delle informazioni non evidenti ad una prima lettura dei testi clinici, ottenendo come obiettivo finale il miglioramento degli stessi.

Nel terzo contributo, Pepe e Toma affrontano un fenomeno di estrema attualità, la presenza dei Minori Stranieri Non accompagnati (MSNA), partendo dai dati mensili dei Report statistici forniti dal Ministero del lavoro e delle politiche sociali. L'obiettivo è quello di analizzare il fenomeno dell'insediamento in Italia degli MSNA in relazione al periodo pandemico scoppiato a marzo 2020. Di tale fenomeno, in particolare, viene studiata l'evoluzione rispetto anche ad alcune caratteristiche come il genere, l'età, la nazione di provenienza e la regione di accoglienza dei minori.

Scarafino & al., propongono l'utilizzo di *Big Data*, *Intelligenza Artificiale*, *Realtà aumentata* e *Realtà virtuale* nell'ambito del settore turistico, al fine di ottenere un tempestivo monitoraggio delle destinazioni, delle performance e dei fruitori delle stesse.

Nel quinto contributo, di Ribecco & al., si cambia totalmente ambito di ricerca passando all'esperienza della didattica a distanza (DAD) in alcune scuole medie superiori di Bari e provincia. Gli autori, attraverso un'indagine che ha visto la partecipazione di quasi 1.600 studenti, hanno indagato alcuni aspetti, tra cui il benessere sociale e la salute mentale, e le connessioni tra la DAD e il rendimento scolastico.

Dopo uno studio da parte di Gallo & al. della curva di concentrazione di Lorenz attraverso un approccio matematico in cui tale curva viene idealmente approssimata ad un arco appartenente ad un particolare tipo di supercerchi o squircle definito appunto "super-ellisse", si torna nuovamente ad un contributo che ha come elemento centrale la pandemia da Covid-19, che da quasi 2 anni ha purtroppo cambiato l'intero tessuto sociale mondiale. D'Uggento e Tarantini partono dalla serie dei decessi nel comune di

Bari dal 2010 ad oggi e, attraverso un approccio classico di analisi delle serie storiche, sviluppano un modello statistico funzionale ai processi organizzativi interni del Comune.

Nell'ottavo contributo, D'Uggento et al. affrontano, invece, i principali risultati di un'analisi statistica condotta sugli effetti di un farmaco biologico, il Benralizumab, che viene utilizzato nella terapia dell'asma cronico. Attraverso uno studio retrospettivo, gli autori individuano le caratteristiche dei pazienti che hanno mostrato una migliore risposta alla terapia con Benralizumab, al fine di ottenere un modello predittivo.

Si prosegue quindi con Carucci e Vannella che intervengono sull'attenzione sempre crescente verso le forme di economia sostenibile da parte delle realtà imprenditoriali. Attraverso i dati forniti dai Censimenti permanenti delle imprese, anche in un'ottica territoriale, sono valutate le azioni di sostenibilità ambientale poste in essere dalle imprese.

Dopo il contributo di Diomede e Tagliatela, che studiano una generalizzazione della funzione logistica, il volume si chiude con Muserra e Grimaldi che, in un'ottica economico-aziendale, analizzano la relazione tra crisi finanziaria e politiche di bilancio delle imprese. I due autori indagano sulla possibilità che, nel contesto italiano, le condizioni macroeconomiche derivate dalla crisi finanziaria del 2008 siano state un incentivo per politiche di manipolazione del reddito.

Certi che il volume, nella sua eterogeneità, quale punto di forza del volume stesso, possa riscontrare il favore dei lettori, si ringraziano tutti coloro che vi hanno contribuito, sia Autori che referee.

Bari, 20/12/2021

*Ernesto Toma*

# A Brief Review on Compositional Inference of Microbiome Data

Massimo Bilancia<sup>1\*</sup>, Fabio Manca<sup>2</sup>, Gianvito Pio<sup>3</sup>

<sup>1</sup>*Ionian Department (DSJGE) – University of Bari Aldo Moro, Via Duomo 259, 74123 Taranto IT,* <sup>2</sup>*Department of Education Science, Psychology, Communication Science (FOR.PSI.COM) – University of Bari Aldo Moro, Palazzo Chiaia/Napolitano, Via Scipione Crisanzio 42, 70122 Bari IT,* <sup>3</sup>*Department of Computer Science – University of Bari Aldo Moro, Via Orabona 4, 70125 Bari IT.*

**Summary:** The microbiota can be defined as the collection of all microbes, bacteria, viruses, and fungi living in a host. Recent evidence suggests that the bacterial microbiota of the human gut plays an important role in the pathogenesis of many diseases by promoting inflammation. Historically, the first structured model for data analysis of the human gut microbiota was proposed in (Holmes et al., 2012). The central idea of this model is that the microbiota of a given individual represents a sample of a bacterial metacommunity, also called enterotype. An enterotype therefore represents a particular composition of the microbiota in terms of the relative abundances of different bacterial species. A similar specification was recently proposed in the unsupervised setting in (Anderlucci & Viroli, 2020). In both cases, the focus is on classification, and it was proposed to marginalize with respect to probability distributions over enterotypes to obtain a class-conditional likelihood that depends only on the specific parameters of each component of the mixture and the latent indicator variable of the metacommunity (enterotype). We review these two models available in the literature to highlight their similarities and differences, also discussing possible future research.

**Keywords:** Human Microbiome Data; Bayesian Statistics; Hierarchical Modeling; Multinomial-Dirichlet Finite-Mixture Model.

## 1. Introduction

The microbiome can be defined as the collection of all microbes, bacteria, viruses, and fungi living in a host. Recent evidence suggests that the bacterial microbiota of

---

\* Corresponding author: massimo.bilancia@uniba.it

The three authors reviewed and revised the manuscript, contributing equally to the draft, and approved the final version by agreeing to submit it for publication.

the human gut plays an important role in the pathogenesis of many diseases by promoting inflammation (Camarota et al., 2020; Duvallet et al., 2017; Lu et al., 2019; Valdes et al., 2018). Furthermore, gut bacterial communities change after transplantation and immunosuppressive therapy, and the prevalence of certain pathogens in these communities may be indicative of the establishment of an alloimmune response and graft rejection (Annavaiah et al., 2019).

Gut microbial communities are profiled by sequencing specific regions, such as 16S ribosomal RNA (or 16S rRNA). This biological macromolecule is part of the prokaryotic ribosome, and the genes that encode it evolve very slowly. Therefore, this sequence is largely similar across bacteria, but it also contains hypervariable subdomains that are specific to each bacterial species. These sequences are grouped into operational taxonomic units (OTUs), i.e. clusters of similar sequences, to represent the abundance of a particular bacterial species in the microbiota while avoiding the excessive sparseness that would result from grouping only identical sequences (Wang et al., 2007).

Historically, the first structured model for data analysis of the human gut microbiota was proposed in (Holmes et al., 2012). The central idea of this model is that the microbiota of a given individual represents a sample of a bacterial metacommunity, also called enterotype. An enterotype therefore represents a particular composition of the microbiota in terms of the relative abundances of different bacterial species and can be associated with particular physiological, pathological or experimental conditions. Since we do not know which enterotype each individual belongs to, it is natural to treat the data generating process as a finite mixture of probability distributions, each of which is a probability distribution over bacterial species and determines their frequency of occurrence in samples belonging to the same enterotype. This model has interesting performance when used for classification, both in supervised and unsupervised settings.

A similar specification was recently proposed in the unsupervised setting in (Anderlucci & Viroli, 2020), which considered marginalization with respect to probability distributions over enterotypes to obtain a class-conditional likelihood that depends only on the specific parameters of each component of the mixture and the latent indicator variable of the metacommunity (enterotype). In this way, the parameters can be easily estimated using the EM algorithm fitted to the Dirichlet-Multinomial distribution.

The aim of this paper is therefore to review in more detail the two models available in the literature and to highlight their similarities and differences. The paper is organized as follows. In Section 2, we describe the Unigram model as a basic

building block for human microbiota data analysis. In Section 3, we present a simple hierarchical extension of the Unigram model, where we have a set of  $k$  Multinomial parameters, each representing a particular enterotype. Section 4 is specifically devoted to the Holmes-Harris-Quince (HHQ) model and the description of its hierarchical structure, while Section 5 focuses on mixtures of Dirichlet-Multinomial distributions and, in particular, on the above-mentioned model of Anderlucci and Viroli, of which we present a simplified version (AV2 model) that has a simpler biological interpretation. Section 6 is devoted to the differences between the two models, comparing their relative hierarchical structures. Section 7 reports a simple case study with real data. Finally, Section 8 discusses possible future research approaches.

## 2. Unigram models

The basic data structure is an OTU table of size  $n \times p$ , where the generic element  $y_{i\ell}$  denotes the number of copies of the  $\ell$ -th OTU ( $\ell = 1, 2, \dots, p$ ) detected in the  $i$ -th sample ( $i = 1, 2, \dots, n$ ). The natural model for the likelihood of OTU abundances detected in each experimental sample is the  $p$ -dimensional Multinomial distribution:

$$p(y_i | \beta) = \frac{(\sum_{\ell=1}^p y_{i\ell})!}{\prod_{\ell=1}^p y_{i\ell}!} \prod_{\ell=1}^p \beta_{\ell}^{y_{i\ell}}, \quad (1)$$

where  $\beta = (\beta_1, \beta_2, \dots, \beta_p) \in \mathbb{R}^p$  is the vector of multinomial parameters that must satisfy the constraints  $\beta_j > 0$  for  $j = 1, 2, \dots, p$  and  $\sum_{\ell=1}^p \beta_{j\ell} = 1$ . The multinomial coefficient is usually abbreviated as:

$$\frac{(\sum_{\ell=1}^p y_{i\ell})!}{\prod_{\ell=1}^p y_{i\ell}!} \equiv \binom{y_{i+}}{y_i}, \quad (2)$$

with  $y_{i+} = y_{i1} + y_{i2} + \dots + y_{ip}$ . The parameter  $y_{i+}$  is a nuisance parameter that only enters the normalization constant and is therefore irrelevant for the inference from the posterior distribution of the parameters.

From the perspective of textual data analysis, the Multinomial model can be interpreted as a Unigram model (Nigam et al., 2000). Suppose we have a vocabulary  $V$  with  $p = |V|$  terms, where in this interpretation an OTU corresponds to a vocabulary term, while the index  $i$  denotes the  $i$ -th document of a corpus of  $N$  documents. A document can be viewed as a stream of  $S$  terms from this vocabulary. The  $\ell$ -th term in  $V$  is represented by a  $p$ -dimensional vector  $w$  of the unit basis, such that the

only non-zero component of  $w$  is the  $\ell$ -th component. In the Unigram model, the conditional likelihood has the expression:

$$w_{is} | \beta \stackrel{\text{ind.}}{\sim} \text{Multinouilli}_p(\beta), \quad s = 1, 2, \dots, S. \quad (3)$$

Since the generic  $w$ -vector has expression  $w_{is} = (0, \dots, 0, 1, 0, \dots, 0)$ , with  $w_{i+} = 1$ , the Multinouilli distribution in (3) indicates a Multinomial distribution on a single trial. It can also be viewed as a generalization of the Bernoulli distribution for the case of  $S > 2$  groups. The specification (3) is infinitely exchangeable (Gelman et al., 2013), and the data-generating mechanism can be viewed as a random mechanism that generates streams of terms such that, regardless of the length of the sequence, two sequences that differ only in the order of occurrence of the terms have the same probability of being generated. This model captures, at a low level, the outcome of the process of classifying 16S rRNA sequences into a bacterial taxonomy and encodes the fact that the only relevant information is whether a particular bacterial taxon is present in the gut bacterial community of an individual. The frequency counts of each OTU can be expressed as:

$$y_i = \sum_{s=1}^S w_{is}, \quad (4)$$

since by summing the unit vectors  $w_{is}$  over all possible terms in the document, we obtain the counts of each term in the vocabulary  $V$ .

The vector  $y_i = (y_{i1}, y_{i2}, \dots, y_{ip})$  can obviously correspond to several different configurations of the indicator vectors  $w_{is}$ . Given the independence and expression of the Multinouilli pmf in (3), the probability of observing  $y_i$  for a single configuration is:

$$\underbrace{\beta_1 \times \beta_1 \times \dots \times \beta_1}_{y_{i1} \text{ times}} \times \underbrace{\beta_2 \times \beta_2 \times \dots \times \beta_2}_{y_{i2} \text{ times}} \times \dots \times \underbrace{\beta_p \times \beta_p \times \dots \times \beta_p}_{y_{ip} \text{ times}} = \beta_1^{y_{i1}} \beta_2^{y_{i2}} \dots \beta_p^{y_{ip}}, \quad (5)$$

and summing this probability over all possible configurations gives the Multinomial likelihood (1), showing complete equivalence between the two representations.

### 3. Hierarchical mixtures of unigrams

We can construct a finite mixture of Unigrams, that is a mixture of Multinomial distributions, to have greater flexibility. In this case, a microbial host community can

belong to one of  $k$  different enterotypes. An enterotype is a probability distribution over bacterial taxa, and different enterotypes therefore have different relative compositions, meaning that bacterial taxa that are highly expressed in one enterotype may be of little importance in another enterotype. In a finite mixture of Unigrams we have  $k$  probability distributions over the vocabulary of terms, each of which represents a different physiological or experimental condition (Cheng et al., 2014). In this case, the Multinomial parameter vector is replaced by a matrix  $\beta \in \mathbb{R}^{k \times p}$ , i.e.:

$$\beta = \{\beta_{j\ell}\}, \quad j = 1, 2, \dots, k, \quad \ell = 1, 2, \dots, p. \quad (6)$$

In other words, each row of the  $\beta$  matrix is a probability distribution representing a Multinomial probability vector (an enterotype). Since we do not know which enterotype each sample belongs to, we can use a finite mixture of Unigrams written in hierarchical form as follows (Gelman et al., 2013):

$$\begin{aligned} y_i | \beta, z_i &\stackrel{\text{ind.}}{\sim} \text{Multinomial}_p(\beta_j), \quad i = 1, 2, \dots, n, \\ z_i | \lambda &\stackrel{\text{ind.}}{\sim} \text{Multinouilli}_k(\lambda), \quad i = 1, 2, \dots, n, \end{aligned} \quad (7)$$

where  $\beta_j \in \mathbb{R}^p$  denotes the row of the matrix  $\beta$  corresponding to the index of the single element in the latent indicator vector  $z_i \in \mathbb{R}^k$  such that  $z_{ij} = 1$ , while  $\lambda \in \mathbb{R}^k$  denotes the mixture weights. Thus, the probability distribution over the  $j$ -th enterotype ( $j = 1, 2, \dots, k$ ) is described by a Multinomial distribution with enterotype-specific Multinomial probabilities selected by the latent indicator vector  $z_i$  specific to the  $i$ -th microbial community present in the  $i$ -th individual likelihood (1),.

#### 4. The Holmes-Harris-Quince (HHQ) model

In a mixture of Unigrams, the Multinomial parameters  $\beta_j$  are constant for all samples generated from the same enterotype. However, this hypothesis may not be valid due to variability in the underlying biological material. To account for this additional source of variability, we can treat the multinomial parameters as random variables by specifying a prior distribution for each row of the  $\beta$  matrix (Subedi et al., 2020; Xia et al., 2013). The first formulation of a hierarchical mixture model for the analysis of the human microbiota goes back to (Holmes et al., 2012), who proposed a model with the following hierarchical structure:

$$\begin{aligned}
y_i | \beta_i &\stackrel{\text{ind.}}{\sim} \text{Multinomial}_p(\beta_i), \quad i = 1, 2, \dots, n, \\
z_i | \lambda &\stackrel{\text{ind.}}{\sim} \text{Multinouilli}_k(\lambda), \quad i = 1, 2, \dots, n, \\
\beta_i | z_i, \theta &\stackrel{\text{ind.}}{\sim} \text{Dirichlet}_p(\theta_j), \quad i = 1, 2, \dots, n,
\end{aligned} \tag{8}$$

with  $\theta_j \in \mathbb{R}^p$ , for  $j = 1, 2, \dots, k$ , and  $\theta_{j\ell} > 0$  for  $\ell = 1, 2, \dots, p$ . In a full Bayesian context, this structure can be completed, e.g., by setting independent Gamma priors over the mixture parameters  $\theta_j$  and a Uniform distribution over each component of the mixing weights  $\lambda$ . However, this hierarchical structure is quite peculiar, since we have a set of Multinomial parameters specific to each sample, and a mixture-like prior distribution over such parameters that depends on enterotype-specific hyperparameters  $\theta_j$ . To better understand this formulation, we observe that by integrating the multinomial parameters into the conditional likelihood in (8) we obtain:

$$\begin{aligned}
p(y_i | z_i, \theta) &= \int p(y_i, \beta_i | z_i, \theta) d\beta_i = \int p(y_i | \beta_i) p(\beta_i | z_i, \theta) d\beta_i = \\
&= \int \binom{y_{i+}}{y_i} \prod_{\ell=1}^p \beta_{i\ell}^{y_{i\ell}} \frac{\Gamma(\sum_{\ell=1}^p \theta_{j\ell})}{\prod_{\ell=1}^p \Gamma(\theta_{j\ell})} \prod_{\ell=1}^p \beta_{i\ell}^{\theta_{j\ell}-1} d\beta_{i\ell} = \\
&= \binom{y_{i+}}{y_i} \frac{\Gamma(\sum_{\ell=1}^p \theta_{j\ell})}{\prod_{\ell=1}^p \Gamma(\theta_{j\ell})} \int \prod_{\ell=1}^p \beta_{i\ell}^{y_{i\ell} + \theta_{j\ell} - 1} d\beta_{i\ell} = \\
&= \binom{y_{i+}}{y_i} \frac{\Gamma(\sum_{\ell=1}^p \theta_{j\ell})}{\prod_{\ell=1}^p \Gamma(\theta_{j\ell})} c^{-1} \underbrace{\int c \prod_{\ell=1}^p \beta_{i\ell}^{y_{i\ell} + \theta_{j\ell} - 1} d\beta_{i\ell}}_{=1},
\end{aligned} \tag{9}$$

where the inverse of the normalization constant  $c$  in (9) has expression:

$$c^{-1} = \frac{\prod_{\ell=1}^p \Gamma(y_{i\ell} + \theta_{j\ell})}{\Gamma(\sum_{\ell=1}^p (y_{i\ell} + \theta_{j\ell}))}. \tag{10}$$

Using standard notation for the multivariate Beta function:

$$B(x) = B(x_1, x_2, \dots, x_p) = \frac{\prod_{\ell=1}^p \Gamma(x_\ell)}{\Gamma(\sum_{\ell=1}^p x_\ell)}, \tag{11}$$

the conditional likelihood (9) can be rewritten as:

$$p(y_i | z_i, \theta) = \binom{y_{i+}}{y_i} \frac{B(y_i + \theta_j)}{B(\theta_j)}. \tag{12}$$

The pmf (12) defines the Dirichlet-Multinomial distribution. It was studied, among others, by Mosimann (1962), who showed that the variance of each marginal component of the class-conditional distribution is:

$$\text{Var}(y_{i\ell}|z_i, \theta) = y_{i+} E(\beta_{j\ell}) E(1 - \beta_{j\ell}) \left( \frac{y_{i+} + \theta_{j+}}{1 + \theta_{j+}} \right), \quad (13)$$

with  $\theta_{j+} = \sum_p \theta_{j\ell}$ . Thus, as a consequence of explicitly accounting for variability in the multinomial probabilities, the variance of the class-conditional likelihood exhibits overdispersion with respect to the variance of the class-conditional likelihood in (7). The extent of this overdispersion, which depends on the heterogeneity of the underlying biological material, is controlled by the term  $\theta_{j+}$ , with higher values of  $\theta_{j+}$  corresponding to lower overdispersion.

## 5. Mixtures of Dirichlet-Multinomial distributions

The generative structure of the HHQ model introduces a finite mixture prior, since the Multinomial parameters of each observation are generated by a finite mixture of Dirichlet distributions. A possible alternative specification is the following:

$$\begin{aligned} y_i | \beta, z_i &\stackrel{\text{ind.}}{\sim} \text{Multinomial}_p(\beta_j), \quad i = 1, 2, \dots, n, \\ z_i | \lambda &\stackrel{\text{ind.}}{\sim} \text{Multinouilli}_k(\lambda), \quad i = 1, 2, \dots, n, \\ \beta_j | \theta &\stackrel{\text{ind.}}{\sim} \text{Dirichlet}_p(\theta_j), \quad j = 1, 2, \dots, k. \end{aligned} \quad (14)$$

In this model we have  $k$  distinct probability distributions  $\beta_j$  over the enterotypes that are explicitly part of the model. As in the case of the HHQ model:

$$\begin{aligned}
p(y_i|z_i, \theta) &= \int (y_i, \beta|z_i, \theta) d\beta = \int p(y_i|\beta, z_i) p(\beta|\theta) d\beta = \\
&= \int p(y_i|\beta_j) \prod_{s=1}^K p(\beta_s|\theta_s) d\beta_s = \\
&= \int p(y_i|\beta_j) p(\beta_j|\theta_j) d\beta_j \underbrace{\int \prod_{-j} p(\beta_{-j}|\theta_{-j}) d\beta_{-j}}_{=1} = \\
&= \int \binom{y_{i+}}{y_i} \prod_{\ell=1}^p \beta_{j\ell}^{y_{i\ell}} \frac{\Gamma(\sum_{\ell=1}^p \theta_{j\ell})}{\prod_{\ell=1}^p \Gamma(\theta_{j\ell})} \prod_{\ell=1}^p \beta_{j\ell}^{\theta_{j\ell}-1} d\beta_{j\ell} = \\
&= \binom{y_{i+}}{y_i} \frac{\Gamma(\sum_{\ell=1}^p \theta_{j\ell})}{\prod_{\ell=1}^p \Gamma(\theta_{j\ell})} \int \prod_{\ell=1}^p \beta_{j\ell}^{y_{i\ell}+\theta_{j\ell}-1} d\beta_{j\ell},
\end{aligned} \tag{15}$$

and then continuing the proof in the same way that applies to the HHQ model, the class-conditional result to be a Dirichlet-Multinomial:

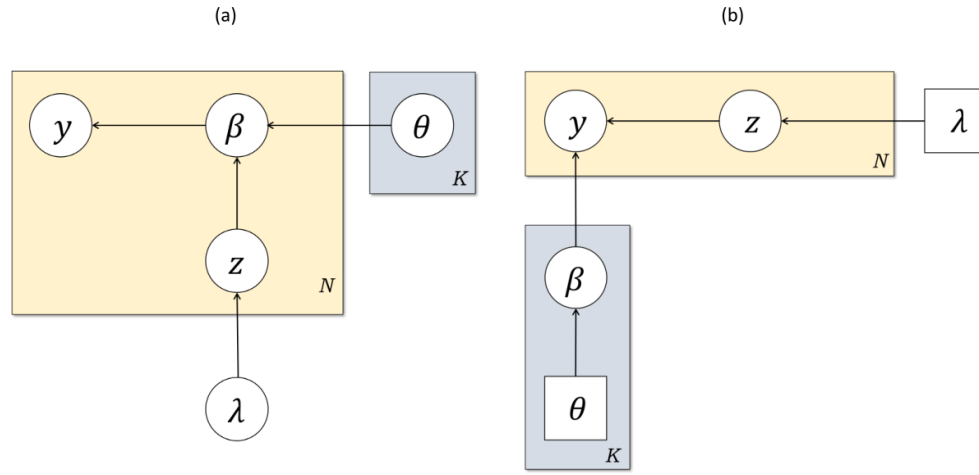
The hierarchical model defined by (14) is a suitably refactored version of the model presented in (Anderlucci & Viroli, 2020). In the originally proposed specification of each  $\beta_j$  prior in (14) also depended on the latent indicator variable  $z_i$ . However, this formulation was redundant since the exchangeable prior distribution in (14) induces a class-conditional likelihood equivalent to that of the model originally proposed in Anderlucci & Viroli (2020). In what follows, we refer to the model we presented in this section as the AV2 model to highlight the subtle differences that exist in our hierarchical specification.

## 6. A comparison of the AV2 and HHQ models

The representation in terms of oriented graphs of the two hierarchical models HHQ and AV2 is shown in Figure 1. The specification of  $k$  probability distributions over the multinomial parameters of the AV2 model is equivalent to a subset of the generative structure of the Latent Dirichlet Allocation (LDA) model introduced in Blei et al. (2003), in which we have  $k$  probability distributions over the multinomial probabilities of occurrence of terms from a vocabulary  $V$ . However, the LDA model is a mixed membership model (Airoldi et al., 2014), where each OTU can be associated with one of the  $k$  probability distributions  $\beta_j$ . Therefore, the starting point of the LDA model is the Multinoulli conditional likelihood (3), and the bacterial community of each individual can be viewed as a mixture of enterotypes with assignment

probabilities determined by the prior distribution over the latent variables  $z_i$ . The view based on the LDA model is discussed in detail in Sankaran & Holmes (2019) and will not be pursued here. In contrast, the AV2 model is based on a biologically simpler assumption, namely that each community can belong to a unique enterotype because the configuration of the latent indicator variable  $z_i$  selects a row of the  $\beta$ -matrix for the entire sample.

**Figure 1.** The Dirichlet-Multinomial mixture model in graphical form. (a) HHQ model (b) AV2 model. The Bayesian specification of the HHQ model can be completed, for example, by setting independent Gamma priors over the mixture parameters  $\theta_j$  and a Uniform distribution over each component of the mixing weights  $\lambda$ . In the AV2 model, the vector parameters  $\theta_j$  and  $\lambda$  are assumed to be fixed but unknown.



Enterotype distributions do not appear explicitly in the model specification, but can be approximated using estimates of the Dirichlet parameters  $\theta_j$ s. To be more precise, define (for  $j = 1, 2, \dots, k$ ):

$$\begin{aligned} \psi_j &= \sum_{\ell=1}^p \theta_{j\ell} \in \mathbb{R}, \\ \alpha_j &= \psi_j m_j \in \mathbb{R}^p, \end{aligned} \tag{16}$$

with:

$$m_j = (m_{j1}, m_{j2}, \dots, m_{jp}), \quad \sum_{\ell=1}^p m_{j\ell} = 1, \tag{17}$$

that is each  $m_j$  is a probability measure (normalized to 1). That every  $m_j$  is actually a probability measure is easy to prove, because:

$$E(\beta_{i\ell}|z_i, \alpha) = \frac{\alpha_{j\ell}}{\sum_{\ell=1}^p \alpha_{j\ell}} = \frac{\psi_j m_{j\ell}}{\sum_{\ell=1}^p \psi_j m_{j\ell}} = \frac{\psi_j m_{j\ell}}{\psi_j \sum_{\ell=1}^p m_{j\ell}} = m_{j\ell}, \quad (18)$$

where  $\alpha = (\alpha_1, \alpha_2, \dots, \alpha_k)$  and the  $j$ -index is determined by the value assumed by the latent indicator  $z_i$ . Thus, the elements of the  $m_j$  vector can be interpreted as the expected values of the Multinomial probabilities of samples belonging to the  $j$ -th enterotype, where  $\psi_j$  acts like a precision with a large  $\psi_j$  giving a small variance around these expected values.

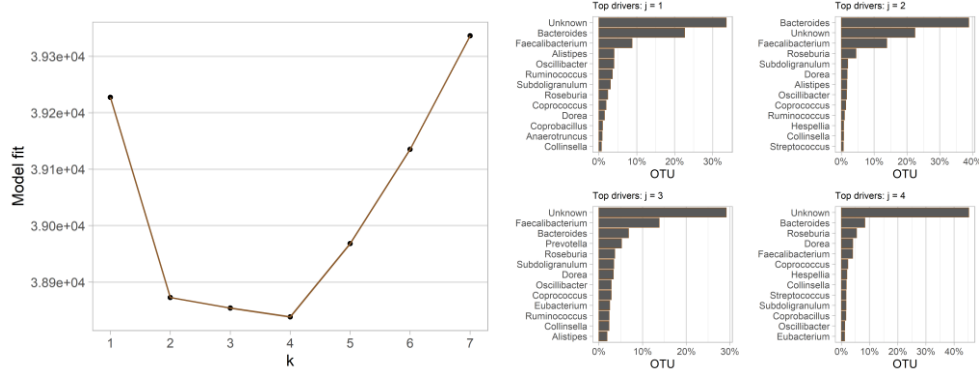
The AV2 model explicitly introduces probability distributions  $\beta_j$  for the enterotypes. However, in Anderlucci & Viroli (2020) a EM-like algorithm for parameter estimation is proposed based on the class-conditional Dirichlet-Multinomial likelihood, where the parameters  $\beta_j$  have been integrated out. A possible alternative approach is not to integrate over the enterotype distributions. In this way, we have more flexibility in modeling because we can structure the distributions over the enterotypes in a more realistic way that better fits the underlying biological processes without the limitation of analytical integrability. The obvious drawback of this approach is that the analytical simplicity of the pure Dirichlet-Multinomial model is lost, and the marginal likelihood is no longer available in closed form.

## 7. The Twin study

In this section, we present a brief illustrative example of the capabilities of the methods we presented using a dataset originally presented in Holmes et al. (2012). The original data were sampled from 154 different individuals characterized by family and body mass index. Each individual was sampled at two time points approximately two months apart. Of the 308 possible samples, some contained no reads after pre-processing, leaving a total of 278 samples available.

Using the HHQ model, the mixture of Dirichlet prior can be used to cluster samples at the metacommunity level. Assuming that each sample represents a unique community, we can infer which metacommunity that sample is most likely to have originated from. The number of mixture components  $k$  is estimated by minimizing the negative log-marginal distribution, approximated using the Laplace quadrature formula (Figure 2, left panel). In this way, we compute the model fit taking into account the model complexity, and the results show a minimum for  $k = 4$ , suggesting that a mixture of four Dirichlet distributions is more appropriate for these data.

**Figure 2.** Left: negative log-marginal distribution for varying  $k$ , estimated using the Laplace quadrature formula. Right: Estimated enterotype distributions.



Enterotype estimates were obtained by expression (18) and are shown in Figure 2 (right panel). They can be used to characterize the bacterial composition of the intestinal flora in the population we studied and possibly compare it with the composition of other populations (e.g., in individuals affected by a particular disease) or to assign each sample to the enterotype for which the respective posterior probability is highest.

## 8. Future developments

If we do not want to marginalize with respect to the  $\beta_j$ -distributions, it is natural to consider a fully Bayesian specification of the AV2 model consisting of the levels (14), and a symmetric Dirichlet prior distribution over the mixture weights:

$$\begin{aligned}
 y_i | \beta, z_i &\stackrel{\text{ind.}}{\sim} \text{Multinomial}_p(\beta_j), \quad i = 1, 2, \dots, n, \\
 z_i | \lambda &\stackrel{\text{ind.}}{\sim} \text{Multinouilli}_k(\lambda), \quad i = 1, 2, \dots, n, \\
 \beta_j | \theta &\stackrel{\text{ind.}}{\sim} \text{Dirichlet}_p(\mathbf{1}_p \theta), \quad j = 1, 2, \dots, k, \\
 \lambda | \alpha &\sim \text{Dirichlet}_k(\mathbf{1}_k \alpha)
 \end{aligned} \tag{19}$$

where  $\alpha, \theta \in \mathbb{R}$  with  $\alpha, \theta > 0$ . We will assume that the hyperparameters  $\theta$  and  $\alpha$  are known and fixed by the researcher based on purely empirical considerations. Exploiting the conditional independence between the marginal components of the likelihood and the prior distributions of the model parameters, the marginal likelihood of the model is:

$$p(y|\theta, \alpha) = \int \prod_{i=1}^n \sum_{z_i} p(y_i|\beta, z_i) p(z_i|\lambda) p(\beta|\theta) p(\lambda|\alpha) d\lambda d\beta, \quad (20)$$

which is clearly intractable. Therefore, the posterior distribution is not available in closed form, and we must resort to appropriate numerical methods for Bayesian estimation of the model parameters. One possibility, which was examined in Bilancia et al. (2021) is to use the methods of variational inference (Blei et al., 2017). The starting point is to approximate the posterior distribution  $p(\beta, z, \lambda|y, \theta, \alpha)$  by a variational distribution  $q(\beta, z, \lambda|\nu)$ , which itself depends on the variational parameters  $\nu$ . Our variational objective is:

$$\nu^* = \operatorname{argmin}_{\nu} \operatorname{KL}(q(\beta, z, \lambda|\nu) \| p(\beta, z, \lambda|y, \theta, \alpha)), \quad (21)$$

where the objective function is the reverse Kullback-Leibler (KL) divergence between the posterior distribution and the variational distribution. The solution of (21) provides the best possible approximation to the intractable posterior distribution with respect to the KL-divergence, and the optimized variational distribution  $q(\beta, z, \lambda|\nu^*)$  is used to approximate the posterior inference of the model parameters. In this way, the posterior inference is treated as an optimization problem rather than a Monte Carlo sampling problem (Ghahramani, 2015).

The search for the optimal approximating distribution can be greatly simplified if we define the Evidence Lower Bound (ELBO) as follows:

$$\operatorname{ELBO}(q) = \mathbb{E}_q[\log p(y, z, \beta, \lambda|\theta, \alpha)] - \mathbb{E}_q[\log q(\beta, z, \lambda|\nu)], \quad (22)$$

which is a functional of the variational distribution  $q$ . By making explicit the computation of the expected values in (22), the ELBO is a function of both the variational parameters and the fixed hyperparameters  $\theta$  and  $\alpha$ . It can then be shown that (Tran et al., 2021):

$$\log p(y|\theta, \alpha) = \operatorname{ELBO}(q) + \operatorname{KL}(q(\beta, z, \lambda|\nu) \| p(\beta, z, \lambda|y, \theta, \alpha)). \quad (23)$$

Since the KL-term is always positive, minimizing the KL-divergence with respect to the variational parameters is equivalent to maximizing the ELBO with respect to the variational parameters, and  $\log p(y|\theta, \alpha) \geq \operatorname{ELBO}(q)$ , that is the ELBO is a lower bound on the marginal log-likelihood. Therefore, minimizing the free energy of the variational distribution with respect to the variational parameters is equivalent to determining the tightest possible lower bound on the marginal log-likelihood. Using an appropriate variational distribution  $q$ , Bilancia et al. (2021) provide an explicit

expression of the ELBO that is maximized for parameter estimation using a coordinate ascent algorithm.

Several other extensions to the proposed model are indeed possible, through the use of nonparametric Bayesian models that would allow flexible modelling of associations between microbial communities and covariates such as environmental factors and clinical characteristics. These extensions are currently under investigation.

## References

- Airoldi, E. M., Blei, D., Erosheva, E. A., & Fienberg, S. E. (Eds.). (2014). *Handbook of Mixed Membership Models and Their Applications*. Chapman and Hall/CRC. <https://doi.org/10.1201/b17520>
- Anderlucci, L., & Viroli, C. (2020). Mixtures of Dirichlet-Multinomial distributions for supervised and unsupervised classification of short text data. *Advances in Data Analysis and Classification*, 14(4), 759–770. <https://doi.org/10.1007/s11634-020-00399-3>
- Annavajhala, M. K., Gomez-Simmonds, A., Macesic, N., Sullivan, S. B., Kress, A., Khan, S. D., Giddins, M. J., Stump, S., Kim, G. I., Narain, R., Verna, E. C., & Uhlemann, A.-C. (2019). Colonizing multidrug-resistant bacteria and the longitudinal evolution of the intestinal microbiome after liver transplantation. *Nature Communications*, 10(1), 4715. <https://doi.org/10.1038/s41467-019-12633-4>
- Bilancia, M., Di Nanni, M., Manca, F., & Pio, G. (2021). Variational Bayes estimation of hierarchical Dirichlet-Multinomial mixtures for unsupervised classification. *Submitted*.
- 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. <https://doi.org/10.1080/01621459.2017.1285773>
- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation Michael I. Jordan. *Journal of Machine Learning Research*, 3, 993–1022.
- Cammarota, G., Ianiro, G., Ahern, A., Carbone, C., Temko, A., Claesson, M. J., Gasbarrini, A., & Tortora, G. (2020). Gut microbiome, big data and machine learning to promote precision medicine for cancer. *Nature Reviews Gastroenterology and Hepatology*, 17(10), 635–648. <https://doi.org/10.1038/s41575-020-0327-3>
- Cheng, X., Yan, X., Lan, Y., & Guo, J. (2014). BTM: Topic Modeling over Short Texts. *IEEE Transactions on Knowledge and Data Engineering*, 26(12), 2928–2941. <https://doi.org/10.1109/TKDE.2014.2313872>
- Duvallet, C., Gibbons, S. M., Gurry, T., Irizarry, R. A., & Alm, E. J. (2017). Meta-

- analysis of gut microbiome studies identifies disease-specific and shared responses. *Nature Communications*, 8(1), 1784. <https://doi.org/10.1038/s41467-017-01973-8>
- Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A., & Rubin, D. (2013). *Bayesian Data Analysis* (Third). Chapman and Hall/CRC.
- Ghahramani, Z. (2015). Probabilistic machine learning and artificial intelligence. *Nature*, 521(7553), 452–459. <https://doi.org/10.1038/nature14541>
- Holmes, Ian, Harris, K., & Quince, C. (2012). Dirichlet Multinomial Mixtures: Generative Models for Microbial Metagenomics. *PLoS ONE*, 7(2), e30126. <https://doi.org/10.1371/journal.pone.0030126>
- Lu, H.-F., Ren, Z.-G., Li, A., Zhang, H., Xu, S.-Y., Jiang, J.-W., Zhou, L., Ling, Q., Wang, B.-H., Cui, G.-Y., Chen, X.-H., Zheng, S.-S., & Li, L.-J. (2019). Fecal Microbiome Data Distinguish Liver Recipients With Normal and Abnormal Liver Function From Healthy Controls. *Frontiers in Microbiology*, 10, 1518. <https://doi.org/10.3389/fmicb.2019.01518>
- Mosimann, J. E. (1962). On the Compound Multinomial Distribution, the Multivariate  $\beta$ -Distribution, and Correlations Among Proportions. *Biometrika*, 49(1/2), 65. <https://doi.org/10.2307/2333468>
- 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. <https://doi.org/10.1023/A:1007692713085>
- Sankaran, K., & Holmes, S. P. (2019). Latent variable modeling for the microbiome. *Biostatistics*, 20(4), 599–614.
- Subedi, S., Neish, D., Bak, S., & Feng, Z. (2020). Cluster analysis of microbiome data by using mixtures of Dirichlet–multinomial regression models. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 69(5), 1163–1187. <https://doi.org/10.1111/rssc.12432>
- Tran, M.-N., Nguyen, T.-N., & Dao, V.-H. (2021). *A practical tutorial on Variational Bayes*. <http://arxiv.org/abs/2103.01327>
- Valdes, A. M., Walter, J., Segal, E., & Spector, T. D. (2018). Role of the gut microbiota in nutrition and health. *BMJ (Online)*, 361, 36–44. <https://doi.org/10.1136/bmj.k2179>
- Wang, Q., Garrity, G. M., Tiedje, J. M., & Cole, J. R. (2007). Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Applied and Environmental Microbiology*, 73(16), 5261–5267. <https://doi.org/10.1128/AEM.00062-07>
- Xia, F., Chen, J., Fung, W. K., & Li, H. (2013). A Logistic Normal Multinomial Regression Model for Microbiome Compositional Data Analysis. *Biometrics*, 69(4), 1053–1063. <https://doi.org/10.1111/biom.12079>

# Applications of text mining in the biomedical field

Claudia Marin<sup>1\*</sup>, Fabio Manca<sup>2</sup>

<sup>1 2</sup>*Dipartimento di Scienze della Formazione, Psicologia, Comunicazione,  
Università degli studi di Bari Aldo Moro*

**Abstract:** Text mining is one of the technologies designed to improve the quality of clinical medicine services through the intelligent analysis of biomedical information. In recent years, clinical natural language processing has received considerable attention due to its role in uncovering information hidden in clinical texts. Adapting and extending natural language processing (NLP) techniques can open the doors to better clinical records.

**Keywords:** text mining; biomedical field, cancer.

## 1. Introduction

Biomedical text mining is an artificial intelligence technology that uses natural language processing to transform free, unstructured text in documents and databases into structured and normalized data suitable for analysis or use by machine learning algorithms. As a research area, it encompasses ideas from natural language processing, bioinformatics, medical informatics, and computational linguistics.

Text mining involves the use of many computer technologies:

- Machine learning.
- Natural language processing.
- Biostatistics.
- Information technology.

---

\* Autore corrispondente: [claudia.marin@uniba.it](mailto:claudia.marin@uniba.it)

Il lavoro qui descritto è frutto di un progetto comune, ma Marin C. ha provveduto alla redazione dei paragrafi 1 e 2, mentre Manca F. ha redatto il paragrafo 3 e 4.

- Model recognition.

We have a large number of biomedical texts, which are a rich source of knowledge for biomedical research. Text mining, which is now widely used in this field, extracts the information we need and thus contributes to the improvement of knowledge. For this reason, many researchers have taken advantage of this text mining technology, as it can also improve the development of biomedical research related to malignant diseases, such as cancer.

Cancer is a deadly human disease that caused 7.4 million deaths in 2008 alone. For biomedical researchers, cancer is one of the most important areas of study.

Cancer-related text mining applications include:

- Identifying biomedical references related to malignant tumours (genes, proteins, etc.).
- The search for relationships between biomedical entities (protein-protein, gene-disease, etc.).
- The extraction of knowledge from text and the generation of hypotheses.
- The construction or improvement of pathways (Zhu et al., 2013).

Extracting the tacit knowledge hidden in the unstructured text and representing it in explicit form is the goal of text mining. It is done by retrieving and extracting information, discovering knowledge, and generating hypotheses that infer unknown biomedical facts from the text.

## 1.1 Information Retrieval

There are traditional information retrieval systems, but also advanced information retrieval systems for knowledge. To improve our understanding of complex biomedical systems, the latter integrate data from multiple resources into a single context. For example, Maier et al. (2011) created a knowledge base on chronic obstructive pulmonary disease (COPD) and developed an integrated knowledge management system to access text mining results and other data. The Salivaomics knowledge base defined the Saliva ontology as a vocabulary of terms and relationships to facilitate retrieval and integration of data into multiple search fields, as well as data analysis and data mining (Ai et al., 2011). QuExT, a PubMed-based document retrieval system, has applied a concept-oriented query expansion methodology to find documents that contain concepts related to query words (Matos et al., 2010).

Thanks to advances in biotechnology and high-throughput methods for genetic analysis, the need for text mining and information retrieval tools will increase to help researchers find articles relevant to their studies.

## 1.2 Relation Extraction and Named Entity Recognition

In knowledge extraction, the most important step is named entity recognition. The general goal of this recognition is to identify specific terms such as gene, protein, disease, and drug.

Although various computer technologies have been used to identify biomedical terms, there are still many obstacles associated with the automatic identification of these terms. This is because a biomedical term may have different spellings even though it refers to the same thing or entity: for example, the word cancer can refer to a disease, but it can also refer to an astronomical sign. Not to be forgotten are the abbreviations, which, as has been shown in other cases, always lead to problems of ambiguity.

Current biomedical technology for named entity recognition falls into three main categories:

- dictionary-based approaches, which tend to overlook indeterminate terms that are not listed in the dictionary (Rebholz-Schuhmann et al., 2011);
- rule-based approaches that require rules to identify terms from the text, and the resulting rules are often not effective in all cases;
- machine learning approaches, which typically require standardized, annotated training datasets that usually require enormous human effort to create (Krallinger et al., 2011).

The machine learning techniques used for named entity recognition are:

- Hidden Markov Models (HMM). Zhou et al. (2004) used this system with biomedical information such as domain knowledge to recognize proteins, DNA, RNA, cell type, and cell line.
- Support vector machines (SVMs). Kazama et al. (2002) used SVMs to identify proteins, DNA, cell type, cell line, and lipids, with an F1 (F score) of 73.6%.
- Conditional random fields (CRFs). Tsai et al (2006) developed a CRF system to extract protein mentions and achieved an F1 rate of 78.4%.
- Maximum entropy (ME). Lin et al. (2004) used ME to recognize 23 categories of biological terms with an F1 rate of 72%.

Researchers have tried many methods to improve performance by combining different approaches and proposing hybrid approaches, performing post-processing after machine learning, and expanding knowledge of the biomedical domain. Nevertheless, for biomedical named entity recognition systems, the best F1 rates are currently not as good as the results of general systems (Dai et al., 2010). Returning to the issue of abbreviations: A biomedical term may not only occur in

this form, but may also have multiple synonyms in the text. The ability to recognize abbreviations and synonyms is useful “for unifying and normalizing biomedical terms in named entity recognition.” (Zhu et al., 2013) One of the systems that addresses abbreviation tagging is the logistic regression by Chang et al. which achieved a recognition rate of 83% and an accuracy rate of 80% with the Medstract corpus (Chang et al., 2002).

A system for recognizing abbreviations and definitions in biological literature was developed by BIOADI. It is based on a machine learning approach and has a hit rate of 95.86% and a recognition rate of 84.64% with the AB3P corpus (Kuo et al., 2009).

Yu et al. (2002) developed a set of pattern matching rules to map an abbreviation in its complete form. They achieved a recognition rate of 70% and a hit rate of 95%. Liu and Friedman’s (2003) system achieved a recognition rate of 88.5% and an accuracy rate of 96.3% based on the finding sites. Currently, research is focused on term identification and normalization. BioCreative III focused on gene normalization, which identifies gene mentions and links these genes to standard identifiers (e.g., database identifiers) (Arighi et al., 2011). Several researchers have focused extensively on relationship extraction. Conventional relationship extraction focuses on studying the extraction of biomedical relationships (protein-protein interaction and gene-disease relationship) starting from biomedical terms such as genes, proteins, diseases, or drugs. The system developed by Ben Abacha et al. (2011) is capable of using MetaMap to determine the correct semantic relationship between each pair of entities to identify medical substances.

The systems developed by Chun et al. (2006) were able to extract relationships between genes and diseases from Medline. Namely, they used a machine learning-based named entity recognition system to remove diseases and false gene names caused by recognizing terms based on dictionary correspondence. They found that improving term recognition performance can also improve report extraction accuracy.

Today, many researchers are interested in extracting gene-gene interactions, protein-protein interactions, and other interactions in genome-wide associations, which provide useful scaffolds for further integrative gene expression analysis and database annotation, and other extended reports.

In 2004, Eskin and Agichtein applied text mining technology and combined it with sequence analysis to discover the subcellular localizations of proteins. The results appeared to be extremely accurate. In 2010, Li et al. used a co-occurrence-based text mining approach to determine interactions from the biomedical literature,

using a naive Bayesian approach to verify the resulting interactions and integrate heterogeneous types of evidence from genomic and proteomic datasets. The systems developed by Agarwal et al. (2011) can instead be used to determine whether an article is associated with protein-protein interactions and to link the interaction to relevant articles.

Finally, Tsai (2011) presented a text mining and visualization system to find the details of protein-protein interactions to provide a deeper understanding of protein function by identifying the amino acid sequence at the interface of a protein interaction.

Regarding gene-disease and protein-subcellular relationships, Srinivasan and Wedemeyer (2003) studied the relationships between diseases and disease domains. Shetty and Dalal (2011) created a citation-based statistical document classifier from MEDLINE to determine whether a drug causes adverse effects. Their systems have contributed to the current drug safety process.

## **2. Cancer research and text mining**

We have numerous papers on biomedical text mining in cancer research. One of the most important topics in this area is DNA methylation, which is “an epigenetic modification of DNA, the process of which involves the binding of a methyl group (-CH<sub>3</sub>) to a nitrogenous base.”

Methylation profiling has been successfully used for early diagnosis and personalized treatment of cancer (Fang et al., 2008). Among the various databases developed for DNA methylation, the best known are:

- PubMeth, a cancer methylation database that uses text mining tools and expert commentary (Ongenaert et al., 2008);
- MeInfoText, where associations between genes, methylation and cancer are extracted from a large body of biomedical literature (Fang et al., 2008).

Cancer is a complex disease precisely because a large number of genes and proteins are involved. Biomedical researchers involved in cancer diagnosis, treatment, and prevention draw information about cancer-related genes and proteins from the literature.

In this context, a study was conducted in 2006 by Chun et al. who developed a maximum entropy-based system that detects the named entities and the relationships between prostate cancer and relevant genes.

In turn, Deng et al. in 2006 used a text-mining approach to identify genes associated with prostate cancer as candidate genes and used the OMIM (Online Mendelian Inheritance in Man) database to verify them. Natarajan et al. (2006) used a text-mining approach to create gene-gene interaction networks for 72 genes. They discovered new insights into the effect of S1P (sphingosine-1-phosphate) on angiogenesis and glioblastoma invasion, which contributed to the understanding of the interaction between invasive glioblastomas and S1P.

Krallinger et al (2010) implemented two cancer-related text mining applications. One was used to extract human genetic mutations of predefined cancers from the literature; the other was used specifically for breast cancer categorization and textual classification of breast cancer genes.

Finally, in 2011, Clancy et al. developed an immunological relevance score classified for all human genes and used it to evaluate gene expression profiles and quantify the immunological component of tumors. They applied it to expression profiles in melanoma to detect early activation of the adaptive immune response and diversity of the immunological component during melanoma progression. Areas related to other biomedical interactions, such as the studies by:

- Kolluru et al. who in 2011 used text mining workflows to automatically extract compounds from microorganisms and their free text habitats. They used conditional random fields to extract microorganisms, habitats, and the interactions between organisms and their habitats from the literature.
- In 2012, Xu et al. used the scientific literature in PubMed to extract experimental data on protein phosphorylation. The resulting information has proven invaluable to biomedical researchers studying cell processes and cancer.

A web-based tool for text mining that can extract the interaction network from PubMed abstracts and refine it with user-defined concepts is PESCADOR. PESCADOR has been used in research on protein aggregation in neurodegenerative diseases and in expanding pathways associated with colon cancer (Barbosa-Silva et al., 2011).

To reduce and control the causes of cancer, early detection and evidence-based strategies for prevention and treatment of patients play an important role. Therefore, in cancer research, it is crucial to conduct risk assessment for this disease. Indeed, through risk assessment, it is possible to determine the likelihood of developing cancer by evaluating the available data. Let us consider some examples:

- Korhonen et al. (2009) applied biomedical text mining technology to cancer risk assessment. They extracted evidence from the literature as features and developed

different classes for risk levels of cancer causes from which researchers can capture risk levels;

- Guo et al. (2011) instead developed classifiers to automatically identify patterns from abstracts to support cancer risk assessment.

Medical records can be used to study cancer and other potential factors. This is because they contain a wealth of information about the diagnosis and treatment of the disease. Studies in this area have been conducted by:

- Ben Abacha et al. (2011), who used a supervised machine learning approach to extract relationships between medical problems, treatments, and tests.
- You et al. (2011), who instead proposed a multiclass classifier to analyze diagnostic syndromes in medical records to improve experience and techniques.
- Lee et al. (2007), who developed a text mining-based system to reveal the relationships between cancer and possible factors. They extracted relationships between diseases and potential factors in medical records using self-organizing maps while using SVM to evaluate them.

In the context of text mining resources for cancer research, an important contribution was made by Thompson et al. (2011), who provided extensive terminology resources for biomedical text mining: it is BioLexicon, which contains over 2.2 million lexical entries, over 1.8 million terminology variants, over 3.3 million semantic relations, including over 2 million synonymy relations. Maqungo et al, on the other hand, developed a database for genes associated with prostate cancer in 2011. The database provides precompiled biomedical text mining information on prostate cancer and also integrates data on molecular interactions, pathways, gene ontologies, gene regulation at the molecular level, and binding sites of transcription factors predicted in prostate cancer promoters implicating genes and transcription factors. HlungDB, an integrated human lung cancer database, contains genes and miRNAs collected through text mining. The database also provides lung cancer-associated networks for further investigation of the molecular mechanisms of lung cancer.

### **3. Text mining and the biology of oncological system**

Systems biology- based networks can be constructed by summarizing associations previously reported in the literature or in various databases. An example is that of Hayasaka et al. (2011), who built a network of genes, genetic diseases, and brain regions based on associations reported in the literature. Sharma et al. in turn collected

a set of known disease-related genes and built an interaction network from the literature, finding 19 genes that were confirmed to be associated with prostate cancer after analysis. These examples illustrate why the use of text mining has become so important to the study of the biology of oncological systems.

The conventional flow of text mining in oncological systems biology research generally involves the following tasks:

- Text acquisition.
- Recognition of bio-entity terms.
- Extraction of complex relationships.
- Discovery of new knowledge.
- Generation of hypotheses in return.

In the previous pages, we have seen how in the general text mining phase of biology of oncological systems, the extracted biomedical text was related to many available sources such as PubMed. Although literature databases provide a compressed data download service, it is common that the included text is not updated in a timely manner and that the amount of available text is limited. However, there are many literature databases that offer an application programming interface through which we can use scripts to automatically download the text from computers: For example, through PubMed E-Utility, users can easily get the updated text.

Thanks to named entity detection tools, it is possible to extract biomedical mentions from the obtained text. Biomedical mentions include terms such as:

- Names of genes.
- Names of proteins.
- mRNA names (RNA message).
- MiRNA (micro RNA) names.
- Terms related to metabolism.
- Cellular terms

Once the biomedical terms are found, metabolic pathways and gene-gene interaction networks can be constructed (in this case, for example, Gene Ontology could be used). Since microRNAs are associated with cancer, one can investigate how miRNAs function in gene-gene interaction.

The next phase will examine how components and structures change in dynamic contexts.

Some networks (e.g., protein-protein interaction networks) and their variations (e.g., metabolic network changes) can be constructed from the text. But there is a problem: networks based on text mining contain a high rate of false negatives. But

the network can still be corrected and optimized thanks to some validation and inference algorithms.

Therefore, at each stage, it is possible to use different resources to validate the network, such as homology, co-expression data, extensive domain data and data on co-biological processes and other information. Later, validation strengthens some nodes and interactions with strong evidence, while removing or updating an incorrect node. In this way, we can develop a protein-protein interactome based on multiple sources of interaction evidence. The networks and components can be used for further studies. Pathway reconstruction plays an important role in understanding molecular mechanisms in cancer. Maps of signalling pathways are generally obtained by manual literature searches, automated text extraction, or canonical pathway databases. Pena-Hernandez et al. implemented an extraction tool to find current gene relationships and signalling pathways from the literature.

### **3.1 Examples of text mining tools**

A variety of instruments are used in cancer research. However, these must be used judiciously to avoid introducing errors into information and findings derived from an unedited text. This happens when the results of text mining tools are not used properly.

Experts can manually edit the text by using different tools. Here are the three most popular tools for the trade:

- Pathway Studio.
- GeneGO.
- Ingenuity.

Pathway Studio allows us to analyse signalling pathways, gene regulatory networks, and protein interaction maps and navigate through molecular networks. The basic knowledge database contains more than 100,000 regulatory events, interactions, and modifications between proteins, cellular processes, and small molecules. It has a natural language processing module, MedScan, that allows Pathway Studio to identify entities and then apply context-free grammar (CFG) rules to extract relationships. Pathway Studio can access the entire PubMed database and online resources, full-text diaries, literature, experimental and electronic notebooks. Many algorithms are available, such as Find Direct Interactions, Find Shorter Paths, Find Common Goals, or Find Common Regulators.

MetaCore, one of the core products of GeneGO, is an integrated knowledge base and software suite for pathway analysis of experimental data and gene lists. The MetaCore knowledge base is a manually edited database derived from extensive

annotation of the full-text literature. GeneGo's MetaMiner, which primarily includes MetaMiner Disease Platforms, MetaMiner Stem Cells, MetaMiner Prostate Cancer, and MetaMiner Cystic Fibrosis, provides knowledge extraction and data analysis platforms for oncology. The main disease reconstruction function is based on three fundamentals: manual annotation of all gene-disease associations, reconstruction of disease trajectories and functional data, and knowledge extraction from published experimental OMIC studies in a disease area. GeneGo also provides APIs for third-party software development.

Ingenuity helps researchers model, analyse, and understand complex biomedical, biological, and chemical systems by integrating data from a variety of experimental platforms. An example of Ingenuity Systems application is the analysis of stem cell-like subpopulations of CD44hi breast cancer using Ingenuity iReport. Ingenuity's background knowledge is also extracted by experts from the full text of the scientific literature, including insights into genes, drugs, biomarkers, chemicals, cellular and pathological processes, and signalling and metabolic pathways. Researchers can search the scientific literature and find the information most relevant to the desired experimental model or question, build dynamic pathway models, and gain confidence in hypotheses and conclusions.

#### **4. Conclusions**

In recent years, tools such as data mining or text mining have gained fundamental importance as they play an important role in the fields of biomedical knowledge and are used to improve the knowledge of diagnostic processes, classify the severity of a disease and determine its prognosis.

#### **References**

- Agarwal, S.; Liu, F.; Yu, H. Simple and efficient machine learning frameworks for identifying protein-protein interaction relevant articles and experimental methods used to study the interactions. *BMC Bioinformatics* 12, S10 (2011). <https://doi.org/10.1186/1471-2105-12-S8-S10>.
- Ai, J.; Smith, B.; Wong, D.T. Saliva Ontology: an ontology-based framework for a Salivaomics Knowledge Base, *BMC Bioinformatics*, 11:302 (2010), doi: 10.1186/1471-2105-11-302.
- Alexopoulos, L. G.; Melas, I. N.; Chairakaki, A. D.; Saez-Rodriguez, J.; Mitsos, A.

- Construction of signaling pathways and identification of drug effects on the liver cancer cell HepG2, 2010 Annual International Conference of the IEEE Engineering in Medicine and Biology, 2010, pp. 6717-6720, doi: 10.1109/IEMBS.2010.5626246.
- Arighi, C.N.; Roberts, P.M.; Agarwal, S. et al. BioCreative III interactive task: an overview. *BMC Bioinformatics* 12, S4 (2011). <https://doi.org/10.1186/1471-2105-12-S8-S4>.
- Barbosa-Silva, A.; Fontaine, J.F.; Donnard, E.R. et al. PESCADOR, a web-based tool to assist text-mining of biointeractions extracted from PubMed queries. *BMC Bioinformatics* 12, 435 (2011). <https://doi.org/10.1186/1471-2105-12-435>.
- Ben Abacha, A.; Zweigenbaum, P. Automatic extraction of semantic relations between medical entities: a rule-based approach. *J Biomed Semant* 2, S4 (2011). <https://doi.org/10.1186/2041-1480-2-S5-S4>.
- Chang, J.T.; Schutze, H.; Altman, R.B. *Journal of the American Medical Informatics Association*, Volume 9, Issue 6, November 2002, Pages 612–620, <https://doi.org/10.1197/jamia.M1139>.
- Chun, H.W.; Tsuruoka, Y.; Kim, J.D.; Shiba, R.; Nagata, N.; Hishiki, T. et al. Extraction of gene–disease relations from Medline using domain dictionaries and machine learning. In: *Citeseer*, 2006, pp. 4–15, [https://doi.org/10.1142/9789812701626\\_0002](https://doi.org/10.1142/9789812701626_0002).
- Chun, H.W.; Tsuruoka, Y.; Kim, J.D. et al. Automatic recognition of topic-classified relations between prostate cancer and genes using MEDLINE abstracts. *BMC Bioinformatics* 7, S4 (2006). <https://doi.org/10.1186/1471-2105-7-S3-S4>.
- Clancy, T.; Pedicini, M.; Castiglione, F. et al. Immunological network signatures of cancer progression and survival. *BMC Med Genomics* 4, 28 (2011). <https://doi.org/10.1186/1755-8794-4-28>.
- Dai, H.J.; Chang, Y.C.; Tzong-Han Tsai, R. et al. New Challenges for Biological Text-Mining in the Next Decade. *J. Comput. Sci. Technol.* 25, 169–179 (2010). <https://doi.org/10.1007/s11390-010-9313-5>.
- Deng, X.; Geng, H.; Bastola, D. R.; Ali, H. H. Link test—A statistical method for finding prostate cancer biomarkers, *Computational Biology and Chemistry*, Volume 30, Issue 6, 2006, Pages 425-433, ISSN 1476-9271, <https://doi.org/10.1016/j.compbiolchem.2006.09.002>.
- Eskin, E.; Agichtein, E. Combining text mining and sequence analysis to discover protein functional regions, *Biocomputing* (2004), pp. 288-299, [https://doi.org/10.1142/9789812704856\\_0028](https://doi.org/10.1142/9789812704856_0028).
- Fang, Y.C.; Huang, H.C.; Juan, H.F. MeInfoText: associated gene methylation and cancer information from text mining. *BMC Bioinformatics* 9, 22 (2008). <https://doi.org/10.1186/1471-2105-9-22>.
- Guo, Y.; Korhonen, A.; Liakata, M. et al. A comparison and user-based evaluation

- of models of textual information structure in the context of cancer risk assessment. *BMC Bioinformatics* 12, 69 (2011). <https://doi.org/10.1186/1471-2105-12-69>.
- Hayasaka, S.; Hugenschmidt, C.E.; Laurienti P.J. A network of genes, genetic disorders, and brain areas, *PLoS One*, 6 (2011), p. e20907, <https://doi.org/10.1371/journal.pone.0020907>.
- Jiménez-Marín, Á.; Collado-Romero, M.; Ramirez-Boo, M. et al. Biological pathway analysis by ArrayUnlock and Ingenuity Pathway Analysis. *BMC Proc* 3, S6 (2009). <https://doi.org/10.1186/1753-6561-3-S4-S6>.
- Kazama, J.; Makino, T.; Ohta, Y.; Tsujii, J. Tuning support vector machines for biomedical named entity recognition. In: Association for computational linguistics Morristown, NJ, USA; 2002. pp. 1–8.
- Kolluru, B.; Nakjang, Sirintra, Hirt, Robert P., Wipat, Anil and Ananiadou, Sophia. Automatic extraction of microorganisms and their habitats from free text using text mining workflows *Journal of Integrative Bioinformatics*, vol. 8, no. 2, 2011, pp. 176-186. <https://doi.org/10.1515/jib-2011-184>.
- Korhonen, A.; Silins, I.; Sun, L. et al. The first step in the development of text mining technology for cancer risk assessment: identifying and organizing scientific evidence in risk assessment literature. *BMC Bioinformatics* 10, 303 (2009). <https://doi.org/10.1186/1471-2105-10-303>.
- Krallinger M.; Leitner F.; Valencia A. Analysis of biological processes and diseases using text mining approaches, *Methods Mol Biol*, 593 (2010), pp. 341-382.
- Krallinger, M.; Vazquez, M.; Leitner, F.; Salgado, D.; Chatr-Aryamontri, A.; Winter, A.; Perfetto, L.; Briganti, L.; Licata, L.; Iannuccelli, M.; Castagnoli, L.; Cesareni, G.; Tyers, M.; Schneider, G.; Rinaldi, F.; Leaman, R.; Gonzalez, G.; Matos, S.; Kim, S.; Wilbur, W.J.; Rocha, L.; Shatkay, H.; Tendulkar, A.V.; Agarwal, S.; Liu, F.; Wang, X.; Rak, R.; Noto, K.; Elkan, C.; Lu, Z.; Dogan, R.I.; Fontaine, J.F.; Andrade-Navarro, M.A.; Valencia, A. The protein-protein interaction tasks of BioCreative III: classification/ranking of articles and linking bio-ontology concepts to full text. *BMC Bioinformatics*. 2011; 12(8): S3. doi: 10.1186/1471-2105-12-S8-S3.
- Kuo, C.J.; Ling, M.H.; Lin, K.T. et al. BIOADI: a machine learning approach to identifying abbreviations and definitions in biological literature. *BMC Bioinformatics* 10, S7 (2009). <https://doi.org/10.1186/1471-2105-10-S15-S7>.
- Lee, C.; Wu, C.; Yang, H. Text Mining of Clinical Records for Cancer Diagnosis, Second International Conference on Innovative Computing, Information and Control (ICICIC 2007), 2007, pp. 172-172, doi: 10.1109/ICICIC.2007.556.
- Li, X.; Cai, H.; Xu, J. et al. A mouse protein interactome through combined literature mining with multiple sources of interaction evidence. *Amino Acids* 38, 1237–1252 (2010). <https://doi.org/10.1007/s00726-009-0335-7>.
- Lin, Y.F.; Tsai, T.H.; Chou, W.C.; Wu, K.P.; Sung, T.Y.; Hsu, W.L. A maximum

- entropy approach to biomedical named entity recognition. In: The 4th ACM SIGKDD workshop on data mining in bioinformatics; 2004, pp. 56–61.
- Liu, H.; Friedman, C. Mining terminological knowledge in large biomedical corpora, *Biocomputing* (2002), pp. 415–426, [https://doi.org/10.1142/9789812776303\\_0039](https://doi.org/10.1142/9789812776303_0039).
- Maier, D.; Kalus, W.; Wolff, M.; Kalko, S.G.; Roca, J.; Marin de Mas, I. et al., Knowledge management for systems biology a general and visually driven framework applied to translational medicine, *BMC Syst Biol*, 5 (2011), p. 5–38, doi: 10.1186/1752-0509-5-38.
- Maungo, M.; Kaur, M.; Kwofie, S. K.; Radovanovic, A. et al. DDPC: Dragon Database of Genes associated with Prostate Cancer, *Nucleic Acids Research*, Volume 39, Issue suppl\_1, 1 January 2011, Pages D980–D985, <https://doi.org/10.1093/nar/gkq849>.
- Matos, S.; Arrais, J.P.; Maia-Rodrigues, J.; Oliveira, J.L. Concept-based query expansion for retrieving gene related publications from MEDLINE, *BMC Bioinformatics*, 11, 212 (2010), <https://doi.org/10.1186/1471-2105-11-212>.
- Natarajan, J.; Berrar, D.; Dubitzky, W. et al. Text mining of full-text journal articles combined with gene expression analysis reveals a relationship between sphingosine-1-phosphate and invasiveness of a glioblastoma cell line. *BMC Bioinformatics* 7, 373 (2006). <https://doi.org/10.1186/1471-2105-7-373>.
- Nikitin, A.; Egorov, S.; Daraselia, N.; Mazo, I. Pathway studio-the analysis and navigation of molecular networks, *Bioinformatics*, Volume 19, Issue 16, 1 November 2003, Pages 2155–2157, <https://doi.org/10.1093/bioinformatics/btg290>
- Ongenaert, M.; Van Neste, L.; De Meyer, T.; Menschaert, G.; Bekaert, S.; Van Crielinge, W. PubMeth: a cancer methylation database combining text-mining and expert annotation, *Nucleic Acids Research*, Volume 36, Issue suppl\_1, 1 January 2008, Pages D842–D846, <https://doi.org/10.1093/nar/gkm788>
- Palakal M.; Bright J.; Sebastian T.; Hartanto S. A comparative study of cells in inflammation, EAE and MS using biomedical literature data mining, *J Biomed Sci*, 14 (2007), pp. 67–85, DOI: 10.1007/s11373-006-9120-8.
- Pena-Hernandez; K.E.; Mahamaneerat, W.K.; Kobayashi, T.; Shyu, C.R.; Arthur, G.; Caldwell, C.W. Mapping biomedical literature with WNT signaling pathway. In: *AMIA annu symp proc*; 2008, p. 1089.
- Rebholz-Schuhmann, D.; Yepes, A.J.; Li, S.; Kafkas, C.; Lewin, I.; Kang, N. et al., Assessment of NER solutions against the first and second CALBC Silver Standard Corpus, *J Biomed Semantics*, 2, S11 (Suppl. 5) (2011), <https://doi.org/10.1186/2041-1480-2-S5-S11>.
- Sharma, P.; Senthilkumar, R.; Brahmachari, V. et al. Mining literature for a comprehensive pathway analysis: A case study for retrieval of homocysteine related genes for genetic and epigenetic studies. *Lipids Health Dis* 5, 1 (2006). <https://doi.org/10.1186/1476-511X-5-1>.

- Shetty, K.D.; Dalal, S.R. Using information mining of the medical literature to improve drug safety, *J Am Med Inform Assoc*, 18 (2011), pp. 668-674, <https://doi.org/10.1136/amiajnl-2011-000096>.
- Srinivasan, P.; Wedemeyer, M.; Mining concept profiles with the vector model or where on earth are diseases being studied. In: Citeseer; 2003.
- Thompson, P.; McNaught, J.; Montemagni, S. et al. The BioLexicon: a large-scale terminological resource for biomedical text mining. *BMC Bioinformatics* 12, 397 (2011). <https://doi.org/10.1186/1471-2105-12-397>.
- Tsai F.S., Text mining and visualisation of Protein–Protein Interactions, *Int J Comput Biol Drug Des*, 4 (2011), pp. 239-244, doi: 10.1504/IJCBDD.2011.041412.
- Tsai, T.; Chou, W.C.; Wu, S.H.; Sung, T.Y.; Hsiang, J.; Hsu, W.L. Integrating linguistic knowledge into a conditional random field framework to identify biomedical named entities, *Expert Syst Appl*, 30 (2006), pp. 117-128, <https://doi.org/10.1016/j.eswa.2005.09.072>.
- Wang, L.; Xiong, Y.; Sun, Y.; Fang, Z.; Li, L.; Ji, H.; Tielu Shi, T. HLungDB: an integrated database of human lung cancer research, *Nucleic Acids Research*, Volume 38, Issue suppl 1, 1 January 2010, Pages D665–D669, <https://doi.org/10.1093/nar/gkp945>.
- Xu, Y.; Teng, D.; Lei, Y. MinePhos: A Literature Mining System for Protein Phosphorylation Information Extraction, in *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 9, no. 1, pp. 311-315, Jan.-Feb. 2012, doi: 10.1109/TCBB.2011.85.
- You M.; Zhao R.W.; Li G.Z.; Hu X. MAPLSC: a novel multi-class classifier for medical diagnosis, *Int J Data Min Bioinform*, 5 (2011), pp. 383-401, <https://doi.org/10.1504/IJDMB.2011.041555>.
- Yu, H.; Hripcsak, G.; Friedman, C. Mapping abbreviations to full forms in biomedical articles, *Journal of American Medical Informatics Association*, 9 (2002), pp. 262-272. <https://doi.org/10.1197/jamia.M0913>.
- Zhou, G.D.; Su, J. Exploring deep knowledge resources in biomedical name recognition. In: *JNLPBA; 2004*, pp. 96–99.
- Zhu, F.; Patumcharoenpol, P.; Zhang, C.; Yang, Y.; Chan, J.; Meechai, A.; Vongsangnak, W.; Shen, B. Biomedical text mining and its applications in cancer research, *Journal of Biomedical Informatics*, Volume 46, Issue 2, 2013, Pages 200-211, ISSN 1532-0464, <https://doi.org/10.1016/j.jbi.2012.10.007>.

# Alcune considerazioni sulla presenza in Italia dei Minori Stranieri Non Accompagnati (MSNA) in relazione alla pandemia Covid-19

Ilaria Pepe<sup>1</sup>, Ernesto Toma<sup>2\*</sup>.

<sup>1</sup> Dipartimento di Scienze Mediche di base, Neuroscienze e Organi di senso, Università degli studi di Bari Aldo Moro, <sup>2</sup> Dipartimento di Economia e finanza, Università degli studi di Bari Aldo Moro

**Riassunto:** Attraverso i dati mensili dei Report statistici sulla presenza dei Minori Stranieri Non Accompagnati (MSNA), forniti dal Ministero del lavoro e delle politiche sociali, si analizza il variare del fenomeno dell'insediamento in Italia degli stessi dal momento di scoppio della pandemia (qui individuato in marzo 2020, momento di inizio del lockdown) rispetto ai circa due anni precedenti (da gennaio 2018). L'obiettivo è quello di verificare se le difficoltà legate al Covid-19 abbiano, ed eventualmente in che modo, influenzato i flussi e la presenza di MSNA in Italia. Saranno prese in considerazione alcune caratteristiche come il genere, l'età, la nazione di provenienza e la regione di accoglienza di tali minori.

**Keywords:** Minori stranieri non accompagnati (MSNA), pandemia.

## 1. Introduzione

*“Ciao, è stata dura ma finalmente possiamo ripartire. Come prima”.*

Questo è lo slogan di una pubblicità che nel periodo post-pandemico lasciava intravedere la possibilità di tornare ad una vita “normale”, ed è la frase che ha colpito Riccardo Noury, portavoce di Amnesty International Italia, e Luca Leone, direttore editoriale e co-fondatore della casa editrice Infinito che hanno curato un saggio dal titolo “I dimenticati”<sup>1</sup>, edito a pochi mesi dallo scoppio in Italia della

---

\* Autore corrispondente: [ernesto.toma@uniba.it](mailto:ernesto.toma@uniba.it).

Il presente lavoro è frutto di un progetto comune ed i due autori hanno contribuito in equal misura alla stesura dello stesso.

<sup>1</sup> <https://www.allonsanfan.it/2020/08/17/dimenticati-covid-amnesty/>

Pandemia.

La ripartenza di cui parla lo slogan, dopo mesi di lockdown, doveva riguardare (e in effetti ha riguardato) tutte le sfere del sociale, spesso intrinsecamente correlate tra loro: il lavoro, l'economia, i rapporti umani, il benessere psico-sociale. Tra queste, degna di attenzione è la questione della presenza straniera in Italia, motivo per cui il focus di questa breve nota riguarderà l'eventuale incidenza di questo periodo di pandemia sull'evoluzione del flusso migratorio dai paesi stranieri verso l'Italia, con particolare riferimento alla presenza nel nostro Paese di una particolare categoria, i cosiddetti Minori Stranieri Non Accompagnati (da ora MSNA).

Ai sensi della l. n. 47/2017, il Minore Straniero Non Accompagnato è “il minore privo di assistenza e di rappresentanza da parte dei genitori o di altri adulti per lui legalmente responsabili in base alle leggi vigenti nell'ordinamento italiano” e successivamente, l'ordinanza n. 9199 del 3 aprile 2019 della VI sez. Civile della Corte di Cassazione, ha chiarito le condizioni che debbano sussistere perché sia soddisfatta la definizione di “minore straniero non accompagnato”, di cui sopra, specificando come sia necessaria la compresenza di entrambe le condizioni menzionate: l'assenza di soggetti che abbiano la cura, la custodia e assicurino il benessere del minore nonché la mancanza di soggetti che possano rappresentarlo e siano formalmente responsabili per lui.

L'obiettivo del presente lavoro è quello di verificare se lo scoppio della pandemia, con tutte le problematiche da essa discese, abbia influito sulla presenza di MSNA sul suolo italiano e se le eventuali variazioni siano correlate a talune caratteristiche socio-demografiche dei minori rilevate dalle fonti ufficiali ministeriali.

## 2. Materiali e metodi

Di seguito saranno esaminati i dati relativi alla presenza degli MSNA dal gennaio 2018 ad ottobre 2021, con particolare attenzione al confronto tra i dati sino a febbraio 2020, quando ancora la pandemia era in stato embrionale, e quelli successivi a tale data, quindi da marzo 2020, mese di esplosione del virus con conseguente lockdown in tutta Italia.

I dati, desunti dai *Report statistici sulla presenza dei MSNA in Italia*<sup>2</sup>, che il Ministero del lavoro e delle politiche sociali pubblica online mensilmente, consentiranno pertanto di fare il punto sulla presenza dei MSNA non solo nel complesso,

---

<sup>2</sup> Per i dati raccolti cfr.: <https://www.lavoro.gov.it/temi-e-priorita/immigrazione/focus-on/minori-stranieri/Pagine/Dati-minori-stranieri-non-accompagnati.aspx>

ma anche rispetto al genere, all'età, alla cittadinanza e alla regione di accoglienza. Queste ultime 3 caratteristiche, inoltre, saranno alla fine brevemente analizzate con riferimento ai soli dati delle MSNA di genere femminile.

### **3. La presenza di MSNA per genere**

Come evidenziato nel Report di Monitoraggio del 30 giugno 2018 della Direzione Generale dell'immigrazione e delle politiche di integrazione – Div. II del Ministero del Lavoro e delle Politiche sociali, la presenza di MSNA in Italia era già in calo rispetto al 2017, con un decremento di circa il 26% tra i mesi di giugno dei 2 anni (13.151 MSNA al 30 giugno 2018 rispetto ai 17.864 del giugno 2017) e tale fenomeno era rilevabile anche, in generale, su tutti gli stranieri immigrati<sup>3</sup>.

Come si può evincere dalla Tabella 1, la diminuzione resta costante per tutto il 2018 e il 2019, assestandosi nei primi mesi del 2020, periodo in cui si è maggiormente sviluppata in Italia la pandemia da Covid-19. Dai quasi 15.000 MSNA presenti nel gennaio 2018, nell'arco di circa 2 anni, si è giunti ad un valore pari a poco più di un terzo, con le circa 5.000 unità di marzo-giugno 2020 (periodo del lockdown totale o parziale in Italia). Da luglio 2020, il trend si è nettamente invertito e la crescita costante dei MSNA, ad ottobre 2021, ha portato in Italia ad un raddoppio delle presenze, con oltre 10 mila unità. Su questo incremento potrebbe aver influito la Circolare del Ministero dell'Interno n. 3393 del 18 marzo 2020 che, per evitare la diffusione e il contagio da virus, ha disposto la prosecuzione “anche a coloro che non hanno più titolo a permanere nei centri” consentendo anche di riammettere nei centri di accoglienza le persone a cui è stata revocata per vari motivi.

Rispetto al genere, la presenza di MSNA femminile è sempre molto ridotta rispetto a quella maschile: in v.a., ad inizio 2018 si contavano 1.101 femmine (rispetto ai quasi 14.000 maschi) con una incidenza percentuale del 7,4%. Tale presenza femminile diminuisce sensibilmente fino al periodo di inizio pandemia, quando si toccano le 260 unità; durante tutto il periodo pandemico, fino ad oggi, la quota assoluta femminile è variata mensilmente di poche unità e attualmente presenta gli stessi valori di inizio pandemia. La componente maschile è, invece, raddoppiata nel periodo pandemico, passando da circa 5 mila unità di marzo 2020 alle oltre 10 mila di ottobre 2021. L'incidenza percentuale della componente femminile, pertanto, ha continuato a decrescere raggiungendo il 2,5%.

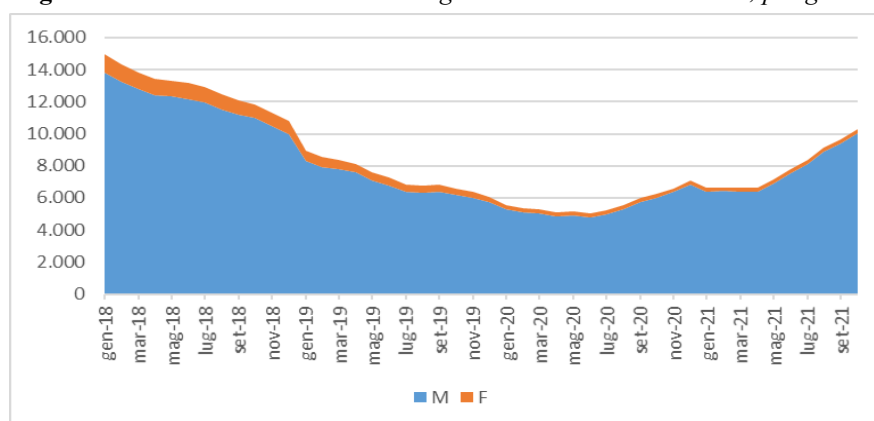
---

<sup>3</sup> Cfr. ISTAT, Statistiche Report, 26 ottobre 2020

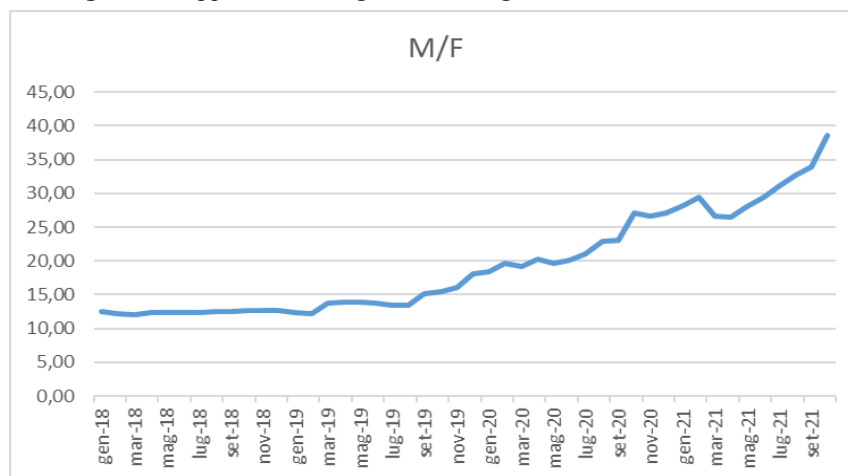
**Tabella 1.** Presenza in v.a. di MSNA da gennaio 2018 a ottobre 2021, per genere

| Periodo<br>pre covid | Genere |       | MF            | % F su<br>MF | Periodo<br>covid | Genere |     | MF            | % F su<br>MF |
|----------------------|--------|-------|---------------|--------------|------------------|--------|-----|---------------|--------------|
|                      | M      | F     |               |              |                  | M      | F   |               |              |
| <b>2018</b>          |        |       |               |              |                  |        |     |               |              |
| Gennaio              | 13.838 | 1.101 | <b>14.939</b> | 7,4          |                  |        |     |               |              |
| Febbraio             | 13.246 | 1.092 | <b>14.338</b> | 7,6          | <b>2020</b>      |        |     |               |              |
| Marzo                | 12.781 | 1.057 | <b>13.838</b> | 7,6          | Marzo            | 5.004  | 261 | <b>5.265</b>  | 5,0          |
| Aprile               | 12.414 | 1.006 | <b>13.420</b> | 7,5          | Aprile           | 4.870  | 241 | <b>5.111</b>  | 4,7          |
| Maggio               | 12.324 | 994   | <b>13.318</b> | 7,5          | Maggio           | 4.907  | 249 | <b>5.156</b>  | 4,8          |
| Giugno               | 12.169 | 982   | <b>13.151</b> | 7,5          | Giugno           | 4.778  | 238 | <b>5.016</b>  | 4,7          |
| Luglio               | 11.966 | 964   | <b>12.930</b> | 7,5          | Luglio           | 4.966  | 236 | <b>5.202</b>  | 4,5          |
| Agosto               | 11.539 | 918   | <b>12.457</b> | 7,4          | Agosto           | 5.308  | 232 | <b>5.540</b>  | 4,2          |
| Settembre            | 11.214 | 898   | <b>12.112</b> | 7,4          | Settembre        | 5.730  | 249 | <b>5.979</b>  | 4,2          |
| Ottobre              | 10.967 | 871   | <b>11.838</b> | 7,4          | Ottobre          | 6.006  | 221 | <b>6.227</b>  | 3,5          |
| Novembre             | 10.506 | 833   | <b>11.339</b> | 7,3          | Novembre         | 6.362  | 239 | <b>6.601</b>  | 3,6          |
| Dicembre             | 10.000 | 787   | <b>10.787</b> | 7,3          | Dicembre         | 6.828  | 252 | <b>7.080</b>  | 3,6          |
| <b>2019</b>          |        |       |               |              | <b>2021</b>      |        |     |               |              |
| Gennaio              | 8.300  | 671   | <b>8.971</b>  | 7,5          | Gennaio          | 6.407  | 227 | <b>6.634</b>  | 3,4          |
| Febbraio             | 7.890  | 647   | <b>8.537</b>  | 7,6          | Febbraio         | 6.414  | 218 | <b>6.632</b>  | 3,3          |
| Marzo                | 7.774  | 568   | <b>8.342</b>  | 6,8          | Marzo            | 6.373  | 239 | <b>6.612</b>  | 3,6          |
| Aprile               | 7.583  | 548   | <b>8.131</b>  | 6,7          | Aprile           | 6.392  | 241 | <b>6.633</b>  | 3,6          |
| Maggio               | 7.072  | 508   | <b>7.580</b>  | 6,7          | Maggio           | 6.913  | 246 | <b>7.159</b>  | 3,4          |
| Giugno               | 6.781  | 491   | <b>7.272</b>  | 6,8          | Giugno           | 7.546  | 256 | <b>7.802</b>  | 3,3          |
| Luglio               | 6.354  | 472   | <b>6.826</b>  | 6,9          | Luglio           | 8.121  | 261 | <b>8.382</b>  | 3,1          |
| Agosto               | 6.306  | 469   | <b>6.775</b>  | 6,9          | Agosto           | 8.860  | 271 | <b>9.131</b>  | 3,0          |
| Settembre            | 6.376  | 422   | <b>6.798</b>  | 6,2          | Settembre        | 9.385  | 276 | <b>9.661</b>  | 2,9          |
| Ottobre              | 6.168  | 398   | <b>6.566</b>  | 6,1          | Ottobre          | 10.056 | 261 | <b>10.317</b> | 2,5          |
| Novembre             | 5.996  | 373   | <b>6.369</b>  | 5,9          |                  |        |     |               |              |
| Dicembre             | 5.737  | 317   | <b>6.054</b>  | 5,2          |                  |        |     |               |              |
| <b>2020</b>          |        |       |               |              |                  |        |     |               |              |
| Gennaio              | 5.278  | 287   | <b>5.565</b>  | 5,2          |                  |        |     |               |              |
| Febbraio             | 5.108  | 260   | <b>5.368</b>  | 4,8          |                  |        |     |               |              |

Fonte: nostre elaborazioni su dati provenienti dai Report statistici sulla presenza dei MSNA in Italia.

**Figura 1.** Presenza in v.a. di MSNA da gennaio 2018 a ottobre 2021, per genere

**Figura 2.** Rapporto M/F negli MSNA da gennaio 2018 a ottobre 2021



Il rapporto dei sessi ( $M/F \cdot 100$ ), illustrato nella Fig.2, evidenzia questo cambiamento importante avvenuto proprio in coincidenza dell'insorgere della pandemia. Da poco più di 12 maschi per ogni MSNA di genere femminile del gennaio 2018, si è passati agli attuali quasi 40 e la crescita di questa proporzione è costante e continua (tranne un breve periodo tra marzo e maggio 2021).

Cosa ha però spinto maggiormente, soprattutto i MSNA maschi, a riprendere questi "viaggi della speranza" verso l'Italia? Varie sono le ipotesi che si possono avanzare, spesso dipendenti soprattutto dalle condizioni dei paesi di origine di questi minori; tra queste, molto verosimilmente si può ipotizzare l'aggravarsi delle condizioni economiche nei paesi di origine, da sempre uno dei motivi principali di questi flussi migratori; per le donne, invece, spesso sono emerse caratteristiche diverse nella storia migratoria, perché quasi sempre vittime di tratta e di sfruttamento sessuale.

#### 4. MSNA per età

Un'altra caratteristica che viene riportata nei Report forniti mensilmente dal Ministero del lavoro e delle politiche sociali è quella dell'età dei MSNA presenti sul territorio italiano. Enormi sono talvolta le difficoltà, in tutti i casi di mancanza di adeguata documentazione, nella identificazione dell'età anagrafica dei minori oggetto

di studio. Su questa problematica è in particolare intervenuta la Legge Zampa del 2017 e a cui si rimanda per approfondimenti<sup>4</sup>.

**Tabella 2.** *Distribuzione percentuale di MSNA dal gennaio 2018 a ottobre 2021, per età*

| Periodo<br>pre covid | 17   | 16   | Età<br>15 | 7-14 | <7  | MF  | Periodo<br>covid | 17   | 16   | Età<br>15 | 7-14 | <7  | MF  |
|----------------------|------|------|-----------|------|-----|-----|------------------|------|------|-----------|------|-----|-----|
| <b>2018</b>          |      |      |           |      |     |     |                  |      |      |           |      |     |     |
| Gennaio              | 59,9 | 24,2 | 9,1       | 6,0  | 0,8 | 100 |                  |      |      |           |      |     |     |
| Febbraio             | 59,4 | 24,6 | 9,0       | 6,3  | 0,8 | 100 | <b>2020</b>      |      |      |           |      |     |     |
| Marzo                | 59,6 | 24,6 | 8,9       | 6,1  | 0,8 | 100 | Marzo            | 62,3 | 25,4 | 7,4       | 4,7  | 0,2 | 100 |
| Aprile               | 59,5 | 24,6 | 8,8       | 6,3  | 0,8 | 100 | Aprile           | 62,4 | 25,4 | 7,2       | 4,7  | 0,3 | 100 |
| Maggio               | 58,9 | 25,0 | 8,9       | 6,5  | 0,8 | 100 | Maggio           | 62,9 | 24,5 | 7,4       | 4,8  | 0,4 | 100 |
| Giugno               | 58,7 | 25,2 | 8,9       | 6,4  | 0,8 | 100 | Giugno           | 63,8 | 23,8 | 7,2       | 4,8  | 0,4 | 100 |
| Luglio               | 58,6 | 25,4 | 8,8       | 6,4  | 0,8 | 100 | Luglio           | 62,9 | 24,3 | 7,6       | 4,6  | 0,7 | 100 |
| Agosto               | 58,9 | 25,2 | 8,7       | 6,2  | 1,0 | 100 | Agosto           | 64,0 | 23,8 | 7,3       | 4,6  | 0,4 | 100 |
| Settembre            | 56,7 | 26,6 | 9,3       | 6,6  | 0,8 | 100 | Settembre        | 64,2 | 23,9 | 7,2       | 4,4  | 0,4 | 100 |
| Ottobre              | 59,6 | 24,8 | 8,5       | 6,3  | 0,8 | 100 | Ottobre          | 65,1 | 22,9 | 7,1       | 4,6  | 0,3 | 100 |
| Novembre             | 59,9 | 25,0 | 8,1       | 6,2  | 0,8 | 100 | Novembre         | 66,2 | 21,9 | 6,9       | 4,7  | 0,4 | 100 |
| Dicembre             | 60,2 | 24,8 | 8,0       | 6,2  | 0,8 | 100 | Dicembre         | 66,9 | 21,8 | 6,6       | 4,4  | 0,4 | 100 |
| <b>2019</b>          |      |      |           |      |     |     | <b>2021</b>      |      |      |           |      |     |     |
| Gennaio              | 61,2 | 23,7 | 7,7       | 6,4  | 0,9 | 100 | Gennaio          | 67,9 | 21,2 | 6,4       | 4,1  | 0,4 | 100 |
| Febbraio             | 61,3 | 23,5 | 7,6       | 6,7  | 0,8 | 100 | Febbraio         | 66,4 | 21,7 | 7,0       | 4,5  | 0,3 | 100 |
| Marzo                | 63,1 | 23,4 | 7,4       | 5,3  | 0,7 | 100 | Marzo            | 65,0 | 22,5 | 7,4       | 4,7  | 0,4 | 100 |
| Aprile               | 63,6 | 22,6 | 7,5       | 5,3  | 0,9 | 100 | Aprile           | 64,7 | 22,5 | 7,5       | 5,1  | 0,3 | 100 |
| Maggio               | 63,6 | 22,5 | 7,3       | 5,7  | 1,0 | 100 | Maggio           | 64,7 | 22,9 | 7,2       | 4,9  | 0,3 | 100 |
| Giugno               | 63,6 | 22,7 | 7,1       | 5,7  | 0,9 | 100 | Giugno           | 64,2 | 23,2 | 7,5       | 4,7  | 0,4 | 100 |
| Luglio               | 63,7 | 22,6 | 6,8       | 5,9  | 1,0 | 100 | Luglio           | 64,4 | 22,8 | 7,9       | 4,6  | 0,3 | 100 |
| Agosto               | 61,7 | 23,6 | 7,6       | 6,1  | 1,1 | 100 | Agosto           | 63,8 | 23,0 | 8,1       | 4,8  | 0,2 | 100 |
| Settembre            | 61,4 | 24,3 | 7,8       | 5,7  | 0,9 | 100 | Settembre        | 63,8 | 23,2 | 8,0       | 4,8  | 0,2 | 100 |
| Ottobre              | 61,5 | 24,6 | 7,8       | 5,3  | 0,8 | 100 | Ottobre          | 64,7 | 23,2 | 7,4       | 4,7  | 0,1 | 100 |
| Novembre             | 61,6 | 25,0 | 7,6       | 5,0  | 0,7 | 100 |                  |      |      |           |      |     |     |
| Dicembre             | 61,5 | 26,1 | 7,2       | 4,5  | 0,7 | 100 |                  |      |      |           |      |     |     |
| <b>2020</b>          |      |      |           |      |     |     |                  |      |      |           |      |     |     |
| Gennaio              | 62,3 | 25,4 | 7,4       | 4,7  | 0,2 | 100 |                  |      |      |           |      |     |     |
| Febbraio             | 62,4 | 25,4 | 7,2       | 4,7  | 0,3 | 100 |                  |      |      |           |      |     |     |

Fonte: nostre elaborazioni su dati provenienti dai *Report statistici sulla presenza dei MSNA in Italia*.

Dalla Tabella 2 si evince che nel corso del 2020 e del 2021 coloro che hanno da 0 a 6 anni sono passati da circa l'1% del periodo pre-covid al solo 0,1% dell'ultimo dato di ottobre 2021. Anche le altre classi di età, 7-14enni, 15enni, 16enni vedono una seppur lieve diminuzione e solo il gruppo di 17enni aumenta di circa 5 punti percentuali. Nel corso del 2018, questi ultimi sono quasi sempre al di sotto del 60% (tranne a dicembre, con il 60,2) mentre, soprattutto verso la fine del 2020-inizio

<sup>4</sup> Cfr., ad es., [https://www.retesai.it/wp-content/uploads/2016/11/FaqAccertamento-et%3%a0\\_rev-legge-Zampa\\_def.pdf](https://www.retesai.it/wp-content/uploads/2016/11/FaqAccertamento-et%3%a0_rev-legge-Zampa_def.pdf)

2021, si nota un innalzamento della percentuale di questa classe fino al quasi 68% di gennaio per poi giungere all'attuale 64,7%. Si è ovviamente consapevoli che l'analisi della presenza dei MSNA per età può essere chiaramente inficiata dal naturale passaggio degli stessi minori da una età alla successiva e, poiché azzardato procedere ad interpretazioni dei trend, si lascerà spazio solo ad una oggettiva descrizione dell'andamento dei dati.

## 5. MSNA per nazionalità

Nella Tabella 3 sono riportate le presenze dei MSNA per nazionalità di provenienza.

**Tabella 3.** *Presenza in v.a. di MSNA rispetto a taluni periodi pre e post covid, da gennaio 2018 a ottobre 2021, per le nazionalità maggiormente presenti nei due periodi*

| Nazionalità di provenienza | Inizio e fine periodo pre covid preso in considerazione |                           | Presenza media mensile | Inizio e fine periodo covid preso in considerazione |                          | Presenza media mensile |
|----------------------------|---------------------------------------------------------|---------------------------|------------------------|-----------------------------------------------------|--------------------------|------------------------|
|                            | Gennaio 2018<br>v.a. e %                                | Febbraio 2020<br>v.a. e % |                        | Marzo 2020<br>v.a. e %                              | Ottobre 2021<br>v.a. e % |                        |
| Afghanistan                | 277 (1,9%)                                              | 70 (1,3%)                 | <b>203</b>             | 86 (1,6%)                                           | 280 (2,7%)               | <b>172</b>             |
| Albania                    | 1.658 (11,1%)                                           | 1.635 (30,5%)             | <b>1.601</b>           | 1.597 (30,3%)                                       | 1.157 (11,2%)            | <b>1.057</b>           |
| Bangladesh                 | 605 (4,1%)                                              | 501 (9,3%)                | <b>358</b>             | 494 (9,4%)                                          | 2.694 (26,1%)            | <b>1.428</b>           |
| Costa d'Avorio             | 1.189 (8,0%)                                            | 231 (4,3%)                | <b>704</b>             | 218 (4,1%)                                          | 343 (3,3%)               | <b>276</b>             |
| Egitto                     | 1.524 (10,2%)                                           | 565 (10,5%)               | <b>889</b>             | 584 (11,1%)                                         | 1.364 (13,2%)            | <b>710</b>             |
| Eritrea                    | 1.092 (7,3%)                                            | 68 (1,3%)                 | <b>572</b>             | 65 (1,2%)                                           | 199 (1,9%)               | <b>108</b>             |
| Gambia                     | 1.712 (11,5%)                                           | 159 (3,0%)                | <b>832</b>             | 143 (2,7%)                                          | 157 (1,5%)               | <b>115</b>             |
| Ghana                      | 291 (2,0%)                                              | 39 (0,7%)                 | <b>135</b>             | 34 (0,6%)                                           | 55 (0,5%)                | <b>41</b>              |
| Guinea                     | 1.305 (8,7%)                                            | 186 (3,5%)                | <b>684</b>             | 187 (3,5%)                                          | 304 (3,0%)               | <b>258</b>             |
| Kosovo                     | 309 (2,1%)                                              | 316 (5,9%)                | <b>324</b>             | 287 (5,5%)                                          | 165 (1,6%)               | <b>188</b>             |
| Mali                       | 735 (4,9%)                                              | 109 (2,0%)                | <b>460</b>             | 104 (2,0%)                                          | 172 (1,7%)               | <b>132</b>             |
| Marocco                    | 230 (1,5%)                                              | 156 (2,9%)                | <b>199</b>             | 154 (2,9%)                                          | 141 (1,4%)               | <b>132</b>             |
| Nigeria                    | 1.102 (7,4%)                                            | 116 (2,2%)                | <b>557</b>             | 110 (2,1%)                                          | 64 (0,6%)                | <b>71</b>              |
| Pakistan                   | 368 (2,5%)                                              | 401 (7,5%)                | <b>472</b>             | 369 (7,0%)                                          | 659 (6,4%)               | <b>490</b>             |
| Senegal                    | 746 (5,0%)                                              | 187 (3,5%)                | <b>453</b>             | 176 (3,3%)                                          | 114 (1,1%)               | <b>130</b>             |
| Somalia                    | 602 (4,0%)                                              | 104 (1,9%)                | <b>352</b>             | 129 (2,5%)                                          | 310 (3,0%)               | <b>246</b>             |
| Tunisia                    | 281 (1,9%)                                              | 247 (4,6%)                | <b>324</b>             | 238 (4,5%)                                          | 1.615 (15,7%)            | <b>872</b>             |
| ....                       |                                                         |                           |                        |                                                     |                          |                        |
| Altre <sup>5</sup>         | 913 (6,1%)                                              | 278 (5,2%)                | <b>624</b>             | 290 (5,5%)                                          | 524 (5,1%)               | <b>382</b>             |

Fonte: nostre elaborazioni su dati provenienti dai *Report statistici sulla presenza dei MSNA in Italia*.

<sup>5</sup> In "Altre nazionalità", a seconda dei mesi di rilevazione, sono ricomprese le seguenti: Algeria, Benin, Bosnia-Erzegovina, Brasile, Burkina Faso, Camerun, Ciad, Cina, Colombia, Congo, Repubblica democratica del Congo, Guinea Bissau, Guinea equatoriale, India, Iran, Iraq, Liberia, Libia, Moldova, Niger, Perù, Sierra Leone, Siria, Sudan, Togo, Turchia, Ucraina, altre non definite e, in un mese, 6 apolidi.

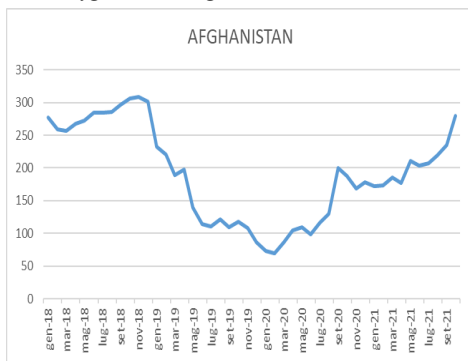
Per facilitare la lettura dei dati, sono state prese in considerazione in maniera dettagliata solo le nazionalità maggiormente presenti, che comunque coprono sempre il 94-95% dell'intera comunità di MSNA in Italia. Per una più semplice comprensione del fenomeno, di tutti i dati mensili, peraltro visualizzabili nelle Figg. da 3.a a 3.s, sono stati solo riportati il valore assoluto e la percentuale, rispetto al totale delle presenze mensili, dei minori nel primo mese di rilevazione (gennaio 2018), nell'ultimo prima del lockdown (febbraio 2020), nel primo in cui è scoppiata la pandemia (marzo 2020) e nell'ultimo di cui i dati erano disponibili al momento della stesura del presente contributo (ottobre 2021). Per il primo periodo di 26 mesi e per il secondo di 20, sono state anche calcolate le presenze medie per nazionalità.

Facendo un'analisi dei dati della Tabella 3, si nota come ad inizio 2018 i MSNA provenivano soprattutto dai seguenti paesi: Gambia (1.712 unità, pari all'11,5%), Albania (1.638, pari all'11,1%), Egitto (1.524, pari al 10,2%), Guinea (1.305, pari all'8,7%) e Nigeria (1.102, pari al 7,4%). Ad inizio 2020, le presenze di MSNA si sono ridotte per quasi tutte le nazionalità, seppure con ritmi diseguali rispetto al totale: tra i paesi considerati; anche quelli che vedono a febbraio 2020 un leggero aumento in valore assoluto rispetto a gennaio 2018 (Kosovo e Pakistan), hanno alle spalle situazioni diverse che saranno analizzate in seguito più nel dettaglio.

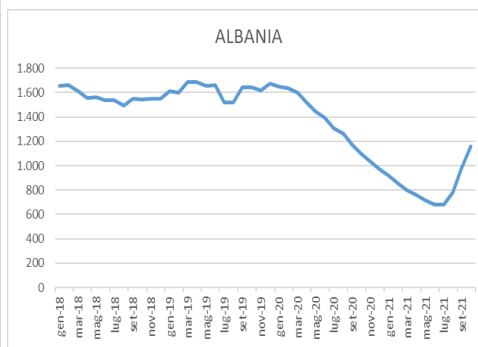
Nei mesi successivi allo scoppio della pandemia, l'incremento delle presenze di MSNA sono soprattutto attribuibili al Bangladesh (che passa da circa 500 unità di inizio 2020 a quasi 2.700 di ottobre 2021), Tunisia (dalle 238 alle oltre 1.600) ed Egitto (che dopo un crollo nel periodo precedente, passa da poco meno di 600 unità di inizio 2020 a quasi 1.400 di ottobre 2021). Le due nazionalità attualmente maggiormente presenti sono quella del Bangladesh (26,1%) e della Tunisia (15,7%) mentre crolla in termini relativi quella albanese la cui presenza relativa di MSNA, pur sempre considerevole in valore assoluto, torna in percentuale a valori di inizio 2018 (dopo essere arrivata a coprire, al momento dello scoppio della pandemia, la quota di quasi un terzo del totale dei MSNA).

Con l'aiuto delle Figg. da 3.a a 3.s, che mostrano l'evoluzione della presenza di MSNA per le 17 nazionalità (oltre al dato delle "altre nazionalità") riportate in Tabella 3, si denota come le situazioni siano piuttosto differenti e che in taluni casi sia evidente un cambiamento intervenuto proprio nel periodo immediatamente successivo allo scoppio della pandemia. Ciò non toglie, come vedremo, che si possano anche ipotizzare altre possibili determinanti di questa evoluzione del fenomeno oggetto di studio.

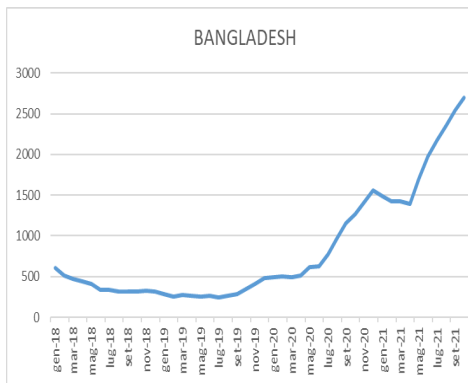
**Figura 3-a.** Presenza in v.a. di MSNA provenienti dall'Afghanistan da gennaio 2018 a ottobre 2021



**Figura 3-b.** Presenza in v.a. di MSNA provenienti dall'Albania da gennaio 2018 a ottobre 2021



**Figura 3-c.** Presenza in v.a. di MSNA provenienti dal Bangladesh da gennaio 2018 a ottobre 2021



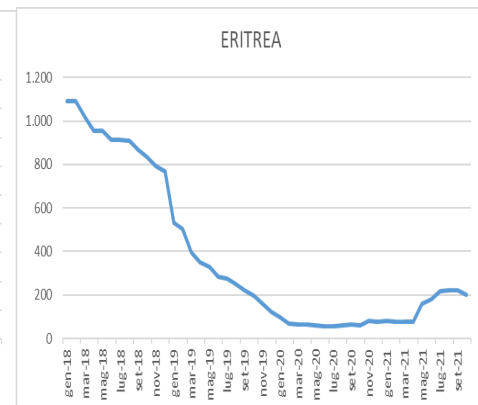
**Figura 3-d.** Presenza in v.a. di MSNA provenienti dalla Costa d'Avorio da gennaio 2018 a ottobre 2021

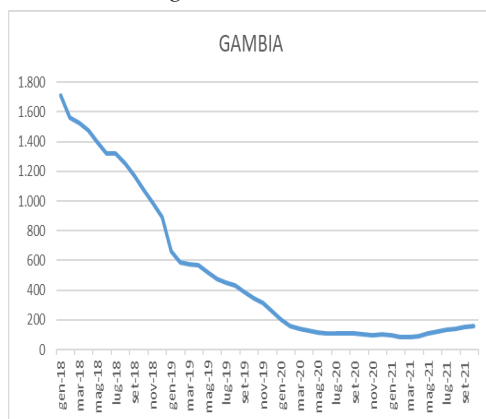
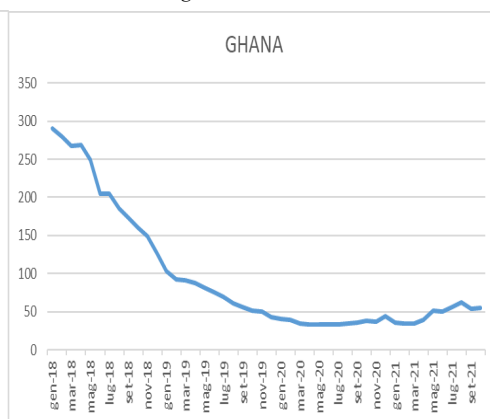
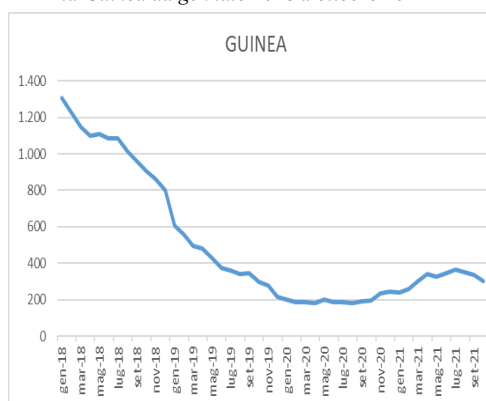
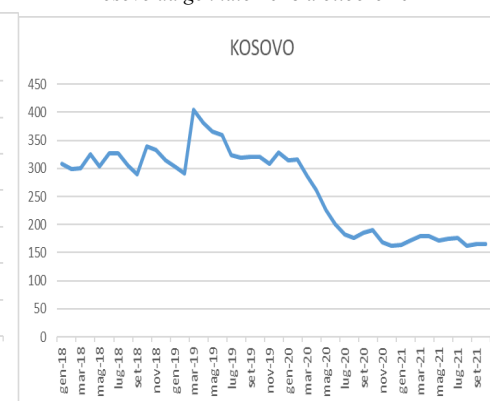
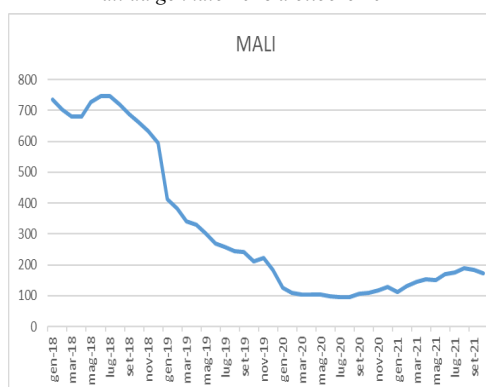
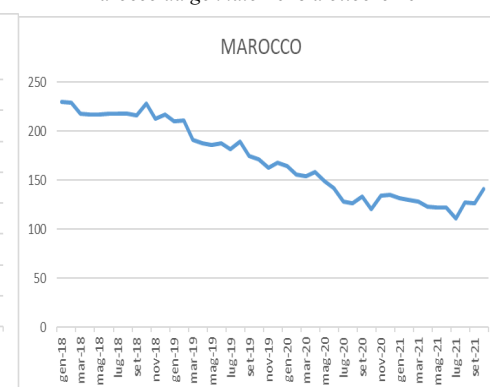


**Figura 3-e.** Presenza in v.a. di MSNA provenienti dall'Egitto da gennaio 2018 a ottobre 2021

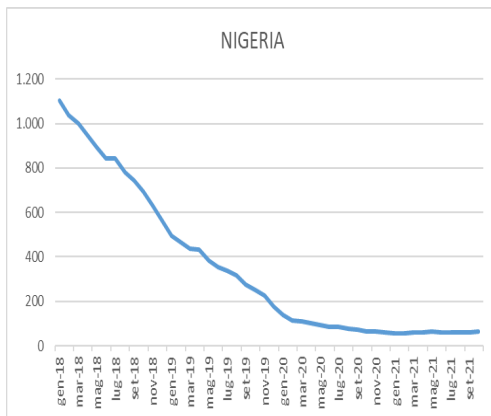


**Figura 3-f.** Presenza in v.a. di MSNA provenienti dall'Eritrea da gennaio 2018 a ottobre 2021

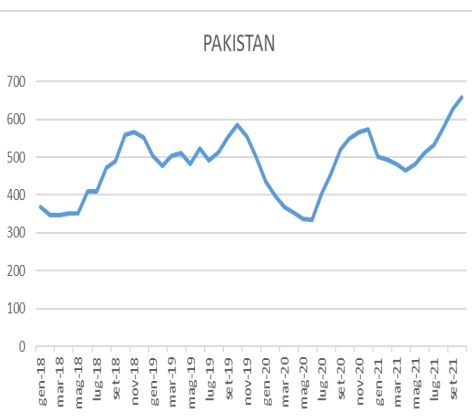


**Figura 3-g.** Presenza in v.a. di MSNA provenienti dal Gambia da gennaio 2018 a ottobre 2021**Figura 3-h.** Presenza in v.a. di MSNA provenienti dal Ghana da gennaio 2018 a ottobre 2021**Figura 3-i.** Presenza in v.a. di MSNA provenienti dalla Guinea da gennaio 2018 a ottobre 2021**Figura 3-l.** Presenza in v.a. di MSNA provenienti dal Kosovo da gennaio 2018 a ottobre 2021**Figura 3-m.** Presenza in v.a. di MSNA provenienti dal Mali da gennaio 2018 a ottobre 2021**Figura 3-n.** Presenza in v.a. di MSNA provenienti dal Marocco da gennaio 2018 a ottobre 2021

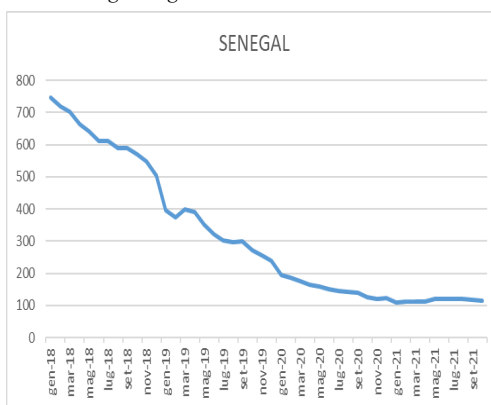
**Figura 3-o.** Presenza in v.a. di MSNA provenienti dalla Nigeria da gennaio 2018 a ottobre 2021



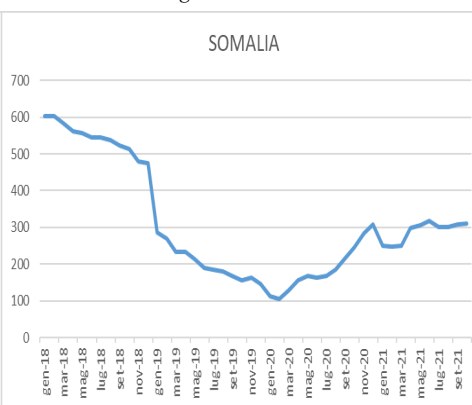
**Figura 3-p.** Presenza in v.a. di MSNA provenienti dal Pakistan da gennaio 2018 a ottobre 2021



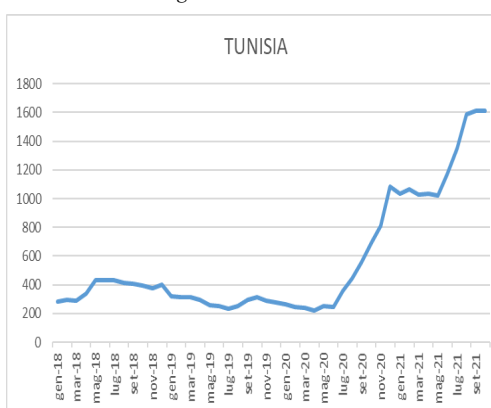
**Figura 3-q.** Presenza in v.a. di MSNA provenienti dal Senegal da gennaio 2018 a ottobre 2021



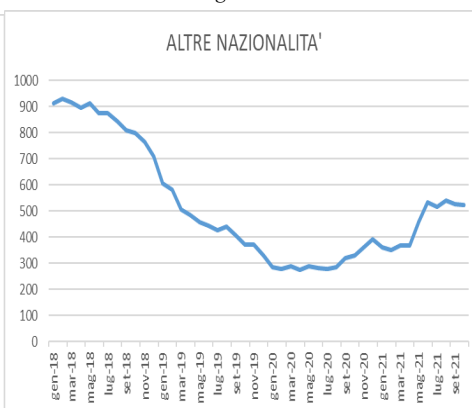
**Figura 3-r.** Presenza in v.a. di MSNA provenienti dalla Somalia da gennaio 2018 a ottobre 2021



**Figura 3-s.** Presenza in v.a. di MSNA provenienti dalla Tunisia da gennaio 2018 a ottobre 2021



**Figura 3-t.** Presenza in v.a. di MSNA provenienti da ALTRE nazionalità da gennaio 2018 a ottobre 2021



I MSNA provenienti dall'Afghanistan (Figura 3-a) diminuiscono in valore assoluto per tutto il 2019 (dai 277 di dicembre 2018 alle sole 70 unità di febbraio 2020), mentre da marzo 2020 si triplicano fino alle attuali 280 unità, toccando in termini percentuali il minimo dell'1,3% a inizio 2020 e raggiungendo attualmente il 2,7%. Tale paese si è spesso contraddistinto nel passato per una forte presenza di MSNA in fuga, a causa di guerre e conflitti con riflessi drammatici sulle condizioni di vita della popolazione: violenza, discriminazione, povertà, assenza di servizi pubblici, e tanto altro ancora hanno spesso indotto le famiglie a far lasciare ai propri piccoli l'Afghanistan, per trasferirsi in Europa, talora dopo un periodo di tempo in Pakistan o in Iran. L'incremento più elevato di presenze di MSNA afgani degli ultimi mesi, è verosimilmente attribuibile al disimpegno delle forze armate statunitensi dal territorio afgano, con tentativi di fuga da un paese tornato sotto il regime talebano, e dalla concomitante drammatica crisi economica aggravata dalla pandemia che ha visto aumentare le persone sull'orlo della carestia dai circa 2,5 milioni di settembre 2019 ai 3,5 milioni di maggio 2020, con un 70% delle famiglie che denunciava, secondo l'UNHCR, una riduzione sia del proprio reddito sia del flusso delle rimesse dall'estero<sup>6</sup>.

Per i MSNA provenienti dal Bangladesh (Figura 3-c), si nota una presenza in diminuzione assoluta dalle 605 unità di inizio 2018 alle 244 di luglio 2019; tale presenza aumenta leggermente fino ad aprile 2020 (quando raggiunge la quota di circa 500 minori) per poi crescere in maniera esponenziale da maggio 2020 ad oggi quando i MSNA provenienti da tale paese sono quasi 2.700. L'incidenza percentuale sul totale dei minori presenti in Italia, che ha valori sempre tra il 4 e il 5% fino al 2019, cresce fino all'attuale 26%, portando il Bangladesh ad essere per distacco la prima nazionalità di provenienza. A differenza di tante altre nazioni, le motivazioni dei provenienti dal Bangladesh sono spesso state soprattutto di tipo economico e solo in misura inferiore di tipo politico<sup>7</sup>. Il Bangladesh, infatti, ha circa il 15% di popolazione in condizioni di povertà assoluta, disponendo di meno di 1,9 dollari al giorno (secondo i criteri definiti dalla Banca Mondiale). Le famiglie del Bangladesh hanno spesso "investito" sui propri figli facendoli emigrare, sperando di poter contare sulle future rimesse prodotte dal loro lavoro. Questa situazione si è ovviamente aggravata nel periodo pandemico portando verosimilmente alla crescita esponenziale di MSNA in fuga dal proprio paese di origine.

<sup>6</sup> Cfr, Osservatorio Nazionale sui Minori Stranieri Non Accompagnati, Afghanistan: un infinito esodo di minori non accompagnati, CESPI, Approfondimento n.9/settembre 2021,

<sup>7</sup> Cfr., ad esempio, <https://www.lenius.it/perche-emigrano-dal-bangladesh/>

Situazione per certi versi simile ai provenienti dal Bangladesh è quella dei MSNA provenienti dalla Tunisia (Figura 3-s) che per il periodo pre-pandemico hanno una presenza assoluta dalle 250 alle 400 unità ma che dall'estate 2020 cresce fino a raggiungere le oltre 1.600 presenze. I MSNA tunisini, attualmente hanno raggiunto una quota del quasi 16% del totale minori, quando nel periodo pre-covid tale percentuale era sempre rimasta contenuta tra il 2 e il 5%. Anche per i tunisini, le motivazioni sono sempre state soprattutto di tipo economico. Come si può evincere dal Rapporto annuale sull'emigrazione dell'organizzazione non governativa Forum tunisino per i diritti economici e sociali, anche per i MSNA tunisini sono le famiglie a far emigrare i figli minorenni in modo tale che, una volta in Europa, possano guadagnare e spedire quindi del denaro nel paese di origine. Quasi la metà dei tunisini emigrati ritiene impossibile avere successo in Tunisia, e sicuramente questa considerazione si è accresciuta con la grave crisi economica dettata da quella pandemica.

Anche per i MSNA provenienti dal Pakistan (Figura 3-p), pur in maniera inferiore rispetto a Bangladesh e Tunisia, si nota un raddoppio della presenza nel secondo periodo, motivata da allentamenti delle misure anti covid19 e da motivazioni economiche simili a quelle sopraindicate: dai meno di 400 di gennaio 2018 (pari al 2,5%), si è adesso a 659 minori che, in proporzione rispetto al totale dei MSNA sul suolo italiano, coprono il 6,4% divenendo la quarta forza presente sul territorio rispetto alla nazionalità.

Pure i MSNA emigrati dall'Egitto, soprattutto negli ultimi sei mesi del 2021, hanno visto un aumento della propria presenza, facendo all'incirca tornare il paese ai livelli del 2018, dopo una diminuzione abbastanza netta nel periodo pre-covid e una sostanziale tenuta nel primo anno di pandemia. Dagli iniziali 1.524 MSNA egiziani (pari al 10,2% del totale MSNA in Italia), si è avuta una riduzione di circa i due terzi attestandosi sulle 500-600 unità per tutto il periodo intermedio che va dall'estate 2019 a quella 2020. Attualmente si riscontra una presenza di 1.364 MSNA egiziani, con un raddoppio negli ultimi mesi (dai 672 di maggio 2021) che ne torna a fare la terza nazionalità per importanza tra i MSNA in Italia col 13,2%.

Una ripresa importante della presenza di MSNA negli ultimi mesi, in particolare negli ultimi quattro, è riscontrabile per i provenienti dal territorio albanese. In termini assoluti, nel periodo pre-covid preso in considerazione, i minori albanesi sono sempre stati tra le 1.500 e le 1.600 unità, manifestando una decrescita inferiore a quella dei provenienti da altre nazionalità. In termini percentuali, ciò aveva portato l'incidenza degli albanesi a triplicarsi da gennaio 2018 (11% circa) fino al momento di scoppio della pandemia (oltre il 30%). Nel periodo pandemico la presenza di

MSNA albanesi scende drasticamente fino alle 680 unità a giugno 2021, quando raggiunge circa il 9% rispetto al totale dei MSNA sul suolo italiano. Negli ultimi 4 mesi, la presenza assoluta si è quasi raddoppiata, passando a 1.157 unità di ottobre 2021 con una quota percentuale dell'11,1%, simile a quella di gennaio 2018. I MSNA albanesi che giungono in Italia, sono spesso vicini alla maggiore età (16-17 anni) e solitamente accompagnati dagli stessi genitori per l'impossibilità dei minori stessi di viaggiare da soli in aereo. I genitori poi rientrano in patria ma consentono in tal modo ai propri figli, già ormai autonomi, di essere accolti in comunità italiane per ottenere il permesso di soggiorno. Il dato abbastanza elevato di MSNA provenienti dall'Albania contrasta con le richieste di protezione dei maggiorenni della stessa nazionalità, che restano sempre su cifre minime (intorno al 4%) e conferma come i minori si muovano dall'Albania, in particolare da zone rurali<sup>8</sup>, a causa dell'impossibilità di ottenere un permesso di soggiorno per motivi di lavoro in Italia. La situazione pandemica, con i conseguenti timori di contagio e le stringenti misure di chiusura, hanno quindi determinato un blocco di arrivi dall'Albania, recentemente ripresi con la normalizzazione della situazione<sup>9</sup>.

La presenza di MSNA di altre nazionalità, ha invece seguito andamenti diversi, con continui decrementi durante l'arco dei 46 mesi presi in considerazione e, solo talvolta, con leggere risalite negli ultimi mesi estivi del 2021. Basti pensare ai MSNA provenienti dal Gambia, oltre 1.700 a gennaio 2018, in quel momento prima nazionalità in assoluto sul territorio italiano con l'11,5%, adesso diminuiti a meno di un decimo (157 ad ottobre 2021, pari all'1,9%). Oppure alla Guinea, che a gennaio 2018 aveva una incidenza dell'8,7% dei MSNA (1.305 minori) e che dopo aver raggiunto il minimo di 179 ad aprile 2020, è leggermente risalita alle circa 300 unità di ottobre 2021. Situazione simile a quanto successo per la Costa d'Avorio, presente con 1.189 MSNA a gennaio 2018 (8%), ed oggi con 343 unità (3,3%) ma comunque in aumento di circa un centinaio di minori nel periodo successivo all'insorgenza della pandemia. Tutto ciò è anche frutto della cosiddetta "chiusura della rotta libica", che ha portato ad una forte contrazione degli arrivi di adulti, e ad un conseguente forte calo di MSNA di provenienza Africa Sub Sahariana: oltre che di Gambia e Guinea, ne è testimonianza anche la Nigeria che dagli oltre 1.100 MSNA di gennaio 2018 (7,4% del totale minori), vede solo 64 presenze ad ottobre

---

<sup>8</sup> Cfr., ad esempio, <https://secondowelfare.it/immigrazione-e-accoglienza/i-flussi-di-minori-stranieri-non-accompagnati-nell'anno-della-pandemia/>

<sup>9</sup> Cfr. <https://www.secondowelfare.it/immigrazione-e-accoglienza/i-flussi-di-minori-stranieri-non-accompagnati-nell'anno-della-pandemia/>

2021 (0,6%) e anche il Senegal che dai 746 minori di inizio 2018, ne ha attualmente solo 114 (Figura 3-q).

Una motivazione profondamente importante, e molto diversa dalle precedenti, che da tempo porta alla fuga dal proprio paese di origine, è quella che spesso spinge i MSNA a fuggire dall'Eritrea. Com'è noto, i giovani eritrei sono obbligati a prestare servizio militare dai 17 anni a tempo illimitato, condizione aspramente criticata negli anni dalle Nazioni Unite. Tale obbligatorietà fu introdotta ufficialmente dal presidente Afwerki nel 1995, per poter affrontare i frequenti conflitti di confine tra l'Eritrea e i suoi vicini. Già dal 2018, però, il governo ha mostrato aperture, peraltro non ancora esaudite, per venire incontro alle richieste delle Nazioni Unite. Dai dati della Tabella 3 e dalla Figura 3-f si evince che la presenza di MSNA eritrei è costantemente in diminuzione, ad eccezione dell'estate 2021, passando dai 1.092 di inizio 2018 ai 199 di ottobre 2021. Tale presenza, quindi, non sembra essere stata influenzata dal periodo di pandemia, verosimilmente a causa della scarsa correlazione del principale motivo dell'emigrazione con essa.

## **5. MSNA per regione di accoglienza**

L'analisi della presenza dei MSNA per regione di accoglienza assume una particolare importanza anche per i risvolti politico-sociali. Come si può desumere dall'analisi della Tabella 4, invero, in Italia ci sono profonde differenze a livello regionale nella presenza di minori, passando, secondo gli ultimi dati disponibili, dalle 46 presenze dell'Umbria alle 3.029 della Sicilia.

La regione che, a prescindere dal mese, è sempre di gran lunga quella con la maggior presenza di MSNA sul proprio territorio, è la Sicilia. Ad inizio 2018 vi erano 6.123 minori (pari al 41% del totale) ma, come si può vedere sia dalla Tabella 4 che dalla Figura 4, il peso della accoglienza dei minori da parte della regione Sicilia era andato successivamente contraendosi fino al periodo di scoppio della pandemia. In termini assoluti, si notano meno di 900 unità nei mesi di febbraio-marzo 2020, con una percentuale di circa il 16% rispetto al totale dei MSNA presenti sull'intero territorio nazionale. L'aumento delle presenze parte dall'estate 2020 e attualmente si riscontrano oltre 3.000 unità nella regione, con una incidenza del quasi 30% che, sicuramente molto più contenuta rispetto ad inizio 2018, fa della Sicilia la regione che deve ancora oggi affrontare il maggior peso del fenomeno della presenza dei MSNA in Italia. In tale regione, grazie anche alla sua posizione al centro del Mediterraneo, si concentrano soprattutto i minori provenienti da Ban-

gladesh, Costa d'Avorio, Tunisia, Guinea, Somalia, Gambia per cui l'aumento delle presenze di minori nell'isola siciliana concorda sicuramente con l'incremento già visto della presenza di alcuni di questi paesi, in particolare del Bangladesh.

**Tabella 4.** *Presenza in v.a. di MSNA rispetto a taluni periodi pre e post covid, da gennaio 2018 a ottobre 2021, per le regioni di accoglienza nei due periodi*

| Territori        | Inizio e fine periodo pre covid preso in considerazione |                           | Presenza media mensile | Inizio e fine periodo covid preso in considerazione |                          | Presenza media mensile |
|------------------|---------------------------------------------------------|---------------------------|------------------------|-----------------------------------------------------|--------------------------|------------------------|
|                  | Gennaio 2018<br>v.a. e %                                | Febbraio 2020<br>v.a. e % |                        | Marzo 2020<br>v.a. e %                              | Ottobre 2021<br>v.a. e % |                        |
| Abruzzo          | 107 (0,7%)                                              | 100 (1,9%)                | <b>95</b>              | 107 (2,0%)                                          | 107 (1,0%)               | <b>104</b>             |
| Basilicata       | 200 (1,3%)                                              | 106 (2,0%)                | <b>120</b>             | 95 (1,8%)                                           | 162 (1,6%)               | <b>116</b>             |
| Calabria         | 1.030 (6,9%)                                            | 105 (2,0%)                | <b>480</b>             | 100 (1,9%)                                          | 732 (7,1%)               | <b>338</b>             |
| Campania         | 561 (3,8%)                                              | 146 (2,7%)                | <b>298</b>             | 159 (3,0%)                                          | 317 (3,1%)               | <b>217</b>             |
| Emilia Romagna   | 929 (6,2%)                                              | 568 (10,6%)               | <b>729</b>             | 547 (10,4%)                                         | 825 (8,0%)               | <b>547</b>             |
| Friuli V. Giulia | 562 (3,8%)                                              | 586 (10,9%)               | <b>633</b>             | 565 (10,7%)                                         | 1.026 (9,9%)             | <b>744</b>             |
| Lazio            | 1.012 (6,8%)                                            | 366 (6,8%)                | <b>679</b>             | 386 (7,3%)                                          | 352 (3,4%)               | <b>332</b>             |
| Liguria          | 230 (1,5%)                                              | 205 (3,8%)                | <b>209</b>             | 194 (3,7%)                                          | 313 (3,0%)               | <b>205</b>             |
| Lombardia        | 1.088 (7,3%)                                            | 812 (15,1%)               | <b>891</b>             | 750 (14,3%)                                         | 1.064 (10,3%)            | <b>737</b>             |
| Marche           | 181 (1,2%)                                              | 125 (2,3%)                | <b>156</b>             | 110 (2,1%)                                          | 196 (1,9%)               | <b>117</b>             |
| Molise           | 102 (0,7%)                                              | 59 (1,1%)                 | <b>76</b>              | 52 (1,0%)                                           | 82 (0,8%)                | <b>80</b>              |
| Piemonte         | 413 (2,8%)                                              | 246 (4,6%)                | <b>333</b>             | 260 (4,9%)                                          | 313 (3,0%)               | <b>241</b>             |
| Pr. aut. Bolzano | 88 (0,6%)                                               | 78 (1,4%)                 | <b>80</b>              | 73 (1,4%)                                           | 73 (0,7%)                | <b>60</b>              |
| Pr. aut. Trento  | 43 (0,3%)                                               | 19 (0,4%)                 | <b>32</b>              | 22 (0,4%)                                           | 24 (0,2%)                | <b>20</b>              |
| Puglia           | 705 (4,7%)                                              | 195 (3,6%)                | <b>418</b>             | 253 (4,8%)                                          | 849 (8,2%)               | <b>483</b>             |
| Sardegna         | 556 (3,7%)                                              | 25 (0,5%)                 | <b>184</b>             | 29 (0,6%)                                           | 57 (0,6%)                | <b>50</b>              |
| Sicilia          | 6.123 (41,0%)                                           | 867 (16,2%)               | <b>3.478</b>           | 833 (15,8%)                                         | 3.029 (29,4%)            | <b>1.825</b>           |
| Toscana          | 633 (4,2%)                                              | 435 (8,1%)                | <b>493</b>             | 404 (7,7%)                                          | 447 (4,3%)               | <b>317</b>             |
| Umbria           | 26 (0,2%)                                               | 36 (0,7%)                 | <b>49</b>              | 36 (0,7%)                                           | 46 (0,4%)                | <b>34</b>              |
| Val d'aosta      | 3 (0,02%)                                               | 12 (0,2%)                 | <b>9</b>               | 13 (0,3%)                                           | 4 (0,04%)                | <b>8</b>               |
| Veneto           | 347 (2,3%)                                              | 277 (5,2%)                | <b>312</b>             | 277 (5,3%)                                          | 299 (2,9%)               | <b>234</b>             |

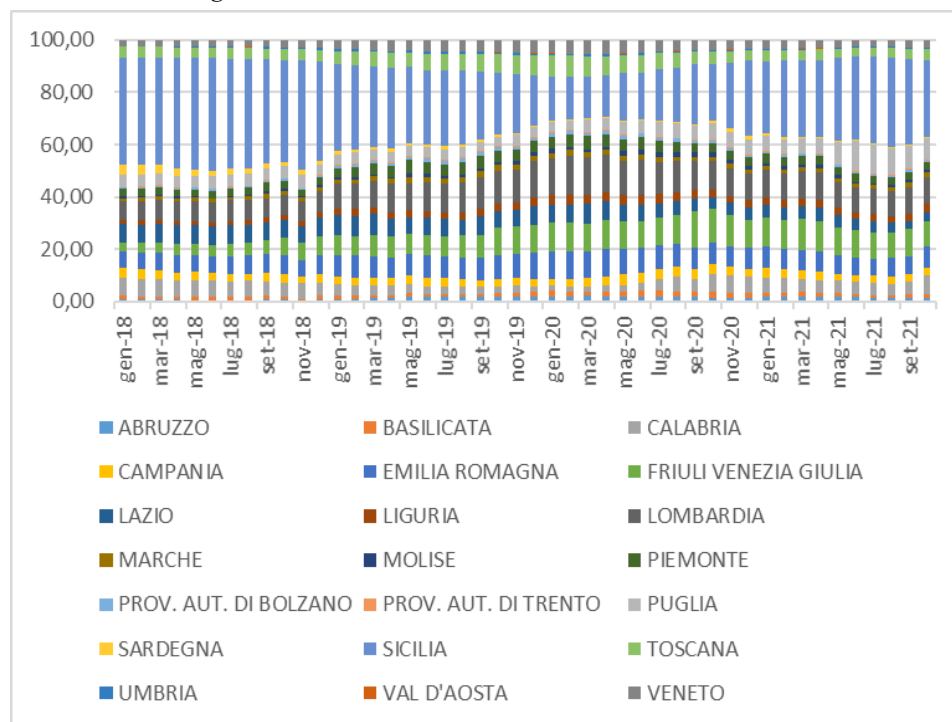
Fonte: nostre elaborazioni su dati provenienti dai *Report statistici sulla presenza dei MSNA in Italia*.

La Lombardia, seconda regione per accoglienza ad inizio 2018 con oltre 1.000 unità sul proprio territorio (pari al 7,3% delle presenze totali), ha visto aumentare il proprio peso percentuale nel periodo precedente allo scoppio della pandemia, arrivando al 14-15% ad inizio 2020, per assestarsi intorno all'attuale circa 10% tornando ad avere oltre 1.000 presenze ad ottobre 2021.

Anche Emilia Romagna e Friuli, pur in un contesto di diminuzione assoluta della presenza dei minori oggetto di studio, vedono invece aumentare la propria incidenza percentuale fino ad inizio 2020 per poi stabilizzarsi su elevate quote (rispettivamente 8 e 10%); in tal caso, tale trend può essere in parte riconducibile al dialogo aperto in maniera autonoma da tali regioni con alcuni degli Stati di provenienza dei minori. Per il Friuli, l'aumento riscontrato è sicuramente anche frutto del fatto che tale regione sia l'approdo naturale dei provenienti dalle zone

balcaniche, che sempre più si riversano in Italia dopo una netta chiusura delle rotte del Mediterraneo centrale.

**Figura 4.** Distribuzione percentuale dei MSNA da gennaio 2018 a ottobre 2021, per regione di accoglienza



Andamento molto diverso quello della presenza dei MSNA nella regione della Calabria: si nota una netta diminuzione nei primi 26 mesi in analisi, che dalle oltre 1.000 unità di inizio 2018 vede scendere la presenza degli MSNA al minimo di 100 (meno del 2%) nel marzo 2020, proprio nel momento di avvio del lockdown in Italia. Successivamente, si nota una costante crescita sia assoluta che percentuale che porta ad oltre 700 unità (7,1% rispetto al totale) i minori in Calabria.

Molto elevata la crescita di MSNA in Puglia negli ultimi mesi, sia in valore assoluto che percentuale. Le presenze si sono praticamente quadruplicate dalle 195 di febbraio 2020 alle 849 di ottobre 2021, superando di fatto ampiamente anche la quota di gennaio 2018 (circa 700 unità). In termini percentuali, la presenza di MSNA copre attualmente circa l'8% del totale di presenze sul territorio italiano.

Anche la Sardegna ha un comportamento difforme dalle altre regioni: la presenza di MSNA crolla già nel 2019 e dai circa 500 di inizio 2018, diventa una

delle regioni con la minore presenza di minori, con valori assoluti sempre inferiori alle 100 unità anche per tutti i mesi successivi in analisi.

A livello regionale, com'è noto, e come già accennato in riferimento ad Emilia Romagna e Friuli, si riscontra una certa frammentazione sia di normative che di pratiche in tema di accoglienza di MSNA che può essere anche parzialmente causa dei diversi trend innanzi evidenziati e sicuramente sarebbe auspicabile un maggior coordinamento anche nazionale per affrontare la presa in carico di tali minori realizzando un'armonizzazione delle politiche locali e nazionali. Andrebbe quindi incentivata una più stretta connessione fra organi istituzionali e non, aumentando protocolli integrati fra Prefetture, Comune, Questure, Tribunali, Servizi Sociali, Aziende Sanitarie e Uffici Scolastici, favorendo una più efficace ed efficiente accoglienza nei confronti dei MSNA. Si può portare ad esempio, quale *best practice*, la creazione della Banca Dati sui MSNA da parte del Garante Regionale della regione Puglia, in stretta connessione con gli organi giudiziari al fine di rendere più semplice il monitoraggio ed il lavoro per la presa in carico dei minori.

Non va trascurato, inoltre, che dietro la maggiore accoglienza da parte di alcune regioni c'è una maggiore presenza di alcune organizzazioni internazionali quali UNICEF e UNHCR che ha aiutato a promuovere i diritti dei MSNA, la loro accoglienza e l'integrazione sul territorio.

## **6. Alcuni dati sulla presenza di MSNA di genere femminile**

Viene infine di seguito fornito un breve cenno su alcune caratteristiche delle sole MSNA di genere femminile.

Come già visto in precedenza, la presenza femminile in questa particolare categoria di minori emigrati sul suolo italiano è sempre stata contenuta: dalle 1.101 unità di inizio 2018 si è passati alle sole 261 di ottobre 2021, con una diminuzione abbastanza costante che infatti vede adesso i maschi non accompagnati essere circa 40 volte più presenti della compagine femminile (vds Tabella 1 e Figura 1). In termini percentuali, dal circa 7% di inizio periodo, attualmente si è scesi addirittura sotto il 3% (2,53 ad ottobre 2021).

La Tabella 5, se confrontata anche con la Tabella 2, fa emergere che tra le MSNA, rispetto ai minori di genere maschile, vi è sempre una maggior percentuale di presenza di età più basse. Le MSNA con età inferiore a 7 anni sono quasi sempre tra il 4 e il 7% (rispetto a percentuali sul totale sempre inferiori all'1%). Questa tendenza è verificata anche nella classe 7-14 anni e, ovviamente, si verifica il con-

trario nelle età più elevate (15, 16 e 17 anni). Una spiegazione a questa più bassa età per le MSNA, da taluni fornita, riguarda purtroppo un collegamento alla tratta di esseri umani, anche se ovviamente non è semplice desumere tutte le possibili cause. Verosimilmente, invece, l'età più elevata dei MSNA di genere maschile può essere legata ad altre motivazioni quali il cercare lavoro o il fuggire dall'obbligo "a vita" del servizio militare (come per gli eritrei), tutte motivazioni che ovviamente sorgono soprattutto con l'avvicinarsi della maggiore età.

**Tabella 5.** *Distribuzione percentuale delle MSNA dal gennaio 2018 a ottobre 2021*

| Periodo<br>pre covid | 17   | 16   | Età<br>15 | 7-14 | <7  | MF  | Periodo<br>covid | 17   | 16   | Età<br>15 | 7-14 | <7  | MF  |
|----------------------|------|------|-----------|------|-----|-----|------------------|------|------|-----------|------|-----|-----|
| <b>2018</b>          |      |      |           |      |     |     |                  |      |      |           |      |     |     |
| Gennaio              | 52,1 | 21,0 | 10,7      | 13,3 | 2,9 | 100 |                  |      |      |           |      |     |     |
| Febbraio             | 55,0 | 26,7 | 6,6       | 6,6  | 5,0 | 100 | <b>2020</b>      |      |      |           |      |     |     |
| Marzo                | 51,3 | 20,9 | 10,4      | 14,5 | 2,9 | 100 | Marzo            | 46,4 | 22,2 | 9,2       | 19,9 | 2,3 | 100 |
| Aprile               | 50,6 | 20,3 | 10,8      | 15,3 | 3,0 | 100 | Aprile           | 45,2 | 21,6 | 9,5       | 19,5 | 4,1 | 100 |
| Maggio               | 48,1 | 22,2 | 10,6      | 15,8 | 3,3 | 100 | Maggio           | 43,4 | 20,1 | 12,9      | 19,3 | 4,4 | 100 |
| Giugno               | 48,3 | 22,7 | 10,9      | 15,1 | 3,1 | 100 | Giugno           | 42,4 | 21,0 | 11,3      | 21,0 | 4,2 | 100 |
| Luglio               | 47,4 | 23,2 | 10,6      | 15,7 | 3,1 | 100 | Luglio           | 39,0 | 19,5 | 12,7      | 20,3 | 8,5 | 100 |
| Agosto               | 47,1 | 23,4 | 10,8      | 15,8 | 2,9 | 100 | Agosto           | 40,9 | 19,0 | 12,9      | 22,0 | 5,2 | 100 |
| Settembre            | 45,4 | 24,5 | 10,9      | 16,3 | 2,9 | 100 | Settembre        | 43,8 | 18,5 | 12,4      | 19,7 | 5,6 | 100 |
| Ottobre              | 46,7 | 23,9 | 10,2      | 16,3 | 2,9 | 100 | Ottobre          | 43,9 | 19,5 | 11,3      | 19,5 | 5,9 | 100 |
| Novembre             | 46,0 | 24,5 | 9,8       | 16,8 | 2,9 | 100 | Novembre         | 41,4 | 20,5 | 13,0      | 18,8 | 6,3 | 100 |
| Dicembre             | 44,6 | 25,4 | 9,4       | 17,4 | 3,2 | 100 | Dicembre         | 43,3 | 21,4 | 9,1       | 19,4 | 6,7 | 100 |
| <b>2019</b>          |      |      |           |      |     |     | <b>2021</b>      |      |      |           |      |     |     |
| Gennaio              | 43,7 | 23,7 | 11,0      | 17,9 | 3,7 | 100 | Gennaio          | 44,1 | 21,1 | 9,3       | 18,5 | 7,0 | 100 |
| Febbraio             | 44,7 | 22,7 | 10,5      | 19,0 | 3,1 | 100 | Febbraio         | 41,3 | 22,0 | 10,1      | 19,7 | 6,9 | 100 |
| Marzo                | 46,7 | 23,6 | 10,0      | 16,2 | 3,5 | 100 | Marzo            | 40,2 | 22,2 | 10,5      | 21,3 | 5,9 | 100 |
| Aprile               | 44,3 | 23,0 | 10,9      | 16,6 | 5,1 | 100 | Aprile           | 41,5 | 19,5 | 12,0      | 21,6 | 5,4 | 100 |
| Maggio               | 44,1 | 22,0 | 10,4      | 17,7 | 5,7 | 100 | Maggio           | 39,4 | 21,1 | 11,8      | 22,4 | 5,3 | 100 |
| Giugno               | 44,2 | 22,6 | 10,2      | 17,1 | 5,9 | 100 | Giugno           | 38,3 | 21,5 | 12,9      | 22,3 | 5,1 | 100 |
| Luglio               | 45,1 | 22,0 | 8,9       | 17,8 | 6,1 | 100 | Luglio           | 39,5 | 20,7 | 13,0      | 22,2 | 4,6 | 100 |
| Agosto               | 47,3 | 21,7 | 7,2       | 17,1 | 6,6 | 100 | Agosto           | 37,3 | 23,6 | 12,2      | 22,9 | 4,1 | 100 |
| Settembre            | 48,8 | 20,9 | 6,6       | 19,2 | 4,5 | 100 | Settembre        | 37,3 | 24,6 | 11,6      | 23,2 | 3,3 | 100 |
| Ottobre              | 50,3 | 18,6 | 7,8       | 18,8 | 4,5 | 100 | Ottobre          | 37,5 | 23,0 | 10,3      | 26,4 | 2,7 | 100 |
| Novembre             | 50,1 | 17,4 | 9,1       | 18,8 | 4,6 | 100 |                  |      |      |           |      |     |     |
| Dicembre             | 47,3 | 19,6 | 9,8       | 18,3 | 5,0 | 100 |                  |      |      |           |      |     |     |
| <b>2020</b>          |      |      |           |      |     |     |                  |      |      |           |      |     |     |
| Gennaio              | 46,0 | 19,9 | 9,8       | 19,2 | 5,2 | 100 |                  |      |      |           |      |     |     |
| Febbraio             | 45,8 | 19,2 | 9,6       | 20,4 | 5,0 | 100 |                  |      |      |           |      |     |     |

Fonte: nostre elaborazioni su dati provenienti dai *Report statistici sulla presenza dei MSNA in Italia*.

Per ciò che concerne la nazionalità di provenienza delle MSNA, la tabella 6 pone in luce ad inizio 2018 una elevatissima concentrazione per le ragazze di origine eritrea e nigeriana. Queste ultime coprono quasi il 40% del totale delle minori con 439 unità mentre le provenienti dall'Eritrea erano 208 con un'incidenza percentuale del 18,9%. Nel corso del periodo preso in esame, la presenza di queste due nazionalità nella compagine femminile delle MSNA si è praticamente azzerata, con

una decrescita continua e costante che vede attualmente solo 19 nigeriane e 12 eritree. L'unica nazionalità di provenienza che vede crescere in maniera importante il proprio peso percentuale tra le MSNA è la Costa d'Avorio, dal 7,5 al 20,3%, pur in presenza di una diminuzione del peso assoluto da 83 a 53. Non disponendo della possibilità di incrociare i dati delle caratteristiche sociodemografiche, è ovvio che non si può sempre desumere i motivi di certi andamenti. In questo caso, per esempio, sarebbe stato anche utile effettuare un'analisi bivariata di età e nazionalità anche per ottenere un riscontro sulle ipotesi innanzi fatte.

**Tabella 6.** *Presenza in v.a. delle MSNA rispetto a taluni periodi pre e post covid, da gennaio 2018 a ottobre 2021, per le nazionalità maggiormente presenti nei due periodi*

| Nazionalità di provenienza | Inizio e fine periodo pre covid preso in considerazione |                           | Presenza media mensile | Inizio e fine periodo covid preso in considerazione |                          | Presenza media mensile |
|----------------------------|---------------------------------------------------------|---------------------------|------------------------|-----------------------------------------------------|--------------------------|------------------------|
|                            | Gennaio 2018<br>v.a. e %                                | Febbraio 2020<br>v.a. e % |                        | Marzo 2020<br>v.a. e %                              | Ottobre 2021<br>v.a. e % |                        |
| Albania                    | 90 (8,2%)                                               | 41 (18,5%)                | <b>74</b>              | 43 (18,0%)                                          | 28 (10,7%)               | <b>39</b>              |
| Costa d'Avorio             | 83 (7,5%)                                               | 33 (14,9%)                | <b>65</b>              | 42 (17,6%)                                          | 53 (20,3%)               | <b>48</b>              |
| Eritrea                    | 208 (18,9%)                                             | 7 (3,2%)                  | <b>109</b>             | 7 (2,9%)                                            | 12 (4,6%)                | <b>7</b>               |
| Guinea                     | 23 (2,1%)                                               | 9 (4,1%)                  | <b>12</b>              | 14 (5,9%)                                           | 17 (6,5%)                | <b>16</b>              |
| Marocco                    | 27 (2,5%)                                               | 10 (4,5%)                 | <b>23</b>              | 10 (4,2%)                                           | 12 (4,6%)                | <b>11</b>              |
| Nigeria                    | 439 (39,9%)                                             | 22 (10,0%)                | <b>232</b>             | 24 (10,0%)                                          | 19 (7,3%)                | <b>29</b>              |
| Somalia                    | 62 (5,6%)                                               | 21 (9,5%)                 | <b>40</b>              | 23 (9,6%)                                           | 22 (8,4%)                | <b>18</b>              |
| ...                        |                                                         |                           |                        |                                                     |                          |                        |
| Altre                      | 169 (15,3%)                                             | 85 (31,5%)                | <b>690</b>             | 85 (32,6%)                                          | 98 (37,6%)               | <b>246</b>             |

Fonte: nostre elaborazioni su dati provenienti da *Report statistici sulla presenza dei MSNA in Italia*.

Infine alcune brevi osservazioni sulla presenza delle MSNA per regione di accoglienza (Tabella 7). Circa la metà delle regioni hanno praticamente una presenza nulla o quasi nulla (massimo 1-2 unità) di donne MSNA, con una forte diminuzione in particolare in Campania dove da 24 si è passati a sole 2 MSNA considerando gli estremi del periodo considerato. Anche il Friuli Venezia Giulia non segue l'andamento generale: considerando anche i maschi minori non accompagnati (Tabella 4) si era riscontrata una percentuale di quasi il 10%, che scende invece allo 0,4% per le donne (una sola presenza assoluta). La quota maggiore, anche più di quanto succede per la compagine maschile, è presente in Sicilia: ad inizio 2018 le MSNA sono 629 (57,1%) mentre a ottobre 2021 sono solo 100, pari comunque a quasi il 40%). Tale diminuzione è stata abbastanza costante e aveva avuto nel corso del 2020-21 una stabilizzazione intorno alle 50-60 unità e solo negli ultimi 4 mesi, dall'estate 2021, si nota una crescita più importante con un raddoppio sia in termini assoluti che percentuali.

**Tabella 7.** *Presenza in v.a. delle MSNA rispetto a taluni periodi pre e post covid, da gennaio 2018 a ottobre 2021, per le regioni di accoglienza nei due periodi*

| Territori        | Inizio e fine periodo pre covid preso in considerazione |                           | Presenza media mensile | Inizio e fine periodo covid preso in considerazione |                          | Presenza media mensile |
|------------------|---------------------------------------------------------|---------------------------|------------------------|-----------------------------------------------------|--------------------------|------------------------|
|                  | Gennaio 2018<br>v.a. e %                                | Febbraio 2020<br>v.a. e % |                        | Marzo 2020<br>v.a. e %                              | Ottobre 2021<br>v.a. e % |                        |
| Abruzzo          | 8 (0,4%)                                                | 2 (0,8%)                  | 3                      | 2 (1,1%)                                            | 2 (0,8%)                 | 2                      |
| Basilicata       | - (-)                                                   | 3 (1,1%)                  | 3                      | 1 (0,4%)                                            | - (-)                    | 0                      |
| Calabria         | 52 (4,7%)                                               | 2 (0,8%)                  | 22                     | 2 (0,8%)                                            | 8 (3,1%)                 | 5                      |
| Campania         | 24 (2,2%)                                               | 9 (3,5%)                  | 16                     | 8 (3,1%)                                            | 2 (0,8%)                 | 3                      |
| Emilia Romagna   | 60 (5,5%)                                               | 38 (14,6%)                | 49                     | 32 (12,3%)                                          | 23 (8,8%)                | 26                     |
| Friuli V. Giulia | 9 (0,8%)                                                | 3 (1,1%)                  | 11                     | 3 (1,1%)                                            | 1 (0,4%)                 | 1                      |
| Lazio            | 43 (3,9%)                                               | 21 (8,1%)                 | 42                     | 24 (9,2%)                                           | 24 (9,2%)                | 22                     |
| Liguria          | 4 (0,4%)                                                | 4 (1,5%)                  | 6                      | 4 (1,5%)                                            | 4 (1,5%)                 | 4                      |
| Lombardia        | 63 (5,7%)                                               | 22 (8,5%)                 | 38                     | 20 (7,7%)                                           | 24 (9,2%)                | 23                     |
| Marche           | 6 (0,5%)                                                | - (-)                     | 6                      | - (-)                                               | 2 (0,8%)                 | 1                      |
| Molise           | 2 (0,2%)                                                | - (-)                     | 3                      | - (-)                                               | - (-)                    | 1                      |
| Piemonte         | 55 (5,0%)                                               | 31 (11,9%)                | 43                     | 31 (11,9%)                                          | 27 (10,3%)               | 28                     |
| Pr. aut. Bolzano | 17 (1,5%)                                               | 8 (3,1%)                  | 13                     | 8 (3,1%)                                            | 4 (1,5%)                 | 6                      |
| Pr. aut. Trento  | 3 (0,3%)                                                | - (-)                     | 2                      | - (-)                                               | - (-)                    | 0                      |
| Puglia           | 47 (4,3%)                                               | 24 (9,2%)                 | 31                     | 3 (11,9%)                                           | 24 (9,2%)                | 30                     |
| Sardegna         | 17 (1,5%)                                               | - (-)                     | 6                      | - (-)                                               | 1 (0,4%)                 | 1                      |
| Sicilia          | 629 (57,1%)                                             | 62 (23,8%)                | 341                    | 66 (23,3%)                                          | 100 (38,3%)              | 72                     |
| Toscana          | 33 (3,3%)                                               | 14 (5,4%)                 | 26                     | 14 (5,4%)                                           | 6 (2,3%)                 | 10                     |
| Umbria           | 4 (0,4%)                                                | 4 (1,5%)                  | 5                      | 4 (1,5%)                                            | 2 (0,8%)                 | 3                      |
| Val d'aosta      | - (-)                                                   | - (-)                     | 0                      | - (-)                                               | - (-)                    | 0                      |
| Veneto           | 26 (2,4%)                                               | 13 (5,0%)                 | 23                     | 10 (3,8%)                                           | 7 (2,7%)                 | 8                      |

Fonte: nostre elaborazioni su dati provenienti dai *Report statistici sulla presenza dei MSNA in Italia*.

## 7. Cenni conclusivi

Pur nel convincimento che i dati forniti dal Ministero del lavoro e delle politiche sociali, desunti dai *Report statistici sulla presenza dei MSNA in Italia*, non possano condurre ad evidenze certe sulla possibile influenza dello scoppio della pandemia sulla presenza dei MSNA sul suolo italiano, è stato comunque possibile individuare un netto trend discendente fino ad inizio 2020 e una ripresa pochi mesi dopo dallo scoppio della pandemia.

L'estate 2020, anche se in maniera differente rispetto alle caratteristiche prese in considerazione (genere, età, nazionalità di provenienza, regione di accoglienza), ha segnato l'inizio della risalita dei flussi di minori che da paesi già in difficoltà economiche, stremati anche dalla pandemia, spesso teatri di guerre civili, hanno visto l'abbandono del proprio paese di origine come unica ancora di salvezza.

Questo aumento della presenza di MSNA non è stato contenuto neanche dalla considerazione che l'Italia, nel periodo del lockdown, si sia definita un luogo non sicuro a causa della impossibilità di garantire una adeguata assistenza sanitaria ai migranti<sup>10</sup> e nonostante l'altra drammatica situazione cui era praticamente impossibile far fronte nel periodo pandemico: in quei mesi, infatti, solo chi era in regola poteva fruire di sussidi economici con le immaginabili conseguenze per chi non lo era e purtuttavia viveva in un'Italia in stato di isolamento nelle proprie abitazioni. Inoltre, la pandemia causata dal Covid-19 ha messo in crisi alcuni settori economici (agricoltura, industria manifatturiera) nei quali da sempre trovano maggiormente collocazione gli immigrati e ciò ha fatto sì che anche coloro i quali fossero in regola, andassero incontro a delle notevoli difficoltà economiche.

Quindi si può verosimilmente ipotizzare che le difficoltà economiche, acuite dal periodo pandemico, oltre ad una diminuzione dei vincoli legislativi per la permanenza sul territorio italiano, siano state tra le principali cause dell'aumento dei MSNA, soprattutto a partire dall'estate 2020.

Per una ricerca più completa, a questo approccio meramente quantitativo ci si propone di affiancare prossimamente uno studio psicologico-sociale su come la pandemia abbia cambiato la vita dei MSNA, già presenti nelle strutture di accoglienza in Italia e su come invece abbiano affrontato le problematiche della pandemia stessa coloro che hanno lasciato i propri paesi per dirigersi verso il nostro, ritenendolo più sicuro rispetto ai territori di origine.

## Riferimenti bibliografici

- AA.VV. *I dimenticati: Coloro che non sono ripartiti dopo la pandemia*, Infinito Edizioni, Formigine (MO), 2020
- AA.VV., Osservatorio sui Minori Stranieri Non Accompagnati in Italia. Primo Rapporto. Centro Studi di Politica internazionale (CeSPI), 2020.
- Brauzzi, F., Galli, D., & Sodano, V. Il ruolo dell'inserimento nei CPIA nel processo di crescita e autonomia dei MSNA. Osservatorio Nazionale sui Minori Stranieri non Accompagnati, CeSPI. *Approfondimento* n. 7/settembre 2020
- Audino, F., & Bianchi, L. (2021). L'accoglienza socio-educativa dei MSNA e dei neomaggioranni nella pandemia da Covid 19. *Educazione interculturale*, 19(1), 78-90.

---

<sup>10</sup> In base alla Dichiarazione di Amburgo del 1979, ogni Stato, allertato dall'autorità di un altro Stato, è obbligato ad intervenire in caso di persone nelle loro acque e a indagare sulla loro identità e sul loro status.

- Carbone, I., Ceschi, S., & Lunardini, M., L'evoluzione della tutela volontaria per minori stranieri non accompagnati. Osservatorio Nazionale sui Minori Stranieri non Accompagnati, CeSPI. *Approfondimento* n. 10/settembre 2021.
- Castiglioni, M., Agostinetto, L., & Bobbo, N. (2020). Migrazioni Minori. Vulnerabilità, salute e pratica educativa. *Journal of Health Care Education in Practice*, 2(1)
- Coin, F., & Banzato, M. (2021). Minori non accompagnati al tempo del covid-19: la tecnologia digitale rafforza o indebolisce le relazioni?. *FORMAZIONE & INSEGNAMENTO. Rivista internazionale di Scienze dell'educazione e della formazione*, 19(1), 505-512.
- Contino, A. C., Di Piazza, S., Federico, A., Federico, M., Frazzica, G., & La Spina, (2021) A. Giovani e legalità in tempo di pandemia. Centro di studi ed iniziative culturali Pio La Torre.
- Corsi, C. (2020). Migranti e immigrati di fronte all'emergenza coronavirus: tra vecchie e nuove fragilità. *Diritto pubblico*, 26(3), 901-942.
- Giampaolo, M. & Ianni, A., Covid-19 e le migrazioni: le percezioni di alcuni stati membri dell'UE. Volti delle migrazioni, Focsiv, novembre 2020.
- Hoseiny, A. A., Afghanistan: un infinito esodo di minori non accompagnati. Osservatorio Nazionale sui Minori Stranieri non Accompagnati, CeSPI. *Approfondimento* n. 10/settembre 2021.
- ISTAT, Statistiche Report, 26 ottobre 2020
- Lunardini, M., Il rischio di una mancata identificazione dell'apolidia nel passaggio alla maggiore età. Osservatorio Nazionale sui Minori Stranieri non Accompagnati, CeSPI. *Approfondimento* n. 11/ottobre 2021
- Masiero, G., & Vitale, G. (2021). La bolla nella bolla: esperienze di bambini stranieri e minori non accompagnati durante il lockdown. *La bolla nella bolla: esperienze di bambini stranieri e minori non accompagnati durante il lockdown*, 104-121.
- Ministero del Lavoro e delle politiche sociali, *La comunità bangladesese in Italia. Rapporto annuale sulla presenza dei migranti*, 2020.
- Ministero del Lavoro e delle politiche sociali, Direzione generale dell'immigrazione e delle politiche di integrazione, Divisione II, *Report mensile Minori Stranieri Non Accompagnati (MSNA) in Italia*, dati mensili da gennaio 2018 a ottobre 2021.
- Ministero del Lavoro e delle Politiche Sociali, Report di monitoraggio MSNA. Dati al 30 giugno 2018, al 31 dicembre 2018, al 30 giugno 2019, al 31 dicembre 2019, al 30 giugno 2020, al 31 dicembre 2020, al 30 giugno 2021
- Pitzalis S. (2020), Il continuum dell'emergenza. Criticità strutturali e mutamenti nel sistema di accoglienza prima e durante la pandemia da covid-19, Supplemento n. 9 a «Illuminazioni», n. 53, luglio-settembre.

Save the Children Italia Onlus (2017), 1° Atlante Minori Stranieri non accompagnati – Prima di tutto i bambini.

Save the Children Italia Onlus (2018), 2° Atlante Minori Stranieri non accompagnati - Crescere lontano da casa.

# Big Data, Intelligenza Artificiale e turismo sostenibile

Aurélie P. Scarafino<sup>1</sup>, Antonella Massari<sup>2\*</sup>, Claudia Marin<sup>3</sup>

<sup>1,2</sup>Dipartimento di Economia Management e Diritto dell'Impresa Università degli Studi di Bari Aldo Moro, <sup>3</sup> Dipartimento di Scienze della Formazione, Psicologia, Comunicazione, Università degli Studi di Bari Aldo Moro

**Riassunto:** Il contributo illustra un'idea progettuale, *InPuglia 360°*, che adopera i *Big Data*, l'*Intelligenza Artificiale* (AI), la *realtà aumentata* (AR) e la *realtà virtuale* (VR), per essere all'avanguardia nel settore turistico nell'ottica della sostenibilità, in linea con i *Sustainable Development Goals* (SDGs) dell'agenda 2030 dell'ONU. L'idea prevede la costituzione di una *Destination Management Organization* (DMOs) per promuovere l'interoperabilità tra gli stakeholder del settore turistico attraverso l'accesso ai dati e un sito che sfrutta l'AI per fornire servizi per il monitoraggio continuo e tempestivo della destinazione, della performance della stessa e dei suoi fruitori. Il progetto comprende delle applicazioni per esperienze in AR e VR, importanti per l'*edutainment* e l'*experience*, e consente di ottenere tutte le informazioni necessarie alla creazione del *dynamic packaging* in un unico portale favorendo la scoperta di itinerari turistici alternativi e slow.

**Keywords:** big data, intelligenza artificiale, interoperabilità, turismo sostenibile.

## 1. Introduzione

Il settore turistico per evolversi, alla luce della recente pandemia, necessita di tre fattori fondamentali per un comparto che presenta una clientela digitalizzata, autonoma ed esigente: *big data*, nuove tecnologie e sostenibilità. L'Intelligenza artificiale (AI) e i big data consentendo la raccolta di dati tempestivi, di ottenere *insight*

---

\* Autore corrispondente: [antonella.massari@uniba.it](mailto:antonella.massari@uniba.it)

Il lavoro è frutto di un progetto comune, tuttavia è possibile attribuire la redazione del sottoparagrafo 2.1 e del paragrafo 3 a Aurélie P. Scarafino, i paragrafi 1, 2 e 5 ad Antonella Massari e il paragrafo 4 a Claudia Marin.

ed effettuare profilazioni accurate degli utenti, possono supportare il turismo nell'ottica della sostenibilità, in linea con alcune delle finalità previste dai *Sustainable Development Goals* (SDGs) dell'agenda 2030 dell'ONU.

Secondo le stime del *World Travel and Tourism Council*, nel 2019, il 9,5% del PIL dell'Unione europea (UE) era direttamente e indirettamente imputabile al turismo. Nello stesso anno le esportazioni di servizi turistici hanno generato per la UE entrate per circa 428 miliardi di euro, pari al 2,6% del PIL. L'impatto della pandemia di Covid-19 sul turismo internazionale in base alle stime dell'Organizzazione mondiale del turismo ha fatto registrare un forte calo dei flussi turistici internazionali nel 2020 compreso tra il 58% e il 78% rispetto all'anno precedente (Banca d'Italia 2020).

Il turismo rappresenta per l'Italia uno dei settori maggiormente rilevanti per l'economia; la sua misurazione è possibile tramite il Conto Satellite del Turismo e la contabilità nazionale; altre misurazioni scaturiscono dalle tecniche usate per la rilevazione delle statistiche del turismo (*enterprise survey*, *accommodation survey*, *household survey* e *frontier survey*), come quelle relative a spostamenti e flussi turistici, e quelle inerenti alle grandezze che misurano il fenomeno fisico della mobilità, come arrivi, presenze e permanenza media (Candela, 2014). I big data, che scaturiscono dalle tracce lasciate dai turisti per mezzo delle nuove tecnologie, sono diventati essenziali per avere stime tempestive e puntuali relative al settore turistico, utili a migliorare l'esperienza turistica proponendo nuove forme di relazione con il viaggiatore (G.E. Feierherd ed altri, 2019) e centrali per lo sviluppo tecnologico del settore stesso.

Grazie agli Open Data, ai progetti di infrastrutturazione e allo sviluppo di applicazioni mobile, si parla sempre più di *Smart tourism* (P. Del Vecchio ed altri 2018); il processo di innovazione tecnologica favorisce la trasformazione della destinazione turistica in *smart destination* rendendola sempre più all'avanguardia, sostenibile e digitale.

L'idea progettuale, *InPuglia 360°*, prevede la costituzione di una *Destination Managemet Organization* (DMOs) che, tramite i dati e la realizzazione di un sito web e applicazioni in *realtà aumentata* (AR) e *virtuale* (VR), punta al miglioramento della gestione della destinazione turistica nell'ottica della sostenibilità. L'obiettivo principale è quello di adoperare *l'Intelligenza artificiale* (AI) e i big data per il monitoraggio tempestivo della destinazione turistica, la performance della stessa e dei suoi fruitori; consentendo di ottenere tutte le informazioni necessarie alla creazione del *dynamic packaging* in un unico portale e favorendo la scoperta di itinerari turistici alternativi e slow.

I dati aggiornati e tempestivi, ottenuti tramite il sito web smart, possono aiutare gli stakeholder nello sviluppo di nuovi servizi e nella profilazione del *prosumer* permettendo di gestire in maniera più efficace i flussi turistici e i fenomeni di sovraffollamento; al contempo la ricerca scientifica disporrebbe di informazioni tempestive necessarie allo studio del fenomeno turistico in una specifica località. La DMO smart presenta una struttura complessa volta a favorire l'interoperabilità, attraverso l'accesso ai dati, da parte dei diversi portatori d'interesse del settore turistico, per una gestione coordinata e sostenibile dello stesso (L. Sheehan, 2016).

## 2. Dalla DMO alla DMO smart

Il Destination Management, secondo la definizione dell'*United Nations World Tourism Organization* (UNWTO), è la gestione coordinata di tutti gli elementi che compongono una destinazione turistica (attrazioni, accesso, marketing, risorse umane, immagine e prezzi). Il ruolo della DMO è quello di guidare e coordinare le attività nell'ambito di una strategia coerente con il perseguimento di obiettivi comuni; è un organismo di natura pubblica o pubblica-privata che promuove, commercializza e gestisce i flussi turistici svolgendo attività di marketing per mezzo di portali e siti web all'interno di un ambito territoriale: Paese, Stato, Provincia, Territorio, Regione, comprensorio o città (C. Tripodi, 2014).

La definizione di DMO, anche grazie all'ausilio delle nuove tecnologie, sta subendo un'evoluzione in *DMO smart* (DMOs) che per il futuro sarà un agente intelligente della destinazione, in grado di indentificare, coinvolgere e apprendere da diversi stakeholder sia all'interno che all'esterno della destinazione, capace di acquisire, filtrare e analizzare dati e informazioni provenienti da varie fonti per creare conoscenza utile alla gestione della destinazione (L. Sheehan ed altri, 2016).

La suddetta gestione deve essere responsabile e sostenibile, ma anche atta a favorire le interazioni tra i visitatori, l'industria turistica e le comunità ospitanti, con l'obiettivo di favorire la costituzione di destinazioni intelligenti che si fondino su cinque pilastri fondamentali: governance, innovazione, tecnologia, accessibilità e sostenibilità (UNWTO, 2019).

### 2.1 La DMOs di "InPuglia 360°"

Il progetto "InPuglia 360°" si propone di istituire una DMOs, di natura pubblico-privata (Associazione Temporanea di Scopo, Fondazione o Consorzio) per ottimizzare l'interoperabilità tra i diversi attori del settore turistico e, sfruttando l'accesso

ai dati e la tecnologia, per produrre informazioni utili e nuovi servizi personalizzati e sostenibili. L'accesso ai dati è reso possibile tramite un sito web *smart* e *user-friendly* che agevola la ricerca delle informazioni e il successivo processo d'acquisto dei servizi da parte dei turisti, considerano che i viaggiatori apprezzano i siti web delle DMO ben strutturati, organizzati, con servizi convenienti e di facile comprensione (J.L. Stienmetz ed altri, 2013). Il sistema, per mezzo degli *human-generated data*, raccoglie dati semi-strutturati e strutturati (D. Bennato, 2015) da analizzare e condividere con i vari stakeholder. L'utilizzo del *machine learning* favorisce la ricerca del turista autoproduttore riducendo il suo prezzo ombra e aumentando la sua permeanza all'interno della destinazione. Il sistema offre un'assistenza h24 a 360°, effettuata da una chatbot che aiuta l'utente fornendo risposte rapide a domande difficili da trovare nei motori di ricerca.

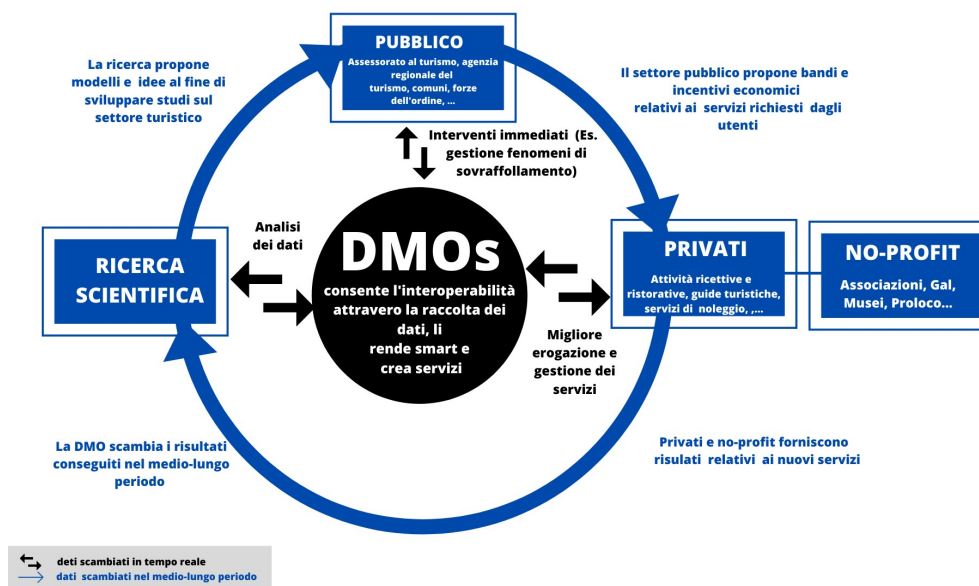
Gli obiettivi che la DMOs del progetto cerca di raggiungere tramite le nuove tecnologie sono relativi ai seguenti cinque punti:

1. *Governance*: migliorare l'interoperabilità tra gli attori del comparto turistico, supportare l'area marketing nelle strategie di comunicazione e sponsorizzazione della destinazione turistica;
2. *Innovazione*: effettuare studi e ricerche scientifiche sulla base dei dati ottenuti dalle interazioni degli utenti con il sito web smart che consentano alla destinazione di innovarsi, supportare gli stakeholder nella progettazione di possibili scenari e di nuovi servizi per mezzo di indagini accurate che aiutino ad anticipare i trend e personalizzare l'offerta;
3. *Tecnologia*: perfezionare l'offerta turistica, l'*experience*, l'acquisto del turista autoproduttore e la creazione del suo *dynamic packaging* in un unico portale, incentivare l'edutainment;
4. *Accessibilità*: migliorare l'accesso ai dati fornendo informazioni di qualità, agevolare l'accessibilità del turista disabile all'interno della destinazione e rendere la destinazione accessibile da remoto;
5. *Sostenibilità*: rendere la destinazione sostenibile dal punto di vista economico sociale e ambientale in linea con i goal dell'agenda 2030 dell'ONU.

Il sito web smart e le due applicazioni in realtà aumentata e virtuale, sono gli strumenti adoperati dalla DMOs per raggiungere i suddetti obiettivi; altresì tali strumenti consentono di migliorare le conoscenze del fenomeno turistico relativo alla destinazione, necessarie alla gestione di una destinazione intelligente e fattore chiave di successo, sia per il pubblico che per il privato. Una DMOs fondata sulle

nuove tecnologie e sulle conoscenze può operare come learning organization (L. Sheehan ed altri, 2016) e raggiungere gli obiettivi prefissati attraverso l'interoperabilità partendo dai dati.

**Figura 1.** Schema dell'interoperabilità tramite l'accesso ai dati della DMOs.



La DMOs raccoglie grandi moli di dati, li rende smart e crea servizi. I dati raccolti vengono scambiati in tempo reale, tramite sito web, con la ricerca scientifica che li acquisisce allo scopo di analizzarli generando modelli e informazioni utili per il settore pubblico, privato e no-profit; gli stessi dati sono accessibili alla pubblica amministrazione, la quale può porre in essere interventi immediati o di lungo periodo come, ad esempio, la gestione di fenomeni di sovraffollamento, per una migliore organizzazione della destinazione turistica; l'interscambio di dati con il settore privato e no-profit agevola l'erogazione e il coordinamento dei servizi turistici. Nel lungo periodo le informazioni scambiate e accessibili a tutti gli stakeholder portano a porre in essere una serie di azioni: i ricercatori dopo aver analizzato il fenomeno propongono modelli e idee per lo sviluppo di nuove politiche economiche, ambientali; il settore pubblico, grazie ai dati provenienti dal sito e da fonti esterne, può proporre bandi ed incentivi economici generando nuove forme di turismo e di business; privati e no-profit possono promuovere nuove *amenities* scam-

biando con la DMOs i risultati di gestione ottenuti con l'implementazione dei servizi richiesti. Le informazioni vengono recepite dalla DMOs e condivise con i ricercatori al fine di ottenere riscontri sui modelli e sulle idee proposte. La DMOs favorendo l'interoperabilità, tramite l'accesso ai dati, porta la destinazione turistica verso una maggiore competitività e la rende un agente intelligente favorendo uno sviluppo sostenibile.

### 3. Il sito web di “InPuglia 360°”

Il progetto prevede la creazione di un sito smart, che attraverso la raccolta e l'analisi dei dati sia capace di aggiornarsi costantemente migliorando la *user experience* e il tasso di conversione (il rapporto tra il numero dei visitatori che accedono al sito web e quanti compiono una determinata azione) e di un'applicazione correlata al sito, dall'architettura semplice e intuitiva, *user-friendly*, con un menù composto da quattro pagine: crea, ispirati, acquista e informati.

Nello specifico le quattro *call to action* inserite nel menù della *landing page* ripropongono il processo decisionale del turista e consentono di effettuare le seguenti azioni:

CREA - Selezionando la parola crea l'utente ha la possibilità, attraverso un questionario a risposta multipla, di ottenere l'automatica generazione di una serie di informazioni e itinerari che rispondano alle sue specifiche esigenze e aspettative. La pagina presenta una serie di domande necessarie a profilare l'utente. In particolare, l'AI, come tecniche di machine learning, fornisce informazioni tramite l'accesso ai dati prodotti dall'interazione con gli utenti e promuove itinerari alternativi, slow, sostenibili e capaci di migliorare l'experience, in coerenza con le esigenze e i desideri del turista. In tal modo la DMOs raccoglie informazioni e grandi moli di dati attendibili, tempestivi e di immediata interpretazione. Successivamente l'utente seleziona i servizi che ritiene interessanti o che desidera contestualmente acquistare ottenendo una mappa con le informazioni e l'itinerario costruito, da stampare, conservare sul proprio smartphone o utilizzare per un successivo acquisto. Attraverso l'intera procedura è possibile ottenere una profilazione del turista nel rispetto delle regole previste dal General Data Protection Regulation, consentire all'AI di apprendere dalle risposte fornite dall'utente favorendo la visualizzazione di itinerari e indicazioni sempre più precise e corrette, facilitare il turista autoproduttore riducendo il suo prezzo ombra (Candela, 2014). Gli itinerari proposti possono incentivare forme di turismo sostenibile e slow, come percorsi culturali, cammini, ciclo-

**Figura 2.** Sito web pagina crea: questionario a risposta multipla, servizi ed itinerari da selezionare, mappa con itinerario scelto e informazioni utili all'utente



ISPIRATI - Scegliendo questa pagina il turista può visionare il diario di bordo di altri viaggiatori o aprire un account pubblicando immagini, video, testi e podcast. I racconti dei viaggiatori, i loro commenti, le loro recensioni, saranno fonte di ispirazioni e anche un modo per attingere, in maniera semplice e veloce, tutte quelle informazioni che comunemente vengono ricercate sui social, sui blog e nei siti di recensione. La creazione di tali contenuti genera dati da adoperare per la *sentiment analysis*.

ACQUISTA - La pagina può essere raggiunta con tre percorsi differenti: in seguito alla creazione della propria mappa; alla sezione informati con la selezione dei ser-

vizi da acquistare oppure entrando nell'*e-commerce*. L'utente in pochi e semplici passaggi acquista una o più *amenities*, ottenendo informazioni in tempo reale sul grado di affollamento di una determinata area ed eventualmente richiedendo un'alternativa al fine di non superare la *carrying capacity* della località stessa. Una volta effettuato l'acquisto il sistema genera un *Qr-code* che consente al turista di utilizzare un unico codice per accedere a tutti i servizi selezionati e che riduce lo spreco di carta.

INFORMATI - La pagina informati è composta da due sezioni: la *chatbot*, che offre assistenza h24 interagendo con l'utente, e una sezione di macro e micro aree di interesse (ristorativo, ricettivo, ecc...) che agevola la ricerca delle informazioni desiderate.

Il sito è progettato in modo da ridurre i tempi di ricerca, migliorare l'usabilità e le asimmetrie informative tipiche dei servizi *experience good* (Candela, 2014).

#### **4. Le applicazioni in realtà aumentata e virtuale**

L'idea progettuale prevede la creazione di due applicazioni in AR e VR che andrebbero ad aumentare l'*experience* dei turisti raccogliendo informazioni utili per i loro interessi e spostamenti all'interno del territorio scelto come destinazione; in tal modo migliora l'accessibilità alla destinazione stessa e si favorisce l'*edutainment* e la sostenibilità. Le aree da preservare, al fine di evitare il loro depauperamento causato dal sovraccarico delle presenze, ne trarrebbero vantaggio in particolar modo in tempo di pandemia.

La AR per mezzo di smartphone, tablet e occhiali, come i *Google Glass*, supporta il turista durante il viaggio guidandolo all'interno della destinazione (A. Mandić ed altri, 2019).

AR "InPuglia360°" è l'applicazione da utilizzare arrivati a destinazione, selezionando la località e la tipologia di turismo, per accedere ad una serie di contenuti da vivere in AR. In questo modo è possibile scoprire il territorio favorendo forme alternative di turismo, come il turismo slow e la scoperta di itinerari enogastronomici, cineturistici, culturali e naturalistici. L'applicazione permettendo al viaggiatore di visitare nuove località spostandosi virtualmente all'interno del territorio potrebbe dar vita a nuove forme di business.

**Figura 3.** AR “InPuglia360°”



La VR consente, tramite un visore, un’interazione con l’ambiente ricreato digitalmente in cui è possibile vedere gli oggetti e manipolarli (Montagna L., 2018) offrendo, in ambito turistico, l’opportunità di sperimentare una destinazione in anticipo favorendo un marketing efficace, nel quale l’attività in VR è un fattore chiave nella commercializzazione della destinazione (Kim ed altri, 2018).

VR “InPuglia360°” è l’applicazione in VR da vivere on location o da remoto, nel secondo caso il sistema consente di conoscere la destinazione permettendo di ridurre le criticità tipiche dei servizi experience good; l’applicazione agevola il marketing nella promozione migliorando la fase del dreaming e del sharing e aumenta l’accessibilità della destinazione, offrendo ai disabili la possibilità di viaggiare senza muoversi da casa e al contempo elimina le barriere architettoniche all’interno della destinazione stessa favorendone l’accessibilità. L’esperienza in VR della destinazione turistica potrebbe essere particolarmente utile in caso di pandemia.

**Figura 4.** VR “InPuglia360°”



#### 4. Conclusioni

La creazione di una *DMO smart* permette alla destinazione turistica di essere all'avanguardia attraverso l'utilizzo delle nuove tecnologie, favorendo l'interoperabilità tra i vari attori del settore turistico migliorando l'accesso ai dati, la gestione della destinazione e la produzione di nuovi servizi nell'ottica della sostenibilità.

In particolare, il sito web smart del progetto consente di ottenere dati sempre aggiornati e scambiati in tempo reale; ottimizza i tempi di ricerca del turista auto-produttore favorendo la commercializzazione della destinazione turistica; promuove nuove forme di turismo sostenibile e contribuisce allo sviluppo della ricerca scientifica in tale settore.

Il turismo sostenibile non rappresenta un nuovo tipo di turismo, ma piuttosto un processo manageriale volto alla massimizzazione dei benefici, anche economici, della comunità locale e non solo a promuovere istanze verdi, coprendo un ampio spettro di attività (Confalonieri, 2008).

Importanti benefici del progetto sono quelli relativi alla destagionalizzazione della destinazione turistica che potrebbe scaturire dalla proposta di itinerari alternativi e slow, da vivere anche in periodi di bassa stagione, e la creazione di nuove possibilità di business e di guadagno.

Un aspetto importante riguarda la conoscenza in tempo reale dei flussi turistici per consentire una migliore gestione dei fenomeni di sovraffollamento, essenziale in particolare in periodo di pandemia, ma sempre necessaria al fine di preservare il patrimonio naturalistico e architettonico di una destinazione turistica attraverso la riduzione della *carrying capacity*. La creazione di itinerari turistici slow promuove forme di turismo responsabile e sostenibile dal punto di vista sociale e ambientale generando nuovi cicli di vita per la destinazione e rigenerando quelli in fase di maturità. La AR migliora l'*experience*, educa divertendo, riduce le barriere architettoniche rendendo la destinazione sempre accessibile. La VR consente, in caso di pandemia, di poter viaggiare da remoto, rendendo la destinazione accessibile anche a soggetti disabili. Il progetto "*InPuglia 360°*" nel raggiungere i suoi obiettivi è in linea con alcuni dei goal previsti dall'agenda 2030 dell'ONU.

La DMOs consentendo una gestione coordinata della destinazione turistica tra più attori del settore, favorisce l'indispensabile interoperabilità tra gli stessi per lo sviluppo coerente e sostenibile del settore turistico, reso altresì possibile dalle nuove tecnologie (AI, AR e VR) e dall'uso dei big data.

L'idea progettuale promuove l'innovazione creata attraverso ecosistemi di dati ed interoperabilità per mezzo dell'accesso ai dati e può rappresentare il punto di partenza per la creazione di infrastrutture digitali che puntino, attraverso l'AI e i dati, al raggiungimento di obiettivi di sostenibilità sociale, economica e ambientale del settore turistico.

### **Riferimenti bibliografici**

- Candela G.; Figini P. (2014). *Economia del turismo e delle destinazioni*, McGraw-Hill Companies, Milano. ISBN 478-88-386-7466-2. Confalonieri M. (2008). *Il turismo sostenibile*. G.Giappichelli editore- Torino. ISBN 978-88-348-8481-2.
- Del Vecchio P., Ndou V., Passiante G. (2018), *Turismo digitale e smart destination. Tecnologie, modelli e strategie per la crescita di un sistema turistico integrato*. FrancoAngeli, Milano. ISBN:9788891776488, 8891776483.
- Feierherd G.E., González F., Viera L.; Romano L. Delia L.N., Hertas F., Depertis B.O. (2019). *Realidad virtual y aumentada, big data y dispositivos móviles: aplicaciones en turismo*, XXI Workshop de Investigadores en Ciencias de la Computación, ISBN: 978-987-3948-85-3.
- Kim M.J., Lee C., Jung T. (2018). Exploring Consumer Behavior in Virtual Reality Tourism Using an extended Stimulus-Organism-Response Model, *Journal of Travel Research* vol. 59 (1) 69-89, DOI: 10.1177/0047287518818915.
- Mandić A. e Praničević D.G. (2019). The Impact of ICT on Actors Involved in Smart Tourism destination Supply Chain, *e-Review of Tourism research (eRTR)* vol. 16, ISSN: 1941-5842. URL <https://journals.tdl.org/ertr/index.php/ertr/article/view/337/109>.
- Montagna L. (2018). *Realtà virtuale e realtà aumentata. Nuovi media per nuovi scenari di business*, Ulrico Hoepli, Milano. ISBN 978-88-203-8317-6.
- Sheehan L., Vargas-Sánchez A., Presenza A., Abbate T. (2016). The use of intelligence in tourism destination management: an emerging role for DMOs. *International journal of tourism research* 18(6), DOI: 10.1002/jtr.2072.
- Stienmetz J.L., Lavy S.E., Boo S., (2013). Factors Influencing the Usability of Mobile destination Management Organization Websites. *Journal of Travel Research*. Vol. 52 Issue 4, 2013. URL <https://doi.org/10.1177/0047287512467698>.
- Tripodi C. (2014). *Costruire il successo delle destinazioni turistiche. Destination management e strategie di sviluppo del territorio*. EGEA S.p.A., Milano. ISBN 978-88-238-4424-7.
- UNWTO (2019). *Guidelines for Institutional strengthening of Destination Management Organization (DMOs)- Preparing DMOs for new challenges*. ISBN: 978-92-844-2083-4.

**Sitografia**

[https://www.bancaditalia.it/pubblicazioni/indagine-turismo-internazionale/2020-indagine-turismo-internazionale/statistiche\\_ITI\\_05062020.pdf](https://www.bancaditalia.it/pubblicazioni/indagine-turismo-internazionale/2020-indagine-turismo-internazionale/statistiche_ITI_05062020.pdf) Banca D'Italia Eurosystem (2020), Indagine sul territorio internazionale, Fascicoli collana Statistiche.

# L'esperienza della didattica a distanza in alcune scuole pugliesi

Nunziata Ribecco<sup>1</sup>, Angela Maria D'Uggento<sup>1\*</sup>,  
Arianna Ruggiero<sup>2</sup>

<sup>1</sup> Dipartimento di Economia e Finanza, Università degli Studi di Bari Aldo Moro,

<sup>2</sup> Dottore in Scienze statistiche

**Riassunto:** La pandemia Covid-19 ha travolto il mondo della formazione in modo del tutto inaspettato e le scuole si sono organizzate al meglio per assicurare la continuità didattica ai propri studenti e consentire loro di seguire le lezioni da casa. Le famiglie hanno cercato di contribuire con gli strumenti tecnologici necessari ma le difficoltà sono state tante e sono andate ad aggiungersi al sentimento di angoscia e isolamento che ha colpito molti giovani. Attraverso una indagine condotta su 1.580 studenti delle scuole superiori di Bari e provincia si è cercato di indagare su aspetti rilevanti quali benessere sociale, salute mentale e sulla relazione tra didattica a distanza (DAD) e rendimento scolastico. Le analisi svolte hanno delineato un quadro assai migliore delle aspettative, sia rispetto al confronto tra DAD e lezioni in presenza, sia sul rendimento degli studenti. L'Analisi delle corrispondenze multiple ha evidenziato che circa l'80% degli studenti non ha mai pensato di abbandonare la scuola mentre coloro che hanno pensato di farlo "qualche volta" o "spesso", si trovavano o in una precaria condizione occupazionale dei genitori o in una condizione emotiva di fragilità psicologica causata dai disagi provati durante la pandemia.

**Keywords:** didattica a distanza; rendimento scolastico; emozioni; pandemia Covid19; analisi delle corrispondenze multiple.

## 1. Introduzione

La Pandemia di Covid19 ha travolto le istituzioni scolastiche in modo del tutto inatteso, imponendo un cambiamento radicale degli approcci didattici tradizionali,

---

\* Autore corrispondente: [angelamaria.duggento@uniba.it](mailto:angelamaria.duggento@uniba.it). Il lavoro qui descritto è frutto di un progetto comune, ma N. Ribecco ha provveduto alla redazione dei paragrafi 1 e 2, A. M. D'Uggento i paragrafi 3.1 e 4, A. Ruggiero i paragrafi 3.2 e 3.3.

principalmente basati sulla didattica frontale, a favore dell'adozione della didattica a distanza (DAD). Tale provvedimento si è reso necessario per superare le criticità dovute alla impossibilità di recarsi a scuola in presenza per rispettare le misure di isolamento adottate per fronteggiare la diffusione del virus COVID19. Da marzo 2020, la chiusura delle scuole a causa della pandemia ha fatto sì che oltre il 90% dei bambini e dei ragazzi iscritti a scuola in tutto il mondo fosse costretta a rimanere a casa.

L'Italia, essendo stato il primo Paese ad essere colpito dalla pandemia, ha sperimentato il lockdown nazionale prima di tutti gli altri stati europei. Tra la fine di febbraio e la prima settimana di marzo 2020, le scuole e le università sono state chiuse subito nell'Italia settentrionale (Lombardia, Emilia-Romagna, Liguria, Piemonte, Veneto e Friuli-Venezia Giulia) e poi in tutto il resto della nazione e tale situazione è perdurata fino a settembre 2020. Una così lunga assenza da scuola ha procurato grandi preoccupazioni tra docenti e genitori circa le lacune formative e il connesso rendimento scolastico da parte degli studenti.

Per assicurare l'apprendimento, seppur a distanza, varie soluzioni sono state adottate, prevalentemente avvalendosi delle piattaforme online gestite da provider internazionali (Google Meet, Cisco Webex, Zoom, etc). Contestualmente, il Ministero dell'Istruzione italiano ha stanziato 85 milioni di euro per le attività di formazione a distanza, di cui 70 milioni per dotare di dispositivi digitali e connettività gli studenti in condizioni economiche disagiate. Le scuole hanno ricevuto una dotazione di 10 milioni di euro per l'acquisizione di piattaforme educative digitali e di 5 milioni per la formazione degli insegnanti.

Secondo l'ISTAT, sulla base dei dati rilevati nel 2019 sulla connettività Internet delle famiglie italiane, è possibile ipotizzare che circa 3 milioni di bambini e ragazzi tra i 6 e i 17 anni potrebbero aver avuto difficoltà nella DAD durante il lockdown per la difficoltà con la connettività o con il nuovo approccio didattico.

Interessanti informazioni di prima mano sull'esperienza di un campione di famiglie che hanno sperimentato didattica e apprendimento da remoto sono fornite dall'indagine condotta nel giugno 2020 nell'ambito di un progetto coordinato dal Joint Research Center della Commissione Europea e da Unicef (Mascheroni et al, 2021). L'indagine è stata condotta mediante questionario somministrato online ad un campione di bambini, ragazzi (tra i 10 e i 18 anni) e genitori italiani che disponevano di accesso a internet. I risultati, malgrado il *bias* dovuto a non aver intervistato ragazzi che non utilizzano Internet, evidenziano delle difficoltà legate alla qualità delle connessioni e all'utilizzo degli stessi dispositivi in condivisione tra genitori e figli, anche nelle stesse fasce orarie e in ambienti sovraffollati. In parti-

colare, il 27% dei genitori ha riferito di non avere abbastanza dispositivi digitali per le necessità scolastiche o lavorative e molti hanno acquistato nuovi dispositivi da utilizzare contemporaneamente o per sostituire dispositivi obsoleti inadeguati. Circa la metà ha acquistato almeno un nuovo smartphone, mentre il 41% delle famiglie ha acquistato almeno un nuovo computer e il 46% delle famiglie intervistate ha ricevuto nuovi dispositivi digitali dalla scuola. Infine, una famiglia su quattro ha utilizzato una connessione internet procurata dalla scuola per le lezioni a distanza.

Questi dati dimostrano l'importanza di investire risorse finanziarie ingenti per l'apprendimento a distanza degli studenti (Mascheroni, G. et al, 2021) e per supportare gli studenti caratterizzati da divario digitale ed economico.

Sulla base di questo quadro a livello nazionale, il presente contributo illustra i principali risultati di una indagine condotta su 1.580 studenti delle scuole di Bari e provincia partecipanti al progetto "Piano Nazionale Lauree Scientifiche"- PLS Statistica. Il progetto PLS Statistica, che nasce con lo scopo di diffondere la cultura statistica tra gli studenti degli ultimi anni della scuola secondaria di 2° grado, prevede, oltre l'attività di "Formazione degli insegnanti" e quella di "Autovalutazione", che ha come obiettivo la verifica della preparazione iniziale con la finalità di accrescere la consapevolezza delle competenze acquisite e di quelle necessarie per affrontare il percorso universitario, quella del "Laboratorio per le scienze di base" che si sviluppa attraverso la realizzazione di un'indagine che ha lo scopo di coinvolgere gli studenti nell'analisi di problemi reali, addestrandoli nella risoluzione di tali problemi con l'ausilio dei metodi statistici e sviluppando il loro spirito critico nell'applicazione di detti metodi e nella interpretazione dei risultati (Ribecco et al, 2019).

L'indagine, svolta nell'anno scolastico 2020/2021, riguarda l'esperienza dell'epidemia da COVID-19 vissuta dai giovani, tema molto attuale che ha suscitato l'interesse da parte degli studenti delle scuole partecipanti e ha determinato sicuramente un maggiore impegno nel raggiungimento dei risultati. Il contributo si articola nel modo seguente: dopo l'introduzione, nel paragrafo 2 sono illustrati l'indagine e i metodi statistici utilizzati; nel paragrafo 3 si discutono i principali risultati ottenuti e, infine, nel paragrafo 4 sono presentate brevi considerazioni conclusive.

## 2. Materiali e metodi

Il questionario, sviluppato in collaborazione tra gli studenti delle scuole partecipanti, gli insegnanti, i docenti e gli studenti tutor iscritti al corso di laurea triennale in Scienze Statistiche dell'Università di Bari coinvolti nel PLS e coordinati dai docenti universitari impegnati nella realizzazione del progetto, è stato implementato con Google moduli ed è stato somministrato online a tutti gli studenti delle scuole partecipanti. Esso è organizzato in sei sezioni finalizzate a cogliere diversi aspetti:

1. Dati anagrafici
2. Fonti informative
3. Benessere sociale e salute mentale
4. Decreti istituzioni e rispetto delle regole
5. DAD e rendimento scolastico
6. Prospettive future.

I dati ottenuti sono stati analizzati dagli stessi studenti con l'ausilio degli insegnanti responsabili per ogni scuola e dei referenti universitari.

Attraverso i dati raccolti sono stati approfonditi gli aspetti riguardanti DAD e rendimento scolastico e, in particolare, il punto di vista degli studenti in merito alla didattica a distanza, quali sono i fattori per i quali preferiscono la didattica in presenza alla DAD. Inoltre, si è analizzato se è presente una associazione tra l'eventualità dell'abbandono scolastico e la condizione occupazionale dei genitori, la fragilità provocata dalla pandemia e il carico di studio durante la didattica a distanza; ove possibile, i dati sono stati confrontati con quelli a livello nazionale. Infine, è stata valutata la significatività delle differenze riscontrate tra il rendimento prima e durante la DAD.

## 3. Risultati

### 3.1 *Analisi descrittiva*

La composizione degli studenti per genere è perfettamente bilanciata (femmine 50,3% vs maschi 49,7%), con età media 16,2 anni ( $\pm 1,4$  anni). Per quanto riguarda le informazioni sulla famiglia, il livello di istruzione dei genitori è medio-alto e la situazione economica non sembra essere preoccupante dal momento che il 94,3% dei padri è occupato, così come il 65,9% delle madri (Tabella 1). Questo dato potrà essere utile per valutare le dotazioni tecnologiche a disposizione per la DAD.

**Tabella 1.** *Distribuzione percentuale del titolo di studio e della condizione occupazionale dei genitori*

| Titolo di studio         | %<br>madre | %<br>padre | Condizione occupazionale  | %<br>madre | %<br>padre |
|--------------------------|------------|------------|---------------------------|------------|------------|
| Post Laurea              | 4,2        | 3,5        | Lavoratore autonomo       | 16,1       | 31,5       |
| Laurea                   | 16,5       | 14,0       | Lavoratore dipendente     | 49,3       | 60,8       |
| Diploma/maturità         | 50,9       | 48,6       | Non occupato /disoccupato | 32,5       | 3,9        |
| Licenza elementare/media | 27,8       | 32,3       | Pensionato/a              | 0,5        | 2,0        |
| n.d.                     | 0,6        | 1,6        | n.d.                      | 1,6        | 1,8        |
| Totale                   | 100,0      | 100,0      | Totale                    | 100,0      | 100,0      |

Gli studenti hanno espresso una preferenza per la didattica in presenza esprimendo quale principale motivazione il maggior livello di attenzione rispetto alla presenza di varie fonti di distrazione che caratterizzano l'ambiente di casa. Come seconda motivazione è stata indicata la possibilità di interagire con il docente e con la classe seguita dalla maggior comprensibilità e chiarezza delle lezioni in aula e dal minor carico di studio associato alla didattica in presenza.

L'unico aspetto per il quale gli studenti potrebbero preferire la didattica a distanza è caratterizzato dalla gestione dell'ansia per verifiche/interrogazioni che è minore se le attività vengono svolte attraverso lo schermo (Tabella 2).

**Tabella 2.** *Confronto sugli approcci didattici*

| Motivazioni                                     | A distanza | In presenza | Indifferente |
|-------------------------------------------------|------------|-------------|--------------|
| Comprendo meglio le lezioni                     | 11,3       | 47,0        | 41,7         |
| Seguo più attentamente                          | 10,7       | 58,3        | 31,1         |
| E' più facile intervenire/partecipare           | 20,3       | 47,8        | 31,9         |
| Il carico di studio è maggiore                  | 54,1       | 14,6        | 31,3         |
| Mi causano più ansia verifiche e interrogazioni | 12,7       | 53,4        | 33,9         |

Nonostante questi risultati mostrino importanti aspetti a favore della didattica in presenza, alla richiesta di esprimere una preferenza in generale tra le tre opzioni il divario tra didattica a distanza (39,5%) e quella in presenza (36,6%) si è ridotto notevolmente a vantaggio della prima; per il restante 23,9% degli studenti intervistati è indifferente la partecipazione in presenza o a distanza. Questo dato induce ad ipotizzare che potrebbero esserci alcuni aspetti che non sono stati presi in considerazione direttamente nel questionario, come ad esempio la distanza dalla scuola frequentata, l'uso di mezzi di trasporto per raggiungere la scuola o ulteriori fattori che abbiano portato gli studenti ad esprimersi in favore della DAD, sebbene la presenza

di più di un terzo degli intervistati a favore dell'approccio tradizionale faccia ben sperare rispetto al desiderio di maggior coinvolgimento di natura sia sociale che educativa che solo la scuola può dare.

Una delle principali criticità certamente ineludibili nella fase iniziale del lockdown è stata quella riguardante i mezzi attraverso i quali poter realizzare la DAD.

Dal Rapporto ISTAT su disponibilità di strumenti e ambienti di studio con dati a livello nazionale, emerge che il 12,3% dei minori in età scolastica non ha un computer a casa, quota che sfiora il 20% nel Mezzogiorno, praticamente circa 850 mila giovani tra 6 e 17 anni. Un ulteriore 57%, pur in presenza di un pc in casa, non ha un suo dispositivo personale e deve condividerlo con gli altri componenti della famiglia. Soltanto il 6,1% dei ragazzi tra 6 e 17 anni vive in famiglie dove è disponibile almeno un computer per componente. A questo si aggiunga che circa il 41,9% dei minori vive in una abitazione sovraffollata e che il 7% di questi affronta anche un disagio abitativo, come ad esempio la poca luminosità e i problemi strutturali.

La pandemia ha portato il nostro Paese in crisi anche per la presenza di numerose criticità pregresse sul fronte delle disuguaglianze economiche e, di conseguenza, digitali. Non tutti i ragazzi italiani partivano dalle medesime possibilità e tale evidenza si è tradotta nel consistente aumento del fenomeno della dispersione scolastica e in un elevatissimo rischio di abbandono degli studi vero e proprio. L'attività scolastica rappresenta un momento fondamentale di crescita nella vita di un individuo e non è accettabile che i ragazzi ne vengano privati: nel momento in cui l'emergenza sanitaria impedisce lo svolgimento delle lezioni in presenza, la formazione deve essere comunque garantita, senza distinzioni di natura economica o sociale.

### ***3.2 Analisi delle corrispondenze multiple***

Le situazioni diverse in cui si sono trovati, unitamente alle emozioni negative provate durante la pandemia, possono aver reso probabilmente meno sopportabili le giornate di studio per alcuni studenti. Risulta, quindi, interessante esplorare le possibili relazioni tra il pensiero dell'abbandono scolastico, la condizione occupazionale dei genitori, la fragilità/instabilità psicologica provocata dalla pandemia e il carico di studio sopportato durante il periodo di didattica a distanza.

Data la multidimensionalità dell'analisi, si è proceduto con la metodologia delle Corrispondenze Multiple (ACM) una delle tecniche statistiche più utilizzate per l'analisi "multipla" di microdati quantitativi, qualitativi o misti. Tale tecnica per-

mette di riassumere l'intreccio delle relazioni di interdipendenza tra le variabili in uno spazio geometrico-statistico di dimensione ridotta.

In altre parole, il metodo perviene alla costruzione di una serie di variabili latenti (o fattori), combinazione delle variabili originali, che esprimono alcuni concetti non direttamente osservabili nella realtà ma frutto della misurazione di un insieme di variabili. Si assegna, quindi, un significato a tali assi, osservando la disposizione delle modalità delle variabili e cercando contrapposizioni in senso orizzontale e verticale.

Lo *scree plot* ci permette di rappresentare graficamente l'inerzia spiegata da ciascuna dimensione. Nel momento in cui la curva diviene tendenzialmente piatta, possiamo trascurare le successive dimensioni.

Sono state prese in considerazione le prime tre dimensioni perché il numero di componenti è determinato nel punto oltre il quale gli autovalori rimanenti sono tutti relativamente piccoli e di dimensioni comparabili.

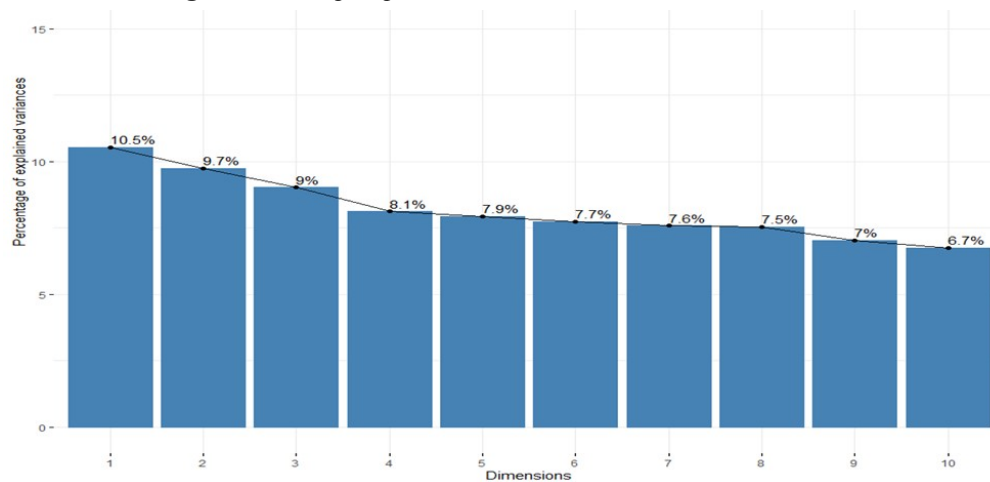
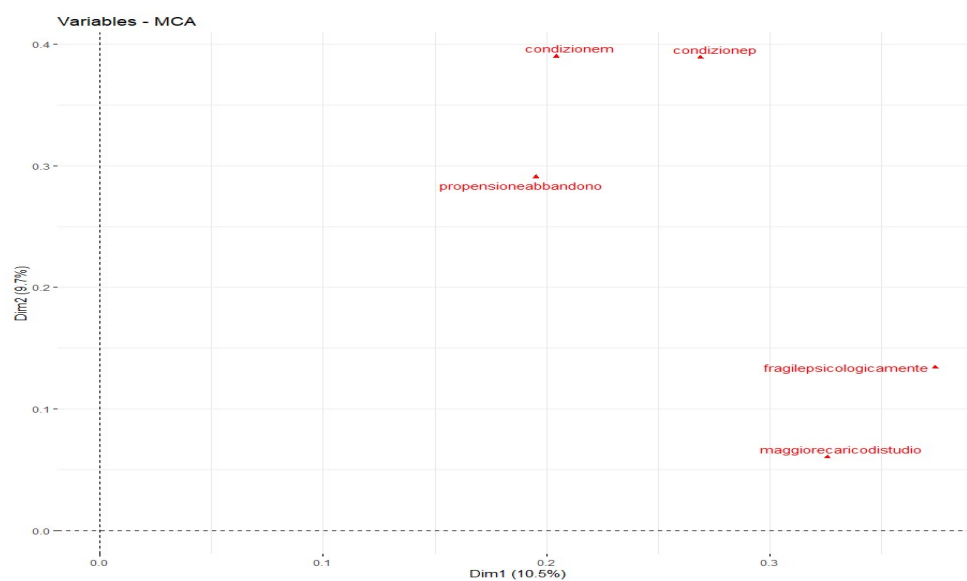
Le dimensioni 1, 2 e 3 spiegano rispettivamente circa il 10,5%, il 9,7% e il 9,0% dell'inerzia totale.

Ciò corrisponde a un totale cumulato del 29,2% dell'inerzia totale trattenuta dalle 3 dimensioni, una quota di variabilità non certamente rilevante.

Per raggiungere una quota di variabilità piuttosto alta e consistente, pari circa all'80%, dovremmo considerare le prime 10 componenti principali, ma non avrebbe senso procedere in tale direzione poiché non sarebbe coerente con la logica di ottenere uno spazio di dimensionalità ridotto.

Inoltre, dato che il fenomeno in oggetto è un fenomeno sociale alquanto complesso e considerato che le variabili sono ottenute attraverso le risposte ad un questionario per lo più politomiche, verosimilmente, sarebbe molto difficile aspettarsi che la variabilità spiegata dalle prime tre componenti possa essere superiore al 30%.

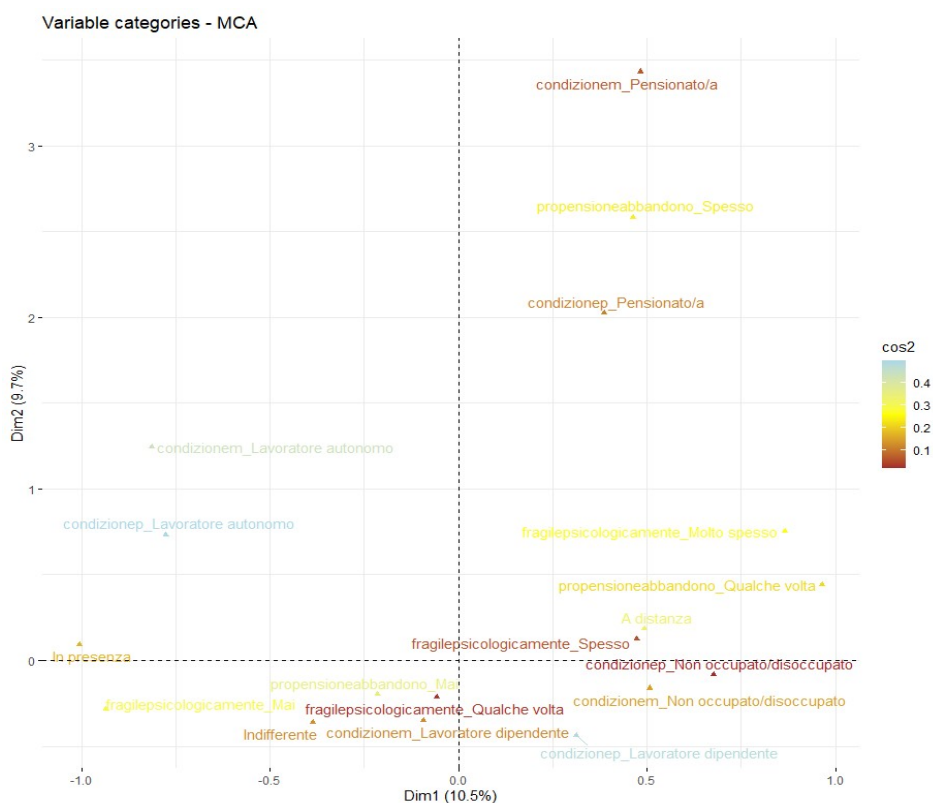
La Figura 2 consente di identificare le variabili più correlate a ciascuna dimensione. Si può notare che le variabili "maggiore carico di studio" e "fragilità psicologica", sono le più correlate con la dimensione 1 mentre le variabili "condizione occupazionale madre", "condizione occupazionale padre" e "propensione abbandono" sono le più correlate con la dimensione 2.

**Figura 1.** *Scree plot per la scelta del numero delle dimensioni***Figura 2.** *Individuazione delle variabili di riferimento per le dimensioni*

La Figura 3 mostra le relazioni tra categorie di variabili attraverso le prime due dimensioni che sono sufficienti a trattenere il 20,2% dell'inerzia totale contenuta nei dati. Il primo asse fattoriale è caratterizzato dal maggior contenuto informativo, il secondo ne ha più del terzo e così via. Inoltre, modalità che nel piano risultano vicine sono maggiormente associate, cioè tendono a coesistere sulle stesse unità statistiche.

Per evidenziare le differenze, colori diversi contraddistinguono le categorie di variabili in base ai loro valori di  $\cos^2$ , che indica il contributo della dimensione alla spiegazione della variabilità di una determinata modalità. Quanto più alto è il coseno al quadrato, tanto più la modalità è ben rappresentata dai due fattori.

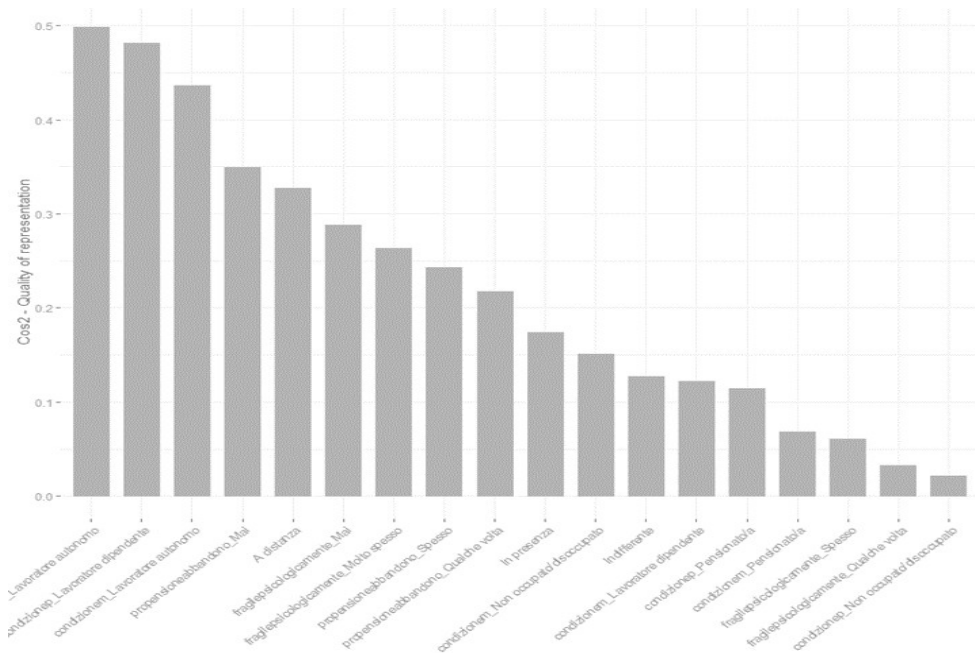
**Figura 3.** MCA: Categorie delle variabili in relazione all'indice  $\cos^2$



Le modalità che sono meglio rappresentate dagli assi fattoriali sono la “condizione di lavoratore autonomo della madre e del padre”, la “condizione di lavoratore dipendente del padre” e “mai” relativo alla propensione all’abbandono scolastico. Le modalità: “condizione di madre e padre pensionati”, “padre disoccupato/non occupato”, “spesso” e “qualche volta” per “fragilità e instabilità psicologica”, risultano avere un contributo relativo prossimo allo zero, per cui esse non sono ben rappresentate dall’asse.

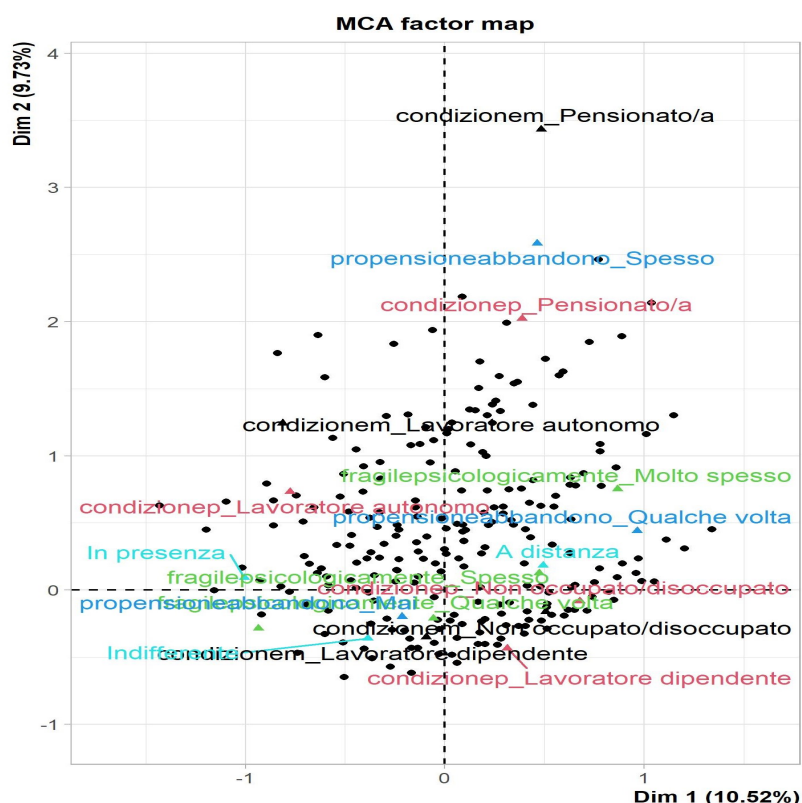
Inoltre, possiamo osservare che il primo asse è in grado di distinguere molto bene la condizione “fragile psicologicamente molto spesso” dalla condizione “fragile psicologicamente mai” e “propensione abbandono mai” dalle altre due modalità “propensione abbandono spesso” e “propensione abbandono qualche volta”. Il secondo asse, invece, distingue molto bene la “condizione occupazionale pensionata”, sia della madre che del padre, dalle restanti condizioni occupazionali.

**Figura 4.** Qualità della rappresentazione delle variabili in ordine decrescente di  $\cos^2$



Dal momento che un individuo può essere considerato come l'insieme delle categorie in cui si trova, può essere collocato al baricentro delle sue categorie. Come è facile notare nella Fig. 5, tutti o quasi tutti gli individui sono collocati nella parte medio bassa.

**Figura 5.** Mappa fattoriale delle variabili di riferimento per le dimensioni 1 e 2



Tale collocazione nel piano evidenzia come la maggior parte degli studenti non abbia mai pensato di abbandonare la scuola, una quota inferiore, invece, ha pensato qualche volta di abbandonare la scuola e ancora meno risultano coloro che hanno pensato spesso di abbandonare la scuola. Nella parte superiore del grafico si collocano le categorie “condizione madre pensionata”, “propensione abbandono spesso” e “condizione padre pensionato” e si posizionano distanti dalla nuvola di punti che racchiude la maggior parte degli individui, perché gli studenti con genitori pensionati e che hanno pensato "spesso" di abbandonare la scuola sono molto pochi. E' interessante osservare, altresì, che la categoria "fragile psicologicamente/ molto

spesso" occupa una posizione completamente opposta alla categoria "fragile psicologicamente/ mai", infatti rispettivamente, le ritroviamo nel primo e terzo quadrante.

Continuando ad osservare la mappa in Fig. 5 sembrerebbe che il pensiero di abbandonare la scuola "qualche volta", rispetto alle posizioni occupate dalla condizione occupazionale del padre, quindi dipendente, autonomo, disoccupato e pensionato, sia più vicina a "condizione occupazionale padre dipendente" e più distante dagli altri; analogo discorso per quanto riguarda la "condizione occupazionale della madre". Inoltre, sempre per quanto riguarda il pensiero di abbandonare la scuola "qualche volta", sembrerebbe che, rispetto alle posizioni occupate dalle modalità relative alla fragilità/instabilità psicologica, ovvero "mai", "qualche volta", "spesso" e "molto spesso", questa sia più vicina a "molto spesso" e più distante dagli altri, mentre, rispetto alle posizioni occupate dalle modalità relative al maggior carico di studio durante la pandemia, ovvero "a distanza", "in presenza" e "indifferente", questa sia maggiormente vicina alla categoria "a distanza" e più distante dalle altre.

Pertanto, si potrebbe affermare che gli studenti che abbiano pensato "qualche volta" di abbandonare la scuola, siano maggiormente coloro che durante la quarantena hanno attraversato frequenti momenti di fragilità psicologica con un carico di studio maggiore e con genitori entrambi non occupati/disoccupati.

Si può ancora notare che la categoria "spesso" relativa alla fragilità/instabilità mentale e la categoria "mai" relativa al pensiero dell'abbandono scolastico sono posizionate molto vicino all'origine, dove si concentra la nuvola di punti rappresentante gli individui. Infatti, la maggior parte degli studenti, non ha mai pensato di abbandonare la scuola e ha durante la quarantena attraversato "spesso" momenti di fragilità psicologica. La categoria "mai" relativa al pensiero di abbandonare la scuola è molto vicina a quasi tutte le categorie della variabile fragilità/instabilità psicologica. L'assenza di una distanza significativa tra queste ci porta a dedurre che la fragilità attraversata durante la pandemia, nella maggior parte dei casi, non abbia comunque distratto i ragazzi dalla volontà di proseguire i propri studi.

### ***3.3 Il rendimento scolastico durante la pandemia***

Un ultimo approfondimento viene dedicato all'analisi del rendimento degli studenti, aspetto che durante la didattica a distanza ha molto preoccupato genitori e docenti. Il sito di "Pagella politica" fornisce, a tal proposito, le informazioni riguardanti i risultati delle prove Invalsi dell'anno scolastico 2020-2021, le prime dopo l'inizio della pandemia di Covid-19, su cui fare importanti considerazioni.

Tali valutazioni riguardano oltre 2,1 milioni di studenti, tra cui circa 475 mila frequentanti l'ultimo anno della scuola secondaria superiore. I risultati riportano che, mentre nelle scuole elementari le valutazioni sono rimaste piuttosto stabili, nelle secondarie di primo e secondo grado si è registrato un calo in Italiano e Matematica, particolarmente marcato in alcune regioni e nelle famiglie più in difficoltà. Al fine di evitare di ipotizzare troppo facilmente dei nessi causali tra pandemia e calo dell'apprendimento è importante sottolineare come tali risultati Invalsi abbiano già segnalato un andamento simile negli ultimi anni. Inoltre, leggere i dati delle prove Invalsi in chiave "anti-dad" sembra un approccio un po' superficiale perché la situazione è così complessa da poter individuare molteplici cause. Le difficoltà registrate sono ascrivibili al contesto pandemico in senso lato e non esclusivamente dalla didattica a distanza. Ciò che emerge è un problema profondo di disuguaglianze, in cui la Dad ha funzionato meglio in famiglie che vivono in determinate condizioni sociali ed economiche. Sarebbe interessante, quindi, leggere i risultati delle prove Invalsi per programmare politiche di ampio spettro di lotta alle disuguaglianze, anche scolastiche.

Ritornando ai dati rilevati sugli studenti partecipanti all'indagine, analizziamo, dunque, come sia cambiato il rendimento degli studenti prima e durante la pandemia.

Dai dati esposti nella Tab. 3 emerge con forte evidenza che il rendimento degli studenti è rimasto assolutamente invariato, salvo un leggerissimo calo nel valore medio che era 7,05 prima della pandemia ed è passato a 7,03 durante la pandemia.

Tale evidenza è anche confermata dai risultati ottenuti dal confronto, mediante il Test di Wilcoxon per campioni appaiati, fra i valori mediani riguardanti il rendimento prima e durante la pandemia ( $p\text{-value} = 0,1049$ ).

**Tabella 3.** *Principali statistiche sul rendimento scolastico degli intervistati prima e durante la pandemia*

| Statistiche  | Votazione in decimi pre-pandemia | Votazione in decimi durante pandemia | Differenza |
|--------------|----------------------------------|--------------------------------------|------------|
| Minimo       | 4,500                            | 4,500                                | 0,000      |
| I quartile   | 7,000                            | 7,000                                | 0,000      |
| Mediana      | 7,000                            | 7,000                                | 0,000      |
| Media        | 7,053                            | 7,034                                | 0,019      |
| III quartile | 7,000                            | 7,000                                | 0,000      |
| Massimo      | 8,500                            | 8,500                                | 0,000      |

## 4. Conclusioni

Malgrado tutte le preoccupazioni espresse da più parti durante il lungo periodo dedicato all'apprendimento da casa, i risultati della nostra indagine sulla didattica a distanza delineano un bilancio complessivamente positivo. Circa il 39% degli intervistati l'ha addirittura preferita alla modalità in presenza ma, probabilmente, nella valutazione complessivamente benevola hanno rivestito un certo peso gli aspetti logistici ed organizzativi legati agli spostamenti casa-scuola, e non solo.

Indubbiamente, gli studenti hanno dovuto affrontare anche alcune difficoltà.

Con riferimento all'esperienza della DAD, è emerso che la maggior parte degli studenti ha riscontrato difficoltà sia nella comprensione delle lezioni che nell'interazione con i docenti, pertanto, ai fini dell'efficacia didattica, la tradizionale modalità in presenza conserva indiscutibili vantaggi, come riconosciuto dagli stessi studenti.

Dalle analisi svolte è, comunque, emerso che circa l'80% degli studenti non ha mai pensato di abbandonare la scuola. Tra coloro che, invece, hanno pensato di farlo con una frequenza pari a "qualche volta" o "spesso", è possibile notare una parziale influenza della condizione occupazionale dei genitori e della condizione emotiva di fragilità psicologica provocata dai disagi e dalle difficoltà vissute durante la pandemia.

Osservando i dati a livello nazionale emerge come il nostro Paese non sia stato decisamente pronto ad affrontare tale emergenza sul fronte delle dotazioni informatiche: secondo ISTAT, solo il 6% dei ragazzi in età scolare vive in famiglie dove è disponibile un computer per componente. Questo, insieme alla non sempre adeguata preparazione dei docenti, è sicuramente uno dei fattori che potrebbe aver inciso più pesantemente sul conseguimento degli obiettivi formativi previsti.

Tuttavia, in ultima analisi e contrariamente alle aspettative, i risultati hanno dimostrato che in Dad il rendimento degli studenti intervistati non è peggiorato e, anzi, è rimasto piuttosto stabile.

## Riconoscimenti

Gli Autori ringraziano per la collaborazione i Dirigenti delle seguenti scuole partecipanti al progetto PLS 2019/2020:

- Liceo G. Bianchi Dottula di Bari;
- I.T.T. Panetti Pitagora di Bari;

- Liceo Scientifico OSA Rita Levi-Montalcini di Molfetta (BA);
- I.T.I.S. Galileo Ferraris di Molfetta (BA);
- Licei Einstein Da Vinci di Molfetta (BA);
- I.T.E.S G. Calò di Francavilla Fontana (BR);
- I.I.S.S. Pertini Anelli di Turi (BA).

In particolare, si ringraziano i docenti e gli studenti per la collaborazione fornita nello svolgimento dell'indagine.

La ricerca qui riportata è stata finanziata dal MIUR, Fondo Progetto Lauree Scientifiche PLS Statistica a.a. 2019/2020.

### **Riferimenti bibliografici**

- Agresti, A. (2002). *Categorical Data Analysis*, 2nd ed. New York: Wiley.
- Beh, E.J., Lombardo, R. (2014). *Correspondence Analysis Theory, Practice and New Strategies*. New York: John Wiley & Sons, Inc.
- Benzécri, J.P. (1992). *Correspondence Analysis Handbook*. New York: Marcel Dekker.
- Greenacre, M. (2016), *Correspondence Analysis in Practice*. Boca Raton: Chapman and Hall/CRC, III ed.
- Greenacre, M.J. (1984). *Theory and Applications of Correspondence Analysis*. London: Academic Press.
- Greenacre, M.J. (1994). Multiple and Joint Correspondence Analysis. In: M. J. Greenacre and J. Blasius (ed.), *Correspondence Analysis in the Social Sciences*. London: Academic Press, 141-161.
- ISTAT (2020) Rapporto annuale 2020. La situazione del paese. Spazi in casa e disponibilità di computer per bambini e ragazzi, Aprile 2020.
- Le Roux, B., Rouanet, H. (2004). *Geometric Data Analysis: From Correspondence Analysis to Structured Data Analysis*. Dordrecht: Kluwer Academic Publications.
- Lebart, L., Morineau, A., Warwick, K.M. (1984). *Multivariate Descriptive Statistical Analysis*. New York: John Wiley & Sons, Inc.
- Mascheroni, G., Saeed, M., Valenza, M., Cino, D., Dreesen, T., Zaffaroni, L. G., Kardefelt-Winther D. La didattica a distanza durante l'emergenza COVID-19: l'esperienza italiana. Centro di Ricerca Innocenti dell'UNICEF, Firenze, 2021.
- Ribecco N., Calculli, C. D'Uggento, A.M. (2019). *Laboratorio di Statistica per l'informazione e la conoscenza: l'esperienza PLS dopo dieci anni di attività*. WIP Edizioni, Bari.

Yelland, P. (2010). An introduction to correspondence analysis. *The Mathematica Journal*, 12, 1-23.

### **Sitografia**

Pagella politica. <https://pagellapolitica.it/>

# Distributive adaptation of the Lorenz curve to a Squirecle

Carlo Cusatelli<sup>1</sup>, Crescenzo Gallo<sup>2\*</sup>, Alessandro Rinaldi<sup>3</sup>

<sup>1</sup> Ionian Department, University of Bari Aldo Moro,

<sup>2</sup> Department of Clinical and Experimental Medicine, University of Foggia,

<sup>3</sup> PhD in Statistics, University of Bari Aldo Moro

**Abstract:** The concentration curve represents the relationship between cumulative relative fractions of statistical units and cumulative relative fractions of intensity of a transferable phenomenon between statistical units. In this article, the Lorenz curve is studied through a mathematical approach and ideally approximated to an arc belonging to a particular type of super circles or squirecle defined precisely as "super-ellipse". From this elliptical curve we can finally deduce all the possible ramifications that the Lorenz curve can assume, starting from the case of equidistribution up to (almost) the case of maximum concentration.

**Keywords:** Lorenz curve; Super-ellipse; Squirecle.

## 1. Introduction

In the statistical field variability can be expressed -in addition to indexes such as variance, deviance or mean squared deviation- through concentration, referring to all phenomena having the property of transferability, i.e. those for which it is possible to transfer part of the total amount of a character from some individuals to others, up to the limit where a single subject can get the full amount of the character itself. In this case we are talking about eterogeneity (maximum concentration case): e.g. income, assets, number of employees of a company, number of cars of a family or number of passengers of an airline.

On the other hand, if the total amount of a character were equally distributed among all subjects, it would be homogeneity (case of fair distribution).

---

\* Corresponding author: [crescenzo.gallo@unifg.it](mailto:crescenzo.gallo@unifg.it)

The work described here is the result of a common project attributable in equal parts to the coauthors.

It is possible to represent the concentration in two ways:

1. an index that shows the degree of concentration of the character or the concentration ratio;
2. the split concentration or Lorenz curve.

The Lorenz curve is used to graph the relationship established between cumulative relative fractions of intensity<sup>1</sup>  $q_i = A_i/A_s$  owned by respective cumulative relative fractions of statistical units  $p_i = n_i/N$ .

If  $N \rightarrow \infty$  the points  $(p_i, q_i)$  for which the concentration section passes become closer and closer to each other, until the splits takes the form of a real curve. In this way, the so-called "Lorenz curve" is obtained, which can assume different ramifications depending on how the phenomenon turns out to be more or less concentrated.

The purpose of this article is to approximate the Lorenz curve to an arc of a super-circle (or squircle) -a special case of superellipse that can be expressed with a Supershape- in such a way as to be able to represent and interpolate (through the family that derives from it) all the possible distributional approximations of Lorenz split curve, i.e. the entire family of Lorenz curves.

From a methodological point of view, the present work has been developed from an exquisitely theoretical perspective, giving a contribution of a purely deterministic and not stochastic nature, with the aim of demonstrating that through a family of super circles it is possible to approximate and interpolate -conceptually and in purely mathematical terms- a Lorenz split curve to an elliptical arc forming part of a Superellipse. From this elliptical curve we can finally deduce all the possible ramifications that the Lorenz curve can assume, starting from the case of equidistribution up to (almost) the case of maximum concentration.

## 2. Explicit representation of the squircle function

The super-circle or squircle is a geometric shape that combines the properties of a square with those of a circle.

In the Cartesian reference system, the supercircle centered on a point  $(a, b)$  with axes parallel to the coordinates is described by the equation:

$$(x - a)^4 - (y - b)^4 = r^4 \quad [1]$$

<sup>1</sup> Assuming that a stochastic  $X$  variable expresses the distribution of  $s$  intensities  $x_1 < x_2 < \dots < x_i < \dots < x_s$ , proceeding with the cumulative sums you have:

$$A_1 = x_1 \cdot n_1; A_2 = A_1 + x_2 \cdot n_2; \dots A_i = A_{i-1} + x_i \cdot n_i; \dots A_s = A_{s-1} + x_s \cdot n_s$$

where  $n_1 + n_2 + \dots + n_i + \dots + n_s = N$ .

where  $r$  stands for the min or radius of the supercircle.

The supercircle is a special case ( $n=4$ ) in the class of geometric shapes known as "supercircles", defined by the equation:

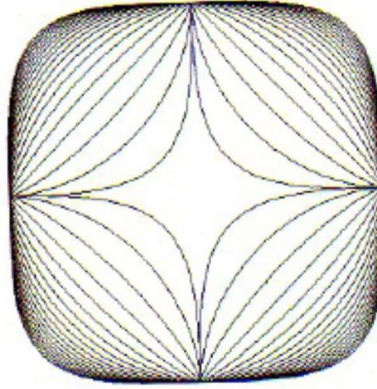
$$|x - a|^n + |y - b|^n = |r|^n \quad [2]$$

This is a special case of superellipse which have as equation (Fig. 1):

$$\left| \frac{x-a}{r_a} \right|^n + \left| \frac{y-b}{r_b} \right|^n = 1 \quad [3]$$

where  $r_a$  and  $r_b$  are the lengths of the major and minor semi-axes, respectively:

**Figure 1.** Multiple branches of a Squire superellipse.



In Euclidean geometry, super-ellipse or "Lamé curve" (Fig. 2) means a geometric figure that generalises the ellipse. In a Cartesian coordinate system it is described as the place of the  $(x, y)$  points that meet the equation:

$$\left| \frac{x}{a} \right|^n + \left| \frac{y}{b} \right|^n = 1 \quad [4]$$

with  $n, a$  and  $b$  positive reals ( $n, a, b \in \mathbf{R}^+$ ). For  $n = 4$  we get the formula:

$$\left| \frac{x}{a} \right|^4 + \left| \frac{y}{b} \right|^4 = 1 \quad [5]$$

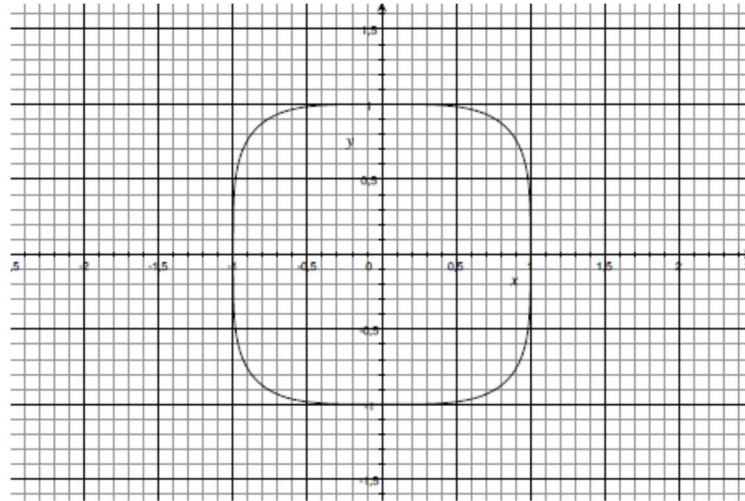
from which it derives:

$$\frac{y^4}{b^4} = 1 - \frac{x^4}{a^4} \quad [6]$$

Finally, by isolating the variable  $y$  you get:

$$y = f(x) = \sqrt[4]{b^4 \left( 1 - \frac{x^4}{a^4} \right)} \Rightarrow b \cdot \sqrt[4]{1 - \frac{x^4}{a^4}} \quad [7]$$

**Figure 2.** Superellipse  $|x|^4 + |y|^4 = 1$  with center in the origin of the axes ( $a=b=1$ ).



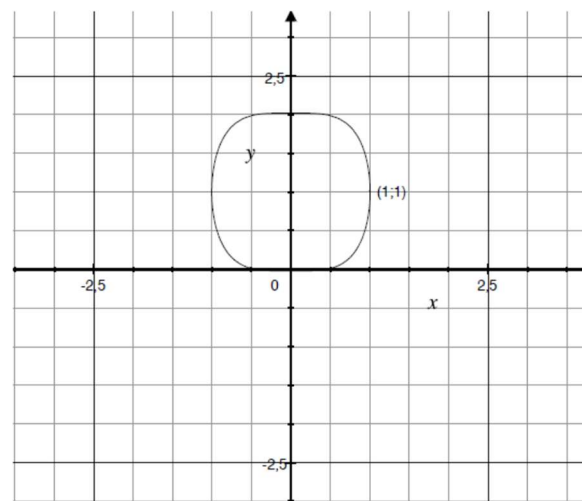
By placing the (6) on the form:

$$\frac{x^4}{a^4} + \frac{4y(y-2)}{b^4} = 1 \quad [8]$$

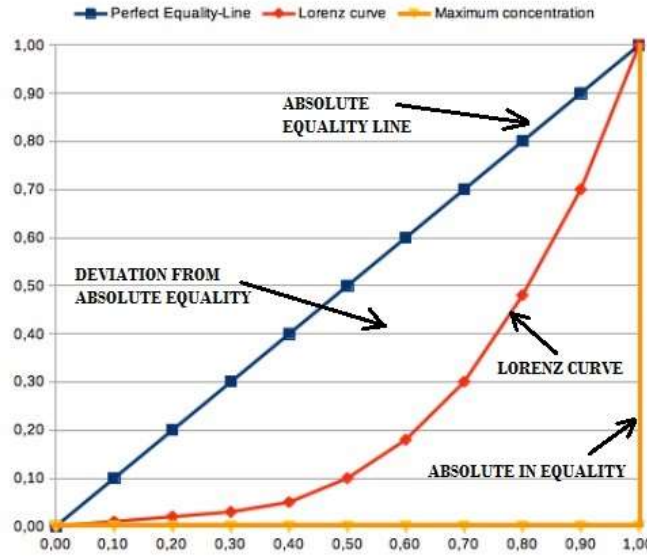
for  $a=0.45$ ,  $b=0.64$  you get the super-ellipse shown in Fig. 3.

Obviously for  $n \rightarrow \infty$  you will have the rectangle resting on the  $x$  axis passing through the  $(0; 0)$ ,  $(0; -1)$ ,  $(-1; 1)$ ,  $(-1; 2)$ ,  $(0; 2)$ ,  $(1; 2)$ ,  $(1; 1)$  and  $(0; 1)$  points. The arc passing through the  $(0; 1)$  point of the super-ellipse represents the *Lorenz Curve* that for  $n \rightarrow \infty$  will touch the axes representing the maximum concentration (Fig. 4).

**Figure 3.** Superellipse passing through the origin of the axes.



**Figure 4.** Representation of the Lorenz curve in the case of equidistribution and in the case of maximum concentration.



### 3. Lorenz curve model and elliptical arc length

The analytical model of the empirical concentration (or Lorenz) curve is given by the expression:

$$q = L(p, d), \quad 0 \leq p \leq 1 \quad [9]$$

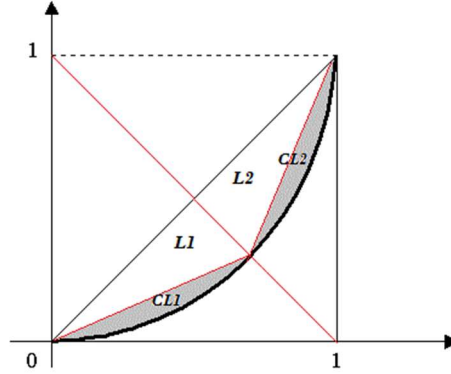
where  $d = (d_1, d_2, \dots, d_m) \in \Theta \subseteq \mathbf{R}^m$  is the parameter vector that identifies the particular element in the function family and  $\Theta$  represents the parametric space subset of the  $m$ -dimensional Euclidean space  $\mathbf{R}^m$  (Tarsitano 1987).

Necessary and indispensable conditions for a Lorenz curve can be comparable or approximate to an elliptical arc of a *squirrel* (or super circle) are 5: positivity (as a real function defined in the closed interval  $[0; 1]$ ), non-decrease, convexity, uniqueness for equidistribution and asymmetry.

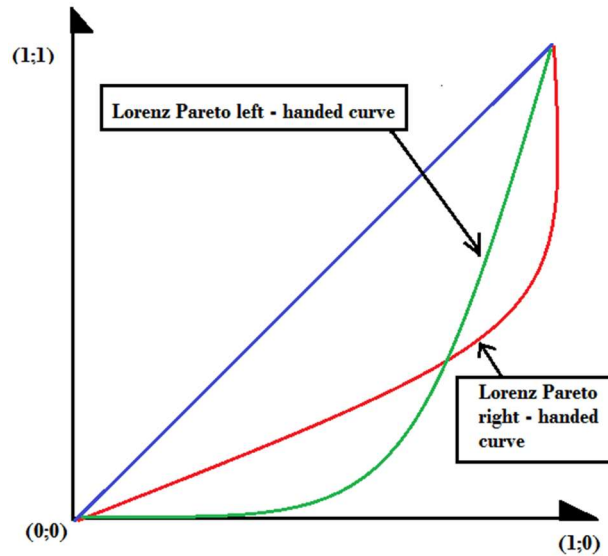
Tarsitano (2000) states and establishes that a Lorenz curve is defined as symmetric if it satisfies the following condition:

$$L(p_D + p) - p = L(p_D - p) - p \quad \text{per } 0 \leq p \leq \min\{p_D, 1 - p_D\} \quad [10]$$

Zanardi (1965) states that a concentration curve is defined as symmetrical if  $L1=L2$ , or if  $C_{L1} = C_{L2}$ , however this requirement does not determine a necessary and sufficient condition for obtaining perfect symmetry as according to (Giurovich, 1959)  $C_{L1}$  and  $C_{L2}$  they can come from different distributions (Fig. 5).

**Figure 5.** *Symmetry of Lorenz curve for sub-areas.*

Condition of symmetry is very important for the case under study, because if this requirement does not occur, it will have a generalization of the Lorenz curve that is obtained when the distribution model is that of Pareto (Fig. 6).

**Figure 6.** *Representation of the Lorenz-Pareto curve in the generalized case.*

Battistini, Paolillo, Servente (2005) define that the concentration of Lorenz curve of a continuous random variable  $X$  as a function:

$$L(p) = \mu^{-1} \int_0^p F_X^{-1}(t) dt, \text{ with } 0 \leq p \leq 1. \quad [11]$$

where  $\mu = E(X)$  represent the expected value,  $F$  indicated cumulative distribution with density  $f$  and  $t$  indicates the time factor.

In Dancelli (1989) a reference is made to the model of the Lorenz curve with 3 parameters for personal incomes distribution indicated by C. Dagum (1977-1980), Fattorini - Lemmi (1979), Dagum - Lemmi (1987) and Lemmi (1987) characterized by the following distribution function:

$$L(p) = \begin{cases} (1 + \theta^\delta x^{-\delta})^{-\beta} & \text{if } x > 0 \\ 0 & \text{somewhere else} \end{cases} \quad [12]$$

From formula (12) we derive the cartesian equation of the Lorenz curve in the variables  $p$  and  $q$  (Dagum 1977-1980 and Dancelli 1986-1989):

$$L(p) = \frac{\int_0^{p^{1/\beta}} z^{\beta+1/\delta-1} (1-z)^{-1/\delta} dz}{\int_0^1 z^{\beta+1/\delta-1} (1-z)^{-1/\delta} dz}, \quad p \in [0,1] \quad [13]$$

where  $z = x^\delta / (\theta^\delta + x^\delta)$  and with  $(\theta, \beta) > 0$  and  $\delta > 1$ .

In formula (13), if the parameters  $\beta$  and  $\delta$  change, the various trends of the Lorenz curve attributable to a super-elliptical arch, can be highlighted illustrated in Dancelli (1989).

As Tarsitano (1987) describes the length of the curve the value:

$$S = \int_0^1 \sqrt{1 + L'(p, d)^2} dp \quad [14]$$

it can be deduced, on the basis of (14), that as  $d$  changes our curve will or will not tend to reach the case of maximum concentration.

From above, we can say that:

$$\left( \int_0^1 \sqrt{1 + L'(p, d)^2} dp \right) 4 = \text{lenght Squire} \quad [15]$$

The index (14) varies between  $\sqrt{2}$  in the case of no concentration, and 2 in the case of maximum concentration.

From formula (15) it can be deduced that the sum of the segments that make up the Lorenz curve coincides with its length, in formula we have:

$$\sum_{i=1}^s \sqrt{(p_i - p_{i-1})(q_i - q_{i-1})} = \int_0^1 \sqrt{1 + L'(p, d)^2} dp \quad [16]$$

The formula (14) implies that the *length of super-ellipse* will be:

$$4 \left( \sum_{i=1}^s \sqrt{(p_i - p_{i-1})(q_i - q_{i-1})} \right) = 4 \left( \int_0^1 \sqrt{1 + L'(p, d)^2} dp \right) \quad [17]$$

On the basis of what has been said, it can be deduced that the relatively modest requirements imposed on the  $L(p, d)$  model give rise to a wide range of analytical forms of Lorenz curves that can be interpreted as arcs from the entire vast range of superellipses-Supershapes referring, however, only to the positive Cartesian quadrant. The purpose of this work is therefore to demonstrate teoretically how a Lorenz curve can be traced, back through appropriate interpolation techniques, to an elliptical arc of a supercircle in the case of perfect symmetry.

#### 4. Superellipse in the general concept of Supershape

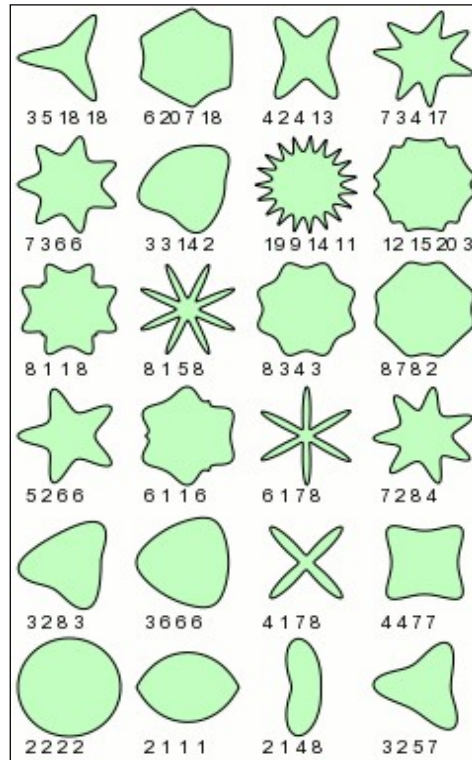
What has been treated up to now is essential in order to be able to trace the Lorenz curve back to a particular type of superellipse defined *Supershape* (which is nothing more than a generalization of circular functions in two dimensions in polar coordinates) that allows -through the modification of some parameters- to obtain a large number of geometric shapes, called "Superformulas" .

The equation of a superformula is by:

$$\frac{1}{p} = \sqrt[n_1]{\left|\frac{1}{a}\cos\frac{m}{4}\phi\right|^{n_2} + \left|\frac{1}{b}\sin\frac{m}{4}\phi\right|^{n_3}} \quad [18]$$

where  $p$  and  $\phi$  are the polar coordinates,  $a, b, m, n_1, n_2, n_3$  are real numbers (Fig. 7).

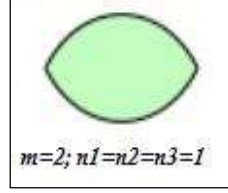
**Figure 7.** Some examples of parameters  $m, n_1, n_2, n_3$  for the superformula (with  $a=b=1$ ).



If in (18) we put  $a=b=n_1=n_2=n_3=1, m=2$  we obtain the equation:

$$\frac{1}{p} = \left| \cos \frac{1}{2}\phi \right| + \left| \sin \frac{1}{2}\phi \right| \quad [19]$$

which is shown in Fig. 8.

**Figure 8.** Example of superformula for  $a=b=n_1=n_2=n_3=1$ ,  $m=2$ .

Being the area of a superellipse given by<sup>2</sup>:

$$A = 4b \int_0^a \left[ 1 - \left( 1 - \frac{x}{a} \right)^n \right]^{\frac{1}{n}} dx = \frac{4^{1+\frac{1}{n}} b \sqrt{\pi} \Gamma\left(1 + \frac{1}{n}\right)}{\Gamma\left(\frac{1}{2} + \frac{1}{n}\right)} \quad [20]$$

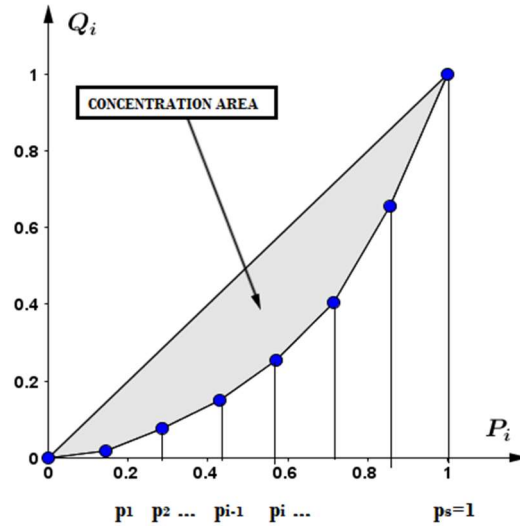
we deduce that the area of a quadrant of the same is given by:

$$\frac{A}{4} = \frac{4b \int_0^a \left[ 1 - \left( \frac{x}{a} \right)^n \right]^{\frac{1}{n}} dx}{4} \quad [21]$$

from which it derives:

$$\frac{4b \int_0^a \left[ 1 - \left( \frac{x}{a} \right)^n \right]^{\frac{1}{n}} dx}{4} + \sum_{i=1}^s (p_i - p_{i-1})(q_i + q_{i-1}) = 1 \quad [22]$$

where the  $\sum_{i=1}^s (p_i - p_{i-1})(q_i + q_{i-1})$  represent the sum of the products of the bases and the heights of trapezes located under the curve corresponding to the various intensity classes of the character (Fig. 9).

**Figure 9.** Representation of the Lorenz curve in the case of a statistical variable by classes.

<sup>2</sup> <http://mathworld.wolfram.com/Superellipse.html>

From formula (22) derives the *concentration ratio* (Girone 2009).

$$1 - \sum_{i=1}^s (p_i - p_{i-1})(q_i + q_{i-1}) = R \text{ (Concentration ratio)} \quad [23]$$

considering the fact that the successive differences of the distribution function correspond to the relative frequencies:

$$p_i - p_{i-1} = y_i \quad [24]$$

is possible use quick formula:

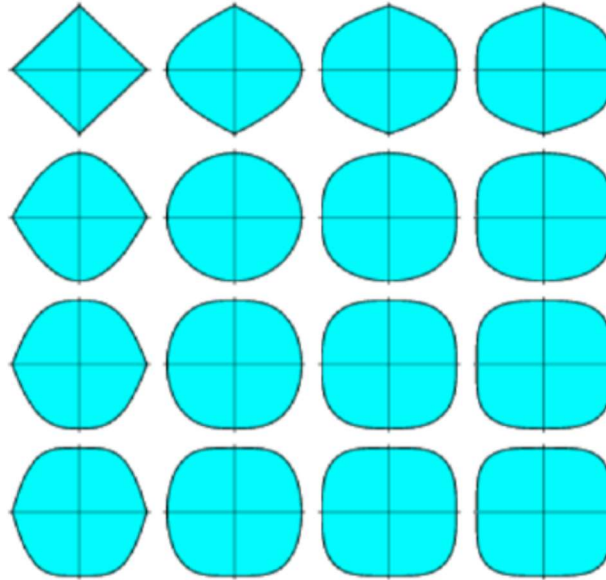
$$1 - \sum_{i=1}^s y_i(q_i + q_{i-1}) = R \quad [25]$$

The graphs in Fig. 10 show a generalization of the super-ellipse provided by the formula:

$$|x|^p + |y|^q \leq 1 \quad [26]$$

for  $p=1, \dots, 4$  and  $q=1, \dots, 4$ .

**Figure 10.** Generalizations of a super-ellipse.



Gielis (2003) considered the further generalization of the super-ellipse given in polar coordinates by:

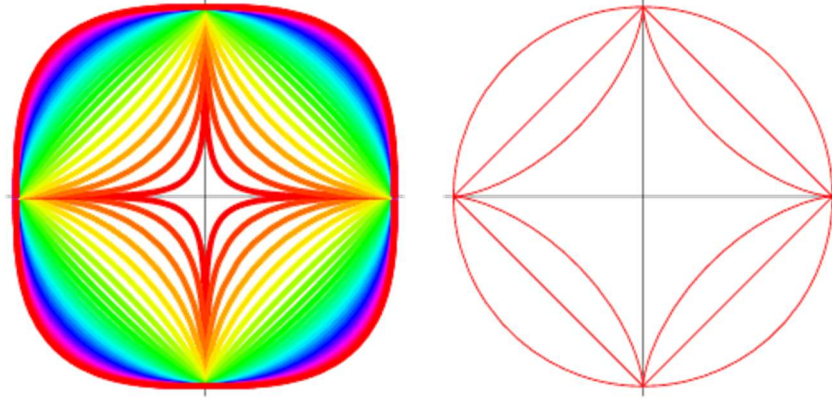
$$r(\theta) = \left[ \left| \frac{\cos(\frac{1}{4}m\theta)}{a} \right|^{n_2} + \left| \frac{\sin(\frac{1}{4}m\theta)}{b} \right|^{n_3} \right]^{\frac{1}{n_1}}. \quad [27]$$

Here, the introduction of the  $m$  parameter and the use of polar coordinates give rise to curves with  $m$ -fold rotational symmetry.

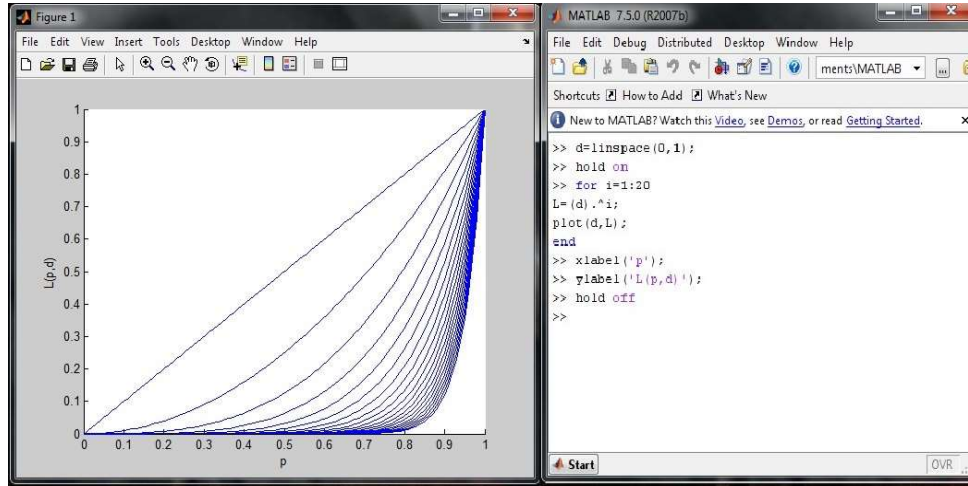
Fig. 11 shows a superellipse representation obtained with the Matlab software (<http://mathworld.wolfram.com/Superellipse.html>). Finally Fig. 12 shows the

various ramifications of the Lorenz curve expressed by the mathematical model  $q=L(p, d)$  with software Matlab and the correlative script.

**Figure 11.** Representation of superellipse in Matlab.



**Figure 12.** Branches of the Lorenz curve in mathematical model  $L(p,d)=q$  in Matlab.



## 5. Final remarks

The present work has the objective of combining and interpolating the Lorenz curve with a super-elliptical arc of Squircle-type belonging to a super-circle, so as to be able to analyze the canonical concentration curve from a deterministic rather than a stochastic point of view.

The statistical topic has been studied and analyzed from a purely mathematical-

deterministic point of view, with a conceptual and cognitive contribution that has served not only to point out some characteristic aspects of the Lorenz curve, but at the same time to explore and analyze interesting types of geometric figures such as the super-ellipses (represented in some cases by Squirele) that, when certain parameters vary, assume the most singular forms.

In conclusion, the purpose at the moment is only to lay the foundations and ideas for reflections and research for further future developments, trying to show how a deterministic-mathematical model is comparable or traceable to a probabilistic empirical stochastic distribution.

### Bibliographic references

- Abramowitz M., Stegun I. A. (1964). *Handbook of mathematical functions: with formulas, graphs, and mathematical tables*, Vol. 55. Courier Corporation.
- Alzer H., Richards K. (2017). *Inequalities for the ratio of complete elliptic integrals*. Proceedings of the American Mathematical Society 145, 4, 1661–1670.
- Battistini E., Paolillo P.L., Servente G. (2005). Nuovi indicatori della dispersione insediativa nella valutazione ambientale strategica: un'applicazione dell'indice di Gini alla pianura cremonese. In: P.L. Paolillo (a cura di), *Rendiconti cremonesi. Il contributo del Politecnico di Milano alla conoscenza delle dinamiche evolutive nel territorio di Cremona*. Milano: Clup, 163–200.
- Carlson B. (2002). *Three improvements in reduction and computation of elliptic integrals*. Journal of research of the National Institute of Standards and Technology 107, 5, 413.
- Dagum C. (1977). *A new model of personal income distribution: specification and estimation*. Economie Appliquée, XXX, 3, 413–436.
- Dagum C. (1980). *The generation and distribution of income, the Lorenz curve and the Gini Ratio*. Economie Appliquée, XXXIII, 2, 327–367.
- Dagum C., Lemmi A. (1987). *A contribution to the analysis of income distributions and income inequality, and a case study: Italy*. Research Paper No.8706, Faculty of Social Sciences, Department of Economics, University of Ottawa, presented at the Econometric Society Meeting held in Chicago, December 28–30, 1987.
- Dancelli L. (1986). *Tendenza alla massima ed alla minima concentrazione del modello di distribuzione personale di Dagum*. Scritti in onore di Francesco Brambilla, Vol.1, Università L. Bocconi, Milano, 249–267.
- Dancelli L. (1989). *Confronti fra le curve di concentrazione  $Z(p)$  e  $L(p)$  nel modello di Dagum*. Statistica applicate, Vol.1, n.4, 1989.
- Duclos J.-Y. (2000). *Gini indices and the redistribution of income*. International Tax

- and Public Finance 7, 2, 141–162.
- Fattorini L., Lemmi A. (1979). *Proposta di un modello alternativo per l'analisi della distribuzione personale del reddito*. Atti delle giornate di lavoro AIRO, Bologna, 89-117.
- Lemmi A. (1987). *Il modello kappa a tre parametri nell'analisi della distribuzione personale del reddito*. Atti del convegno su: la distribuzione personale del reddito: problemi di formazione, di ripartizione e di misurazione, Brescia, 13-14 June 1985, a cura di M.Zenga, Vita e Pensiero, Milano, 165-189.
- Fernandez Guasti M. (1992). *Analytic Geometry of Some Rectilinear Figures*. Int. J. Educ. Sci. Technol. 23, 895-901.
- Fernández Guasti M., Meléndez Cobarrubias A., Renero Carrillo F.J., Cornejo Rodríguez A. (2005). *LCD Pixel Shape and Far-Field Diffraction Patterns*. Optik 116, 265-269. Sloane, N.J.A. Sequence A186642 in The On-Line Encyclopedia of Integer Sequences.
- Gardner M. (1977). *Piet Hein's superellipse. Mathematical Carnival. A New Round-Up of Tantalizers and Puzzles From Scientific American*, Vintage Press, New York, 240–254. <http://mathworld.wolfram.com/Superellipse.html>.
- Gielis J. (2003). *A generic geometric transformation that unifies a wide range of natural and abstract shapes*. American Journal of Botany 90, 3, 333–338.
- Gini C. (1931). *Intorno alle curve di concentrazione*. Hijos de MG Hernández.
- Gini C. (1971). *Variability and mutability, contribution to the study of statistical distributions and relations*. Studi economico-giuridici della R. Università di Cagliari (1912). Reviewed in: Light, RJ, Margolin, BH: An analysis of variance for categorical data. J. American Statistical Association 66, 534–544.
- Girone G., Nannavecchia A. (2015). *The distribution of the concentration ratio for samples from a uniform population*. Applied Mathematics 6, 01, 57.
- Girone G. (2009). *Statistica*. Cacucci Editore.
- Giurovich G. (1959). *Su un indice approssimato di asimmetria della curva di concentrazione*. Statistica. Vol. 19, pp. 463-482.
- Greenhill A.G. (1892). *The applications of elliptic functions*. London, New York, Macmillan.
- Gridgeman N.T. (1970). *Lamé ovals*. The Mathematical Gazette 54, 387, 31–37.
- Hagerbaumer J.B. (1977). *The Gini concentration ratio and the minor concentration ratio: A two parameter index of inequality*. The Review of Economics and Statistics, 377–379.
- Hancock H. (1917). *Elliptic integrals*. No. 18. John Wiley & Sons, Incorporated.
- King L.V. (2016). *On the direct numerical calculation of elliptic functions and integrals*. Cambridge University Press.
- Koshevoy G., Mosler K. (1997). *Multivariate Gini indices*. Journal of Multivariate Analysis 60, 2, 252–276.

- Lin S., Zhang L., Reddy G.V., Hui C., Gielis J., Ding Y., Shi P. (2016). *A Geometrical Model for Testing Bilateral Symmetry of Bamboo Leaf with a Simplified Gielis Equation*. Ecology and Evolution 6, 6798-6806.
- Lorenz M.O. (1905). *Methods of measuring the concentration of wealth*. Publications of the American statistical association 9, 70, 209–219.
- Loria G. (1910). *Spezielle algebraische und transcendente ebene Kurven*. Leipzig: Teubner.
- Martínez-Panero M., García-Lapresta J.L., Meneses L. C. (2016). *Multidistances and dispersion measures*. Fuzzy Logic and Information Fusion. Springer, pp. 123–134.
- Ogwang T. (2000). *A convenient method of computing the Gini index and its standard error*. Oxford Bulletin of Economics and Statistics 62, 1, 123–129.
- Palmitesta P., Provasi C., Spera C. (2000). *Confidence interval estimation for inequality indices of the Gini family*. Computational Economics 16, 1-2, 137–147.
- Rasche R. et al. (1980). *Functional forms for estimating the Lorenz curve: comment*. Econometrica 48, 4, 1061–1062.
- Shalit H. (1985). *Practitioners' corner calculating the Gini index of inequality for individual data*. Oxford Bulletin of Economics and Statistics 47, 2, 185–189.
- Shi P.J., Huang J.G., Hui C., Grissino-Mayer H.D., Tardif J.C., Zhai L.H., Wang F.S., Li B.L. (2015). *Capturing spiral radial growth of conifers using the superellipse to model tree-ring geometric shape*. Frontiers in Plant Science 6, Art. 856, 1-13, Oct. 15.
- Subramanian S. (2002). *An elementary interpretation of the Gini inequality index*. Theory and Decision 52, 4, 375–379.
- Tarsitano A. (1987). *L'interpolazione della spezzata di Lorenz*. Statistica. Anno XLVII, N.3, pp. 437-451.
- Tarsitano A. (2000). *Gli indici di asimmetria della curva di Lorenz per dati raggruppati*. In Atti della XL Riunione Scientifica della SIS. Firenze, 26-28 aprile 2000, 219-222.
- Xu K. (2003). *How has the literature on Gini's index evolved in the past 80 years?* Dalhousie University, Economics Working Paper.
- Yitzhaki S. (1983). *On an extension of the Gini inequality index*. International economic review, 617–628.
- Weisstein E.W. (2003). *Superellipse*. MathWorld, Wolfram Research.
- Weisstein E.W. (2011). *Squircle*. MathWorld, Wolfram Research.
- Zanardi G. (1965). *L'asimmetria statistica delle curve di concentrazione*. Ricerche Economiche 19, 3-4, 355–396.

# L'approccio modellistico per una programmazione più efficace dei processi amministrativi: il caso della mortalità da Covid-19

Angela Maria D'Uggento<sup>1\*</sup>, Marina Tarantini<sup>2</sup>

<sup>1</sup>Dipartimento Economia e Finanza, Università degli Studi di Bari Aldo Moro, <sup>2</sup>Comune di Bari, Ripartizione Servizi Demografici, Elettorali e Statistici

**Riassunto:** La gestione della recente esperienza della pandemia ha evidenziato la necessità di accelerare il cambiamento culturale in atto nella pubblica amministrazione e di fondare il processo decisionale sulla conoscenza e analisi dei fenomeni attraverso dati affidabili e tempestivi. Il presente lavoro intende dare un contributo in tal senso, evidenziando l'importanza di utilizzare i dati, che solitamente vengono rilevati per finalità amministrative, per supportare anche la programmazione strategica. In particolare, ricostruita in formato elettronico la serie dei decessi nel comune di Bari dal 2010 ad oggi partendo dalle denunce di morte, l'approccio classico di analisi delle serie storiche, con l'individuazione delle singole componenti, ha consentito di sviluppare un modello che potrà essere utilizzato in maniera funzionale ai processi organizzativi interni, conservando la sua validità anche cambiando la tipologia di dati in input.

**Keywords:** analisi classica serie storiche; decomposizione; pandemia Covid19; mortalità.

---

\* Autore corrispondente: [angelamaria.duggento@uniba.it](mailto:angelamaria.duggento@uniba.it). Il lavoro qui descritto è frutto di un progetto comune, ma Angela Maria D'Uggento ha provveduto alla redazione dei paragrafi 1 e 2, mentre Marina Tarantini ha redatto i paragrafi 2.1, 3 e 3.1.

## 1. Introduzione

La gestione della recente esperienza della pandemia ha evidenziato la necessità di accelerare il cambiamento culturale in atto e di fondare il processo decisionale, anche di breve periodo, sulla conoscenza della struttura interna dei fenomeni attraverso la continua lettura dei dati. Il presente lavoro intende dare un contributo in tal senso, evidenziando l'importanza di utilizzare i dati, prioritariamente raccolti a fini amministrativi, per supportare i processi decisionali di allocazione delle risorse, sempre più scarse nel settore pubblico, e di diffondere nel Paese una nuova "data governance" (Mochi Sismondi, 2021).

Assumere decisioni basate sull'evidenza dei dati è un fattore cruciale per le amministrazioni pubbliche "smart", le quali non solo devono dotarsi delle opportune tecnologie, ma anche di adeguate professionalità e puntare alla creazione di una rete tra le varie istituzioni presenti sul territorio, favorendo l'accessibilità e l'interoperabilità dei dati, nell'ottica degli Open data.

Molte amministrazioni hanno implementato la sezione Open data sul loro portale, o sono in procinto di completarla, valorizzando il proprio patrimonio di dati e favorendo le interazioni con realtà del settore privato o pubblico (*business government*), migliorando i servizi pubblici e conferendo al dato, anche integrato con quelli di altre fonti, la potenzialità di supportare le politiche programmatiche dell'ente. Tale processo condiviso implica uno sforzo congiunto ai diversi livelli di governo, dallo Stato centrale alle amministrazioni comunali, improntato alla politica della trasparenza e della collaborazione fondate sulla condivisione dei dati ad uso pubblico.

Con particolare riferimento al tema della interoperabilità dei dati, il presente contributo intende illustrare una proposta di ulteriore utilizzo dei dati rilevati dagli Uffici comunali - Ripartizione Servizi Demografici, che quotidianamente rilevano i dati sui decessi, ai sensi del DPR n.396/2000, e li condividono con ISTAT, ASL e con il sistema di Monitoraggio della mortalità giornaliera che li elabora in maniera aggregata per differenti livelli di dettaglio territoriale.

I Comuni costituiscono, dunque, il primo attore nel processo di produzione dell'informazione statistica sulla mortalità e giocano un ruolo di primo piano poiché, disponendo del dato aggiornato in tempo reale, potrebbero individuare eventuali situazioni anomale o potenziali incrementi del fenomeno osservato già nelle primissime fasi di crescita della curva di mortalità.

Partendo da tali premesse, il presente contributo si propone di illustrare uno studio multidisciplinare sulla mortalità nel Comune di Bari che, muovendo dall'analisi

dei dati sui decessi rilevabili a partire dalla dichiarazione di morte raccolte, porti alla costruzione di indicatori e/o modelli statistici robusti e sensibili che, applicati ai dati giornalieri, consentano di stimare con ragionevole anticipo fasi di crescita anomale della curva dei decessi rispetto ad andamenti attesi.

La realizzazione di tale modello troverebbe una valida applicazione nell'ambito della gestione delle risorse e dei servizi da erogare, con particolare riferimento ai servizi essenziali ed indifferibili, quali le dichiarazioni di morte, potendo disporre di uno strumento di programmazione dell'attività amministrativa; si pensi, ad esempio, alla possibilità di disporre per tempo un ampliamento del servizio di accettazione delle dichiarazioni di morte attraverso l'apertura di un maggior numero di sportelli oppure per fasce orarie più estese.

Lo studio potrebbe rivelarsi utile proposta anche su un piano politico e sanitario (adozione di compagne di sensibilizzazione, comunicazioni ai cittadini, adozione di provvedimenti restrittivi, ecc). Il contributo si articola nel modo seguente: dopo l'introduzione, nel paragrafo 2 si analizza la mortalità durante il periodo pandemico in Italia e nel comune di Bari attraverso l'approccio classico all'analisi delle serie storiche; nel paragrafo 3, per mera esigenza di completezza della trattazione, si presenta una applicazione del modello stimato per effettuare previsioni ad un anno, pur nella consapevolezza del mutato scenario epidemiologico, grazie alla massiccia diffusione dei vaccini, evidenziando tuttavia la notevole incertezza dovuta alle varianti.

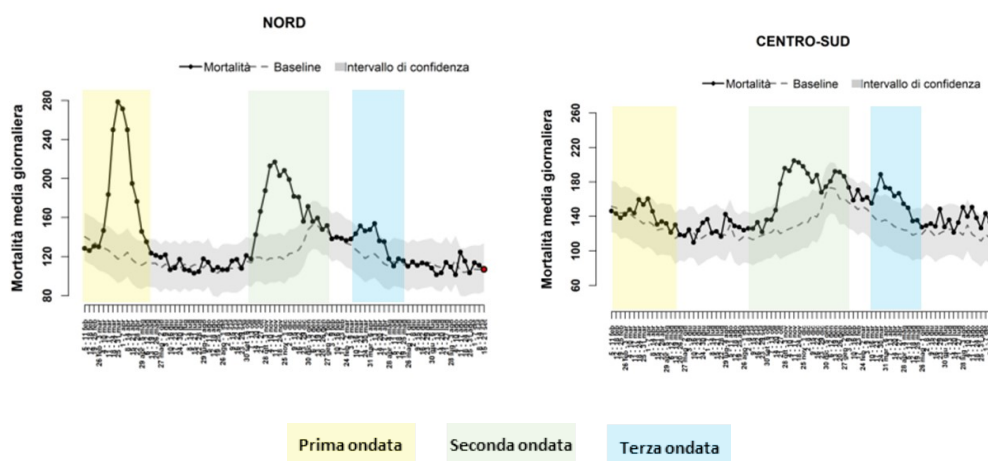
## **2. La mortalità nel periodo Covid**

Il presente studio riassume gli esiti di un'attività sistematica di osservazione dell'andamento dei decessi registrati nella città di Bari che ha avuto avvio nel corso dei primi mesi dell'anno 2020, quando iniziavano ad essere evidenti i primi effetti della pandemia da Coronavirus, e che prosegue tutt'oggi, con il fine di monitorare in tempo reale l'evoluzione del fenomeno e di predisporre opportunamente i servizi comunali connessi alla gestione del dato amministrativo.

Nel corso degli ultimi due anni, infatti, l'osservazione costante delle denunce di morte ha permesso di pianificare e attivare per tempo le misure necessarie per far fronte ai periodi di maggiore richiesta delle attività ad esse connesse che, come si vedrà innanzi, non potevano essere programmate basando l'osservazione su quanto accadeva a livello nazionale, in quanto gli effetti della pandemia da Coronavirus si sono manifestati in Italia con modalità e tempistiche differenti nei diversi territori.

L'ultimo rapporto elaborato dagli esperti del Dipartimento di Epidemiologia SSR Regione Lazio e del Ministero della Salute (Sistema di sorveglianza mortalità giornaliera, SiSMG) ha bene evidenziato la differente evoluzione del fenomeno della mortalità in Italia in relazione all'epidemia da Covid-19 (Fig.1).

**Figura 1.** *Mortalità da Covid-19 in Italia per macroarea e periodo*



Dai grafici elaborati dai ricercatori del SiSMG emerge chiaramente il diverso impatto che ha caratterizzato a livello geografico la recente e ancora attuale emergenza sanitaria: la prima ondata di contagi - che si fa partire dalla scoperta del primo caso positivo del Paese e arriva fino a maggio 2020 - ha avuto delle conseguenze in termini di decessi molto più gravi per le regioni settentrionali rispetto agli effetti rilevati nel Sud Italia. Si rammenta al riguardo che le cronache del periodo riportavano quotidianamente le difficoltà delle regioni del Nord nella gestione a livello sanitario ed ospedaliero dovute proprio ad una velocità di contagio che non aveva eguali in Italia negli anni passati.

Successivamente all'attuazione delle misure restrittive introdotte dal primo lockdown ed al conseguente temporaneo rientro dell'emergenza, si è assistito, dapprima, ad una riduzione, nell'autunno del 2020, alla ripresa del fenomeno dei contagi e, con essa, dell'incremento di mortalità: in questa seconda ondata, a differenza di quanto accaduto nell'inverno precedente, le regioni più colpite sono state quelle del Centro-Sud che hanno dovuto far fronte ad un forte inasprimento della mortalità giornaliera che è proseguito, quasi senza soluzione di continuità, anche per tutta la terza ondata della primavera 2021. Al Nord invece, dopo il picco di

mortalità della seconda ondata, le campagne di vaccinazione unitamente alle misure restrittive legate all'istituzioni delle zone rosse in diverse regioni, hanno permesso di contenere gli effetti dell'ultima ondata di contagi (Fig. 2).

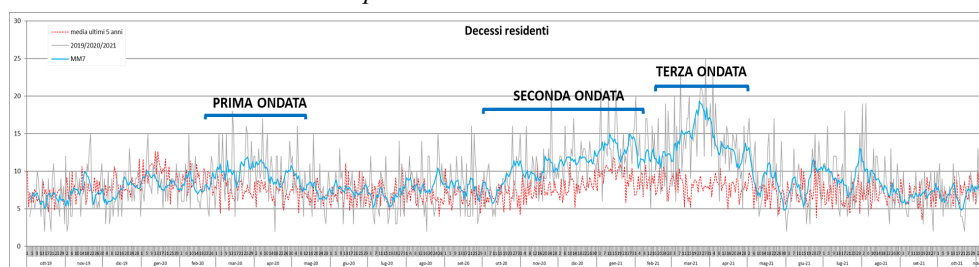
Come si illustrerà più approfonditamente nel paragrafo 2.1, nel medesimo periodo, l'andamento della mortalità nel Comune di Bari ha seguito quello delle regioni del Centro-Sud, facendo emergere, tuttavia, delle importanti differenze in termini di impatto delle tre ondate sulla curva dei decessi.

### 2.1. La mortalità nel Comune di Bari

Il presente paragrafo analizza la mortalità giornaliera nel Comune di Bari, come desunta dalle denunce di decesso pervenute giornalmente agli uffici preposti e, in coerenza con la metodologia utilizzata dal SiSMG, è riferita alle sole dichiarazioni relative ai residenti.

In particolare, dopo aver delineato la situazione da un punto di vista generale, prendendo in considerazione l'intero periodo dei due anni interessati dalla emergenza epidemiologica, si esamineranno le differenze che hanno caratterizzato le tre ondate di contagi registrate fino ad oggi in termini di intensità, durata e popolazione colpita. La Fig.2 illustra l'andamento dei decessi giornalieri relativi dei cittadini baresi residenti nel periodo ottobre 2019 – ottobre 2021 (linea grigia) e consente una diretta comparazione con il livello “atteso” di riferimento rappresentato dalla mortalità media giornaliera riferita ai 5 anni immediatamente precedenti (linea rossa). Per una più agevole lettura del fenomeno, si riporta nel grafico anche la media mobile a 7 giorni dei decessi giornalieri registrati nel periodo ottobre 2019 – ottobre 2021 (linea azzurra).

**Figura 2.** Andamento effettivo dei decessi giornalieri, atteso e media mobile 7 giorni dei residenti nel Comune di Bari nel periodo ottobre 2019–ottobre 2021



Già da una prima analisi dei dati grezzi si può notare come la curva che descrive l'evoluzione della mortalità degli ultimi 2 anni segua molto da vicino l'andamento

atteso fino al mese di febbraio 2020, periodo in cui inizia a registrarsi un allontanamento dai livelli di riferimento per poi assestarsi sistematicamente al di sopra della curva attesa per tutto il mese di marzo e, in modo più attenuato, per tutto aprile 2020; successivamente, grazie agli effetti delle misure restrittive attuate a livello nazionale con il primo lockdown, la curva osservata sembra ritornare ai livelli attesi e ad assestarsi su valori normali per tutto il periodo estivo (ad eccezione di una breve risalita in corrispondenza della seconda ondata di calore relativa al mese di agosto 2020). Con la fine dell'estate 2020 diviene evidente la brusca ripresa del fenomeno: a partire dai primi giorni di ottobre 2020, infatti, si verifica una vera e propria esplosione dei decessi caratterizzata da una durata notevolmente superiore rispetto a quanto osservato durante la prima ondata di contagi dell'inverno 2020 e con picchi giornalieri elevatissimi (si osservi al riguardo che la curva azzurra resta sistematicamente al di sopra del livello atteso per oltre sei mesi, fino a raggiungere un valore massimo in corrispondenza del giorno 6 marzo 2021 in cui si sono osservati 28 decessi di residenti).

I dati appena analizzati evidenziano, dunque, come l'eccesso di mortalità nel Comune di Bari si sia verificato nel corso della prima fase della pandemia con una modalità, seppur consistente, più contenuta rispetto a quanto accaduto in corrispondenza della seconda (ottobre 2020–gennaio 2021) e soprattutto della terza ondata di contagi (marzo 2021–maggio 2021) che si sono manifestate in misura notevolmente più intensa e durevole.

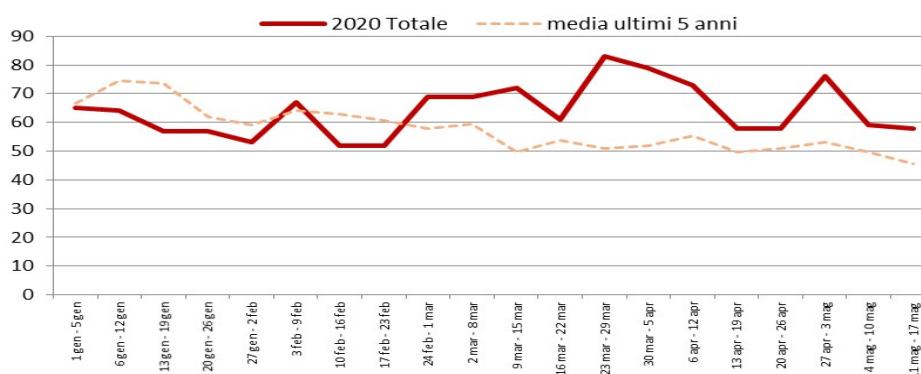
La Tab. 1 in appendice riporta i dati relativi ai decessi registrati nel periodo centrale di ciascuna delle tre ondate epidemiologiche. La scelta di rappresentare i dati con cadenza settimanale è dettata dall'esigenza di limitare gli andamenti accidentali dei dati giornalieri. Emerge chiaramente come in tutti e tre i periodi analizzati la mortalità settimanale sia significativamente superiore rispetto ai valori di riferimento rappresentati dalla media dei decessi settimanali rilevati nei 5 anni immediatamente precedenti e, soprattutto, come tale incremento sia tanto più consistente nella seconda e terza ondata, in cui la variazione percentuale media della mortalità settimanale si è attestata sul 41% e sul 65% rispettivamente. Si evidenzia, in particolare, che nel momento più grave della terza ondata i decessi settimanali hanno più che raddoppiato i livelli normali generalmente osservati per il periodo.

#### *Prima ondata*

L'evoluzione del fenomeno nel corso della prima fase dell'emergenza sanitaria da Coronavirus è analizzata distinguendo i differenti effetti sulle varie fasce d'età della popolazione e confrontando i decessi settimanali nel periodo marzo-maggio

2020 e la media dei decessi settimanali registrati nel medesimo periodo dei 5 anni precedenti. È evidente come la curva della mortalità settimanale osservata superi la curva attesa a partire dall'ultima settimana di febbraio 2020 per poi rimanere sistematicamente al di sopra di essa per tutto il primo periodo dell'emergenza sanitaria. Si precisa, al riguardo, che il primo caso accertato di positività al Covid-19 in Puglia è stato riscontrato il 26 febbraio 2020, pertanto è verosimile ipotizzare che il virus abbia iniziato a circolare già dalle settimane precedenti e che in corrispondenza dell'ultima settimana di febbraio ne siano divenuti evidenti gli effetti.

**Figura 3.** Andamento dei decessi giornalieri e dato medio negli ultimi 5 anni dei residenti nel Comune di Bari nella prima ondata della pandemia.



Per il comune di Bari, il picco massimo è stato raggiunto nella settimana 23-29 marzo 20 con 83 decessi di residenti, ben il 62,7% in più del valore medio osservato per il medesimo periodo nell'ultimo quinquennio. Come accaduto nel resto d'Italia, l'eccesso di mortalità osservato nella prima ondata di contagi ha colpito in maniera differente le diverse fasce di popolazione, come evidenziato in Fig. 4.

Per una più agevole interpretazione dei grafici per età, nella Tab.2 in appendice sono riportati i valori osservati ed i corrispondenti valori attesi nel primo periodo della pandemia. I dati mostrano come nel periodo in esame l'incremento di mortalità più elevato in termini assoluti sia ascrivibile essenzialmente alle fasce di età 65-84 e 85+, anche se i relativi grafici mostrano diversi andamenti nella prima ondata di contagi per le due compagini:

- con riferimento alla classe dei grandi anziani (85+) si sono registrati un primo picco nel periodo 24 febbraio – 8 marzo e, successivamente, un secondo punto di massimo nelle settimane comprese tra il 23 marzo ed il 19 aprile; a partire da tale data si è osservato un graduale ritorno a livelli normalmente attesi per il periodo;
- per la fascia d'età 65-84 anni, lo scostamento dalla curva dei valori medi si è

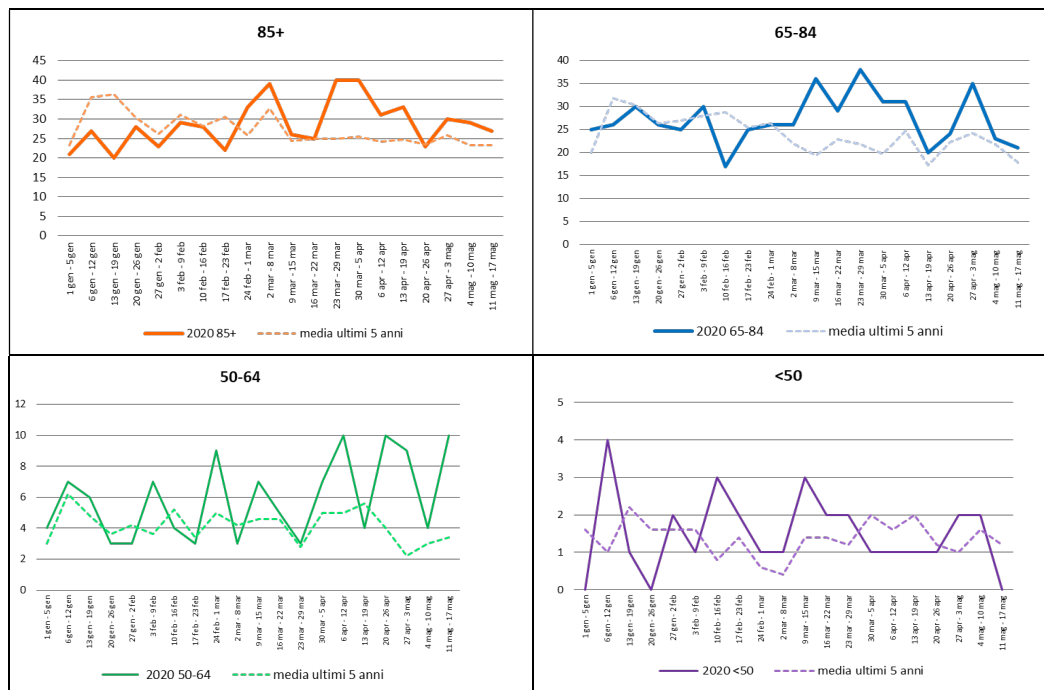
manifestato in misura maggiore e con una frequenza superiore rispetto a quanto accaduto per la classe 85+: per il contingente degli anziani si osservava, infatti, una prima fase di picchi di mortalità nel periodo febbraio-marzo e un terzo punto di massimo nella seconda metà di aprile; è plausibile ritenere, a tal proposito, che la fascia d'età in esame abbia subito questo peggioramento del numero dei decessi, che invece non si è osservato per i grandi anziani, poiché, inizialmente, si riteneva che la malattia avesse esiti fatali solo su individui fragili o in presenza di patologia pregresse, situazioni tipiche delle età più avanzate, e pertanto la società e le istituzioni hanno concentrato l'attenzione e le misure di prevenzione prevalentemente a favore di tale categoria di soggetti, sottovalutando, di fatto, gli effetti che il virus avrebbe poi avuto sul resto della popolazione. A conferma di ciò si osservi quanto registrato in termini di decessi settimanali per i residenti appartenenti alla fascia di età 50-64: questo gruppo è invero quello per il quale si è assistito al maggiore incremento di mortalità in termini relativi: con 32 decessi in più rispetto a quelli normalmente osservati nello stesso lasso di tempo, per il gruppo in esame si calcola un incremento del 64% rispetto ai valori attesi, con maggior concentrazione dei decessi proprio nella seconda metà della prima ondata.

Completa l'analisi l'osservazione dei decessi nel restante contingente di popolazione, quella con meno di 50 anni, che, almeno nella prima fase della pandemia non sembra aver risentito degli effetti.

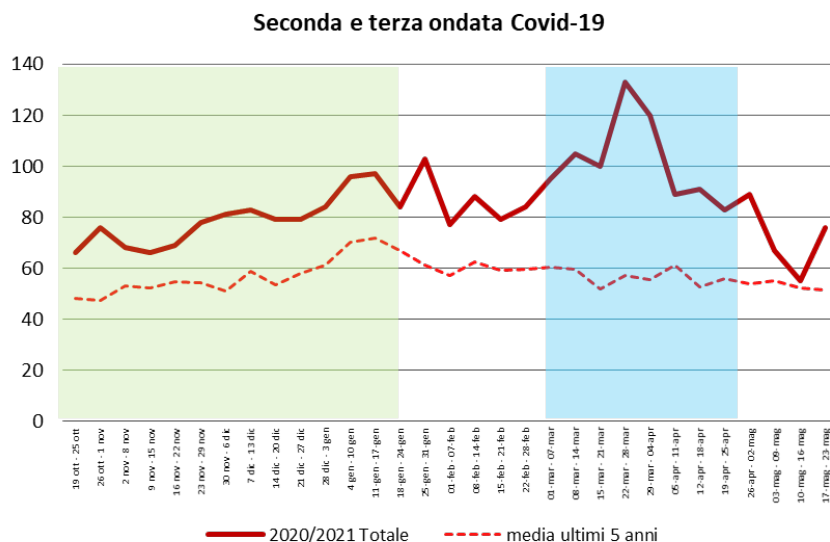
#### *Seconda e terza ondata*

Come appreso dalle statistiche diffuse nel corso di questi mesi, anche la seconda e la terza ondata della pandemia sono state particolarmente intense e hanno colpito, seppure con differenti modalità, tutte le diverse fasce d'età della popolazione. Considerando il fenomeno nel suo complesso, dal grafico che segue si può comprendere molto chiaramente quale sia stata la portata delle nuove ondate registrate tra la fine del 2020 e la prima metà del 2021 per il Comune di Bari: il numero di decessi settimanali di cittadini residenti è stato in media il 50% in più del valore atteso per un periodo di oltre 8 mesi, con punte superiori al 100% in corrispondenza dell'ultima violenta ondata di contagi (marzo-aprile 2021).

**Figura 4.** Andamento dei decessi giornalieri e dato medio negli ultimi 5 anni dei residenti nel Comune di Bari nella prima ondata della pandemia per fasce d'età.



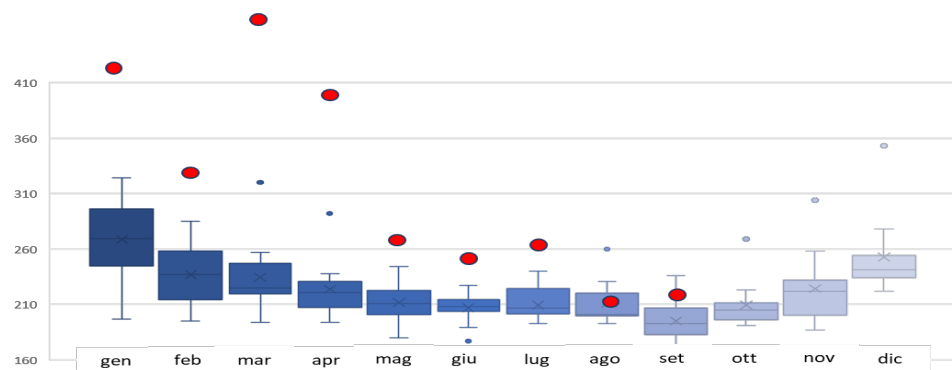
**Figura 5.** Andamento dei decessi e dato medio negli ultimi 5 anni dei residenti nel Comune di Bari nella seconda e terza ondata della pandemia.



È proprio nella primavera 2021 che, difatti, nel Comune di Bari si sono registrati i livelli più elevati di mortalità: si pensi, ad esempio, che agli atti degli uffici demografici risultano ben 489 dichiarazioni di morte relative al solo mese di marzo 2021, valore che risulta essere oltre il doppio dei normali livelli osservati nell'ultimo quinquennio e, in assoluto, il più elevato nelle statistiche disponibili sugli ultimi 10 anni di rilevazione dati da parte degli sportelli dell'ufficio di stato civile. È interessante rilevare, a tal proposito, il consistente scostamento delle osservazioni dei decessi mensili relativi all'anno 2021 rispetto ai valori di riferimento, allargando momentaneamente il campo di osservazione ai dieci anni precedenti.

Nella Fig.6 è presentato il confronto tra il dato dei decessi osservato per ogni mese nell'anno 2021 (indicato dai pallini rossi) e il complesso delle osservazioni disponibili per il medesimo mese negli ultimi 10 anni, rappresentato dal relativo boxplot. Appare confermata la straordinarietà dei livelli di mortalità registrati nel corso dei primi mesi dell'anno 2021 rispetto a quanto accaduto nell'ultimo decennio: i dati mensili del 2021 superano di gran lunga anche le osservazioni outliers che, tra l'altro, sono proprio riferite alla prima (marzo-aprile 2020) ed alla seconda fase della pandemia (ottobre-dicembre 2020), a riprova della maggiore gravità che ha contraddistinto le ondate più recenti.

**Figura 6.** *Decessi nel 2021 e decessi complessivi negli ultimi dieci anni per mese dei residenti nel Comune di Bari.*



Riprendendo l'analisi del numero settimanale di decessi nella seconda e terza ondata, si descrive di seguito il diverso impatto rilevato per le varie fasce d'età (Fig.7).

Dai relativi grafici si evince agevolmente che:

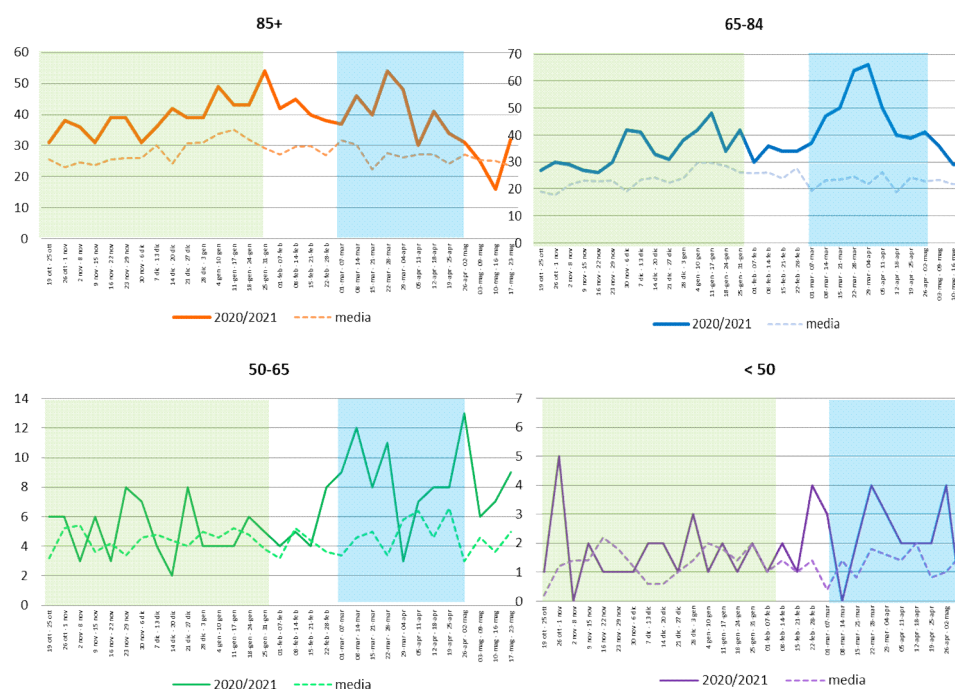
- la seconda ondata ha avuto esiti negativi soprattutto per le fasce d'età più elevate (85+ e 65-84), mentre le classi di età più giovani hanno registrato

incrementi di mortalità molto contenuti (solo 9 casi in più per la classe 50-64 e 9 casi in più per la classe <50);

- la terza ondata, per contro, ha generato un aspro incremento dei livelli di mortalità settimanale per tutte le fasce d'età tranne che per quella dei grandi anziani per i quali, seppur rimanendo su livelli superiori a quelli medi, non si è assistito all'esplosione che invece ha caratterizzato i gruppi <50, 50-64 e 65-84; è verosimile ritenere che la campagna vaccinale, che nelle prime fasi ha coinvolto le fasce d'età più anziane della popolazione, abbia agito con efficacia quale fattore di protezione, mentre le classi di età più giovane sono state più esposte alla nuova ondata di epidemia e al contestuale diffondersi della nuova variante del virus;

- per la prima volta nel periodo analizzato, il gruppo degli individui con età inferiore a 50 anni fa registrare incrementi significativi di mortalità, con un numero di decessi doppio rispetto a quelli registrati in media nelle stesse settimane degli ultimi cinque anni. I dati di dettaglio sono riportati nella Tab. 3 in appendice.

**Figura 7.** Andamento dei decessi giornalieri e dato medio negli ultimi 5 anni dei residenti nel Comune di Bari nella seconda e terza ondata della pandemia per fasce d'età.



### **3. Una proposta di utilizzo dei dati di fonte amministrativa per lo studio dei fenomeni complessi.**

Da oltre 10 anni l'Istituto Nazionale di Statistica e l'Unione Statistica Comuni Italiani promuovono l'impiego degli archivi amministrativi disponibili e dei dati prodotti dagli enti locali per conoscere il territorio e per pianificare le politiche locali. Nel corso di numerosi dibattiti sul tema è, infatti, emersa unanime la convinzione che *“l'amministrazione moderna dei poteri locali deve supportare le sue funzioni con apparati conoscitivi e informativi, disponibili per i decisori e per i cittadini, calibrati sulle dimensioni territoriali di competenza, con caratteristiche di tempestività, pertinenza, fruibilità, robustezza tecnica e ampiezza tematica”* (Innocenti, 2016). In ambito internazionale è riscontrabile già dai primi anni 2000 la tendenza verso l'acquisizione centralizzata di archivi amministrativi, orientata alla costruzione di un sistema di archivi integrati o integrabili, che risponda ai fabbisogni interni. In alcuni casi i dati amministrativi sono utilizzati per sostituire in parte o completamente le unità da rilevare mediante indagine diretta, ipotizzando una fase assimilabile a quella di raccolta/acquisizione, durante la quale i concetti sottostanti all'archivio amministrativo diventano dati osservati e interpretati in termini di variabili statistiche (ISTAT, 2016)

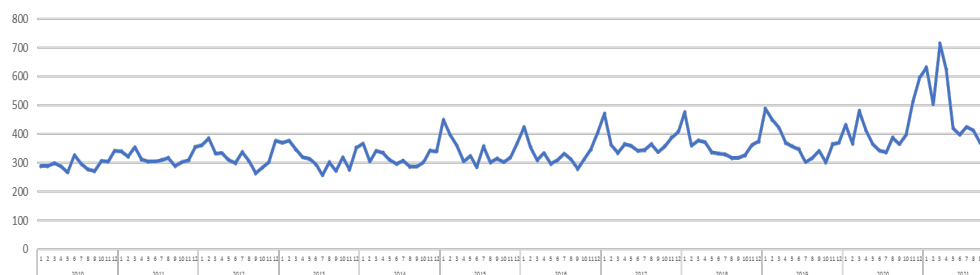
Con la presente analisi, partendo dai dati amministrativi desumibili dai registri degli atti di morte e contenenti tutte le informazioni acquisite dagli ufficiali di stato civile in sede di accettazione delle dichiarazioni di morte avvenute dal 2010 ad oggi nel territorio comunale barese, si è inteso ricostruire la relativa serie storica dei decessi per poi analizzare i dati secondo l'approccio classico di analisi delle serie storiche basato su metodi di scomposizione. Il fine ultimo è quello di individuare eventuali pattern sistematici da proiettare in periodi successivi al fine di produrre una previsione.

Il modello di stima del numero di decessi mensili ottenuto risulterebbe uno strumento particolarmente prezioso nell'ottica della programmazione delle attività degli uffici e dell'organizzazione complessiva dei flussi informativi tra enti, supportando il processo decisionale nel modo più efficiente.

In coerenza con l'obiettivo, in questa specifica analisi verrà considerato il complesso delle dichiarazioni di morte pervenute mensilmente agli uffici comunali, e non soltanto quelle relative ai cittadini residenti come in precedenza, dal 2010 al 2021 (Tab. 4 in appendice). Partendo dall'osservazione del *time plot* dei dati relativi ai decessi mensili (Fig. 8), è agevole individuare delle regolarità di comportamento delle osservazioni: il grafico evidenzia infatti un lieve trend

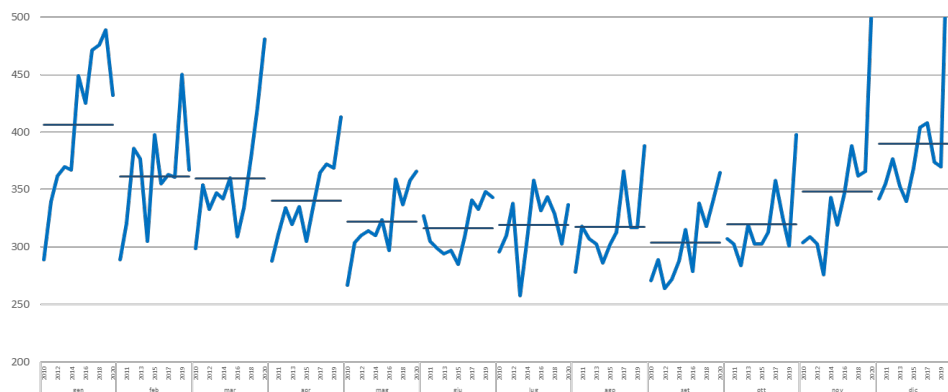
crescente nel lungo periodo nonché la presenza del fenomeno della stagionalità riscontrabile nell'avvicinarsi di picchi e contrazioni della curva in corrispondenza dei medesimi periodi.

**Figura 8.** *Decessi nel Comune di Bari dal 2010 al 2021*



Per ciascun anno di osservazione è riscontrabile, infatti, un picco più alto in corrispondenza dei mesi di gennaio e febbraio, generalmente riconducibile all'incremento di mortalità connesso all'influenza stagionale, e un punto di massimo più contenuto nel periodo luglio-agosto, legato all'aumento dei decessi causato dalle ondate di calore. A definitiva conferma della presenza di un'importante componente dovuta alla stagionalità si riporta il *seasonal plot* della serie in esame (Figura 9), che evidenzia dei picchi di mortalità nel periodo invernale e, seppur molto più contenute, nel corso dei mesi estivi.

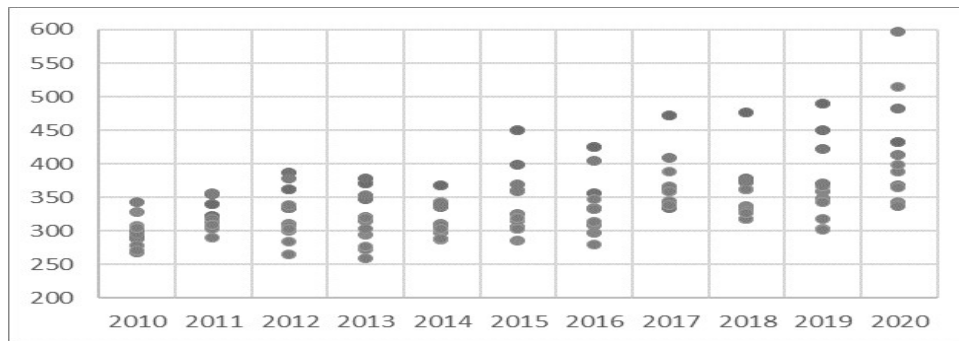
**Figura 9.** *Stagionalità dei decessi nel Comune di Bari dal 2010 al 2020*



Analogamente all'individuazione della componente stagionale, è possibile identificare la presenza di un trend delle osservazioni che, nel caso di specie, evidenzia un andamento di fondo tendenzialmente in aumento, che è imputabile

essenzialmente al graduale invecchiamento della popolazione e, solo per l'anno 2020, all'eccesso di mortalità indotto dalla pandemia (Fig.10).

**Figura 10** Decessi nel Comune di Bari nei 12 mesi dell'anno dal 2010 al 2020

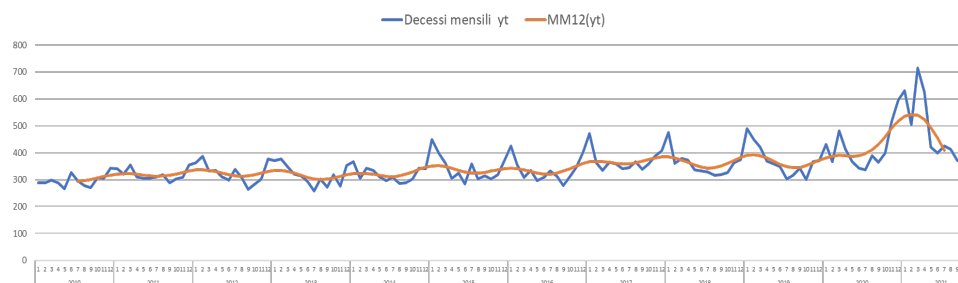


Ipotizzando, secondo il classico approccio di analisi delle serie temporali, la presenza di un modello che sottende ai dati della nostra serie storica, ciascuna osservazione verrà considerata come il risultato dell'azione di elementi sistematici o deterministici da un lato (quali appunto il trend e la stagionalità) e di oscillazioni di disturbo o di natura casuale:  $y_t = f(T_t, S_t, e_t)$ , dove  $y_t$  è il dato riferito al periodo  $t$ ,  $T_t$  rappresenta il ciclo-trend, ovvero la variabile che rappresenta l'andamento di fondo e che tiene conto di eventuali oscillazioni congiunturali,  $S_t$  è la componente stagionale mentre  $e_t$  costituisce la componente accidentale o di disturbo al tempo  $t$ .

Si procederà, dunque, a stimare le componenti costitutive della serie temporale oggetto di studio che presentano regolarità di comportamento e che, pertanto, si ritiene possano continuare a verificarsi nel tempo al netto dell'azione di elementi accidentali o di disturbo. Nel caso in esame si sceglie di considerare le osservazioni temporali come il risultato di un *modello moltiplicativo* poiché l'ampiezza dell'oscillazione stagionale varia con il livello della serie, per cui il modello diventa:  $y_t = T_t \cdot S_t \cdot e_t$ .

La stima delle singole componenti verrà effettuata mediante l'impiego di metodi perequativi e, in particolare, una prima stima iniziale del ciclo-trend è stata realizzata mediante il calcolo di medie mobili centrate a 12 termini,  $MM12(y_t)$ , espediente che consente di eliminare le oscillazioni stagionali, e gran parte di quella erratica. Si precisa che nella presente analisi si è scelto di utilizzare medie mobili ponderate applicando un sistema di pesi che permettesse di dare maggior risalto al mese su cui centrare la media mobile e pesi via via decrescenti ai mesi precedenti e successivi del medesimo anno mobile.

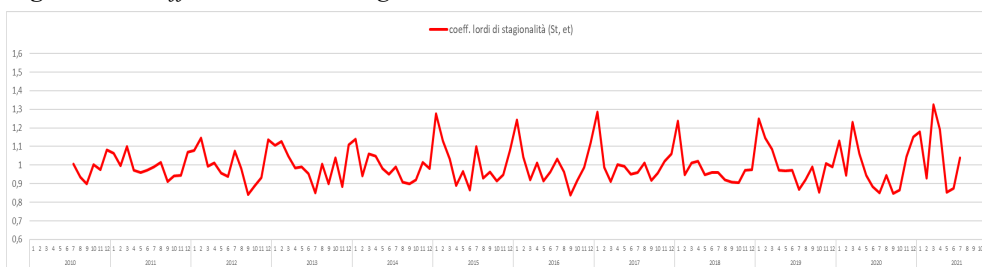
**Figura 11.** Decessi mensili nel Comune di Bari e medie mobili a 12 termini dal 2010 al 2021



La curva delle medie mobili centrate (linea arancione), pur seguendo da vicino “la traccia” della curva originaria, presenta un andamento più liscio, per via dell’effetto smussamento introdotto dal metodo perequativo e legato all’annullamento delle oscillazioni erratiche e stagionali.

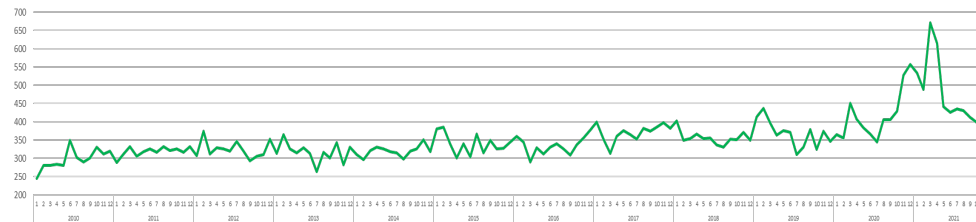
Depurando, per rapporto, la serie originaria dalla componente ciclo-trend appena calcolata si ottiene una stima della serie della stagionalità mista ad errore ( $S_t, e_t$ ) costituita dai cosiddetti coefficienti lordi di stagionalità.

**Figura 12.** Coefficienti lordi di stagionalità dal 2010 al 2021

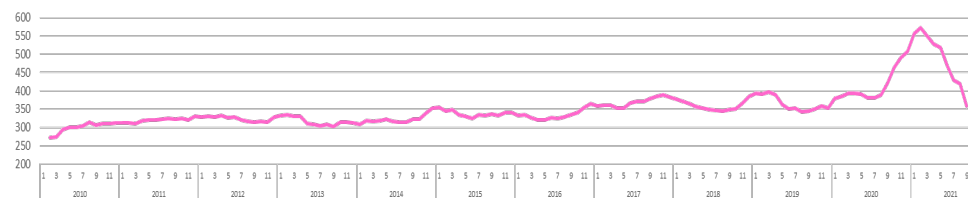


Per giungere all’individuazione della componente stagionale sarà sufficiente eliminare la componente accidentale alla serie dei valori ( $S_t, e_t$ ) mediante il calcolo dei 12 coefficienti netti di stagionalità ( $\bar{S}_m$ ) ciascuno ottenuto come media di tutti i coefficienti lordi riferiti al medesimo mese. Questo approccio si basa sull’ipotesi di stagionalità costante, ovvero sull’assunto che  $S_t = S_{t+12} = S_{t+24} = \dots$ , ipotesi confermata dal fatto che la media di tutti i coefficienti  $\bar{S}_m$  è uguale ad 1 e che implica che effettivamente le oscillazioni stagionali esauriscono il loro effetto all’interno dell’anno (Tab.5 in appendice).

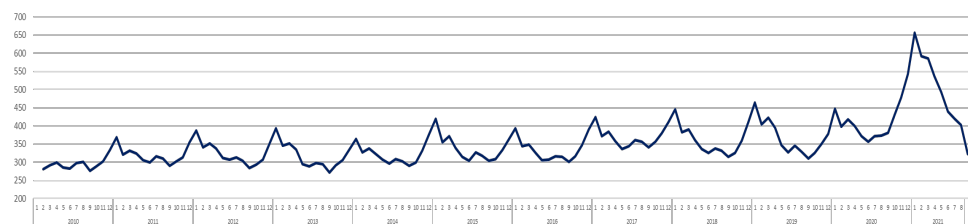
La definizione dei predetti coefficienti consente il calcolo della serie destagionalizzata  $D_t$  che contiene il pattern del ciclo-trend nonché l’effetto di disturbo, come in Fig.13.

**Figura 13.** *Serie destagionalizzata dei decessi dal 2010 al 2021*

Infine, la Fig. 14 illustra l'andamento del pattern sistematico del ciclo trend  $T_t$  ottenuto applicando il calcolo di medie mobili a 5 termini sulla serie dei dati stagionalizzati.

**Figura 14.** *Componente ciclo-trend dei decessi dal 2010 al 2021*

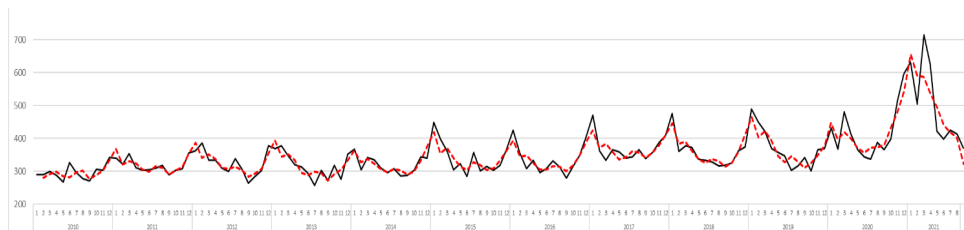
Combinando il ciclo-trend  $T_t$  e la componente stagionale  $S_t$  appena determinate attraverso l'applicazione di un modello di tipo moltiplicativo, è possibile pervenire alla determinazione della componente sistematica  $Y'_t$  della serie temporale ovvero all'andamento teorico che avrebbe avuto il fenomeno se non avessero agito cause accidentali o elementi di disturbo.

**Figura 15.** *Decessi "teorici" dal 2010 al 2021*

Il modello ottenuto sembra presentare un buon livello di adattabilità ai dati osservati, i quali evidenziano un lieve aumento tendenziale del livello del fenomeno del lungo periodo, verosimilmente legato alle dinamiche connesse al progressivo invecchiamento della popolazione, e un graduale incremento delle oscillazioni stagionali, poiché i picchi ricorrenti tendono ad assestarsi su livelli

sempre maggiori, generando un incremento della variabilità osservata infrannuale. È possibile valutare la bontà di adattamento del modello ai dati da un punto di vista grafico, confrontando le due curve appena illustrate (Figura 16), analizzando i residui, o, ancora, le principali misure di fitting.

**Figura 16.** Decessi osservati e “teorici” dal 2010 al 2021

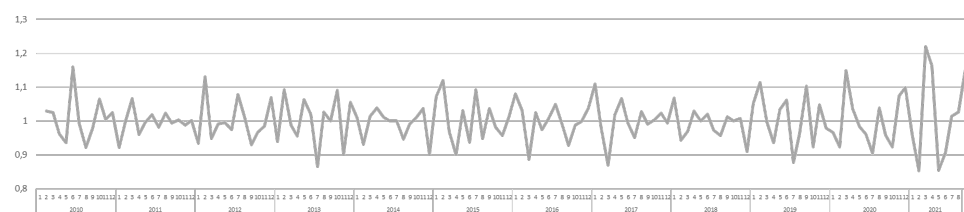


Attestando la validità della scomposizione della serie storica, il grafico dei residui in funzione del tempo non evidenzia andamenti sistematici, oscillando casualmente attorno al valore neutro, pari a 1 per i modelli moltiplicativi e a 0 per quelli additivi. Degna di nota è la maggior variabilità nell’ultima parte del grafico, corrispondente agli anni 2020-2021, che sappiamo, tuttavia essere, imputabile agli effetti della pandemia.

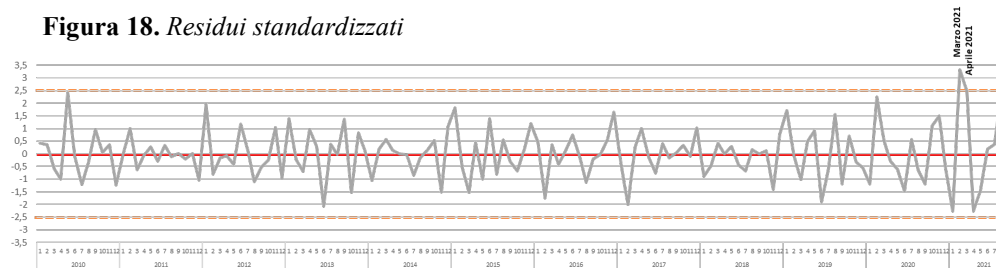
È possibile valutare a questo punto, per ulteriore conferma, l’ipotesi di normalità dei residui, ovvero l’ipotesi che la distribuzione dei residui segua un andamento normale con effetto medio nullo.

Dopo opportuna standardizzazione, si riporta il time plot dei residui standardizzati dal cui esame è possibile individuare eventuali valori anomali.

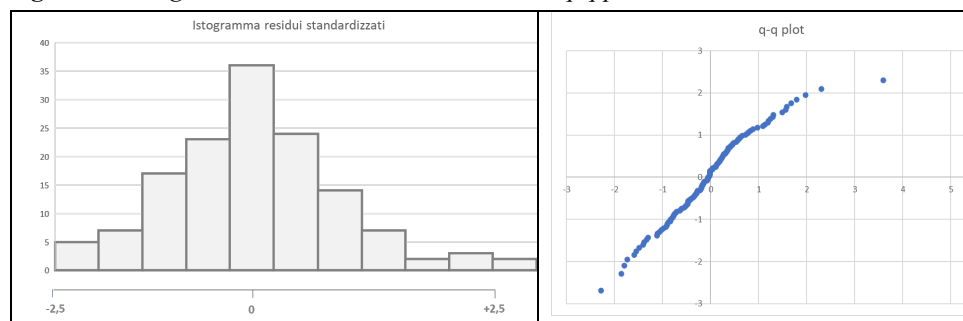
**Figura 17.** Residui



Dalla Fig. 18 emerge che solo due osservazioni (marzo e aprile 2021) risultano essere anomale poiché esterne alla banda di confidenza del 99% (compresa tra  $-2,5$  e  $+2,5$ ). Come già evidenziato, questi due *outliers* sono determinati dagli effetti della terza ondata che, per il comune di Bari, ha rappresentato il momento più critico dall’inizio dell’emergenza sanitaria.

**Figura 18.** *Residui standardizzati*

È appena il caso di aggiungere che, nella terza ondata, la Direzione della Ripartizione Servizi Demografici, in ragione dell'evidente aumento delle denunce emerso dal monitoraggio dei decessi e della conseguente attività di formazione di atti di morte, ha potenziato il servizio incrementando le unità di personale al servizio di accettazione delle denunce e disponendo l'apertura di nuovi sportelli dedicati. La rapida adozione del nuovo assetto organizzativo ha consentito di rispondere prontamente alle esigenze contingenti, garantendo elevati livelli di servizio a fronte di una situazione del tutto inattesa. Ad ulteriore conferma dell'andamento normale della componente erratica, si riportano l'istogramma ed il q-q plot della serie temporale dei residui standardizzati dai quali è possibile evincere l'effettiva normalità dei residui calcolati a partire dal modello stimato e, di conseguenza, la bontà della scomposizione.

**Figura 19.** *Istogramma dei residui standardizzati e q-q plot*

### 3.1 L'impiego del modello per previsioni future

Ai soli fini di completezza dell'analisi e per dimostrare la validità del modello appena implementato per effettuare previsioni future, si ipotizza di voler prevedere l'andamento del fenomeno nell'anno  $t+1$ , pur consapevoli che, nel caso di specie, il fenomeno dell'eccesso di mortalità non è destinato a ripetersi con le medesime proporzioni perché l'adozione di una efficace politica sanitaria nazionale, e in

particolare la diffusione dei vaccini, ha consentito di fronteggiare la diffusione della pandemia. Come negli approcci più comuni in letteratura, si procederà alla previsione della singola componente (trend-ciclo, stagionalità) per poi ricomporre il dato futuro della serie come prodotto dei valori delle singole componenti, in virtù dell'ipotesi di un modello moltiplicativo sotteso alla serie temporale.

Tuttavia, poiché generalmente il *trend-ciclo* risulta essere la componente più difficile da prevedere, soprattutto a causa dell'oscillazione ciclica che non ha andamenti regolari periodici, e poiché nel caso di studio di può legittimamente ritenere che il ciclo abbia intensità trascurabile rispetto all'ordine di grandezza del *trend*, si procederà a stimare il *trend* mediante una funzione analitica del tempo  $f(t)$ .

Indicata con  $n$  la lunghezza della serie temporale, ovvero il numero di osservazioni disponibili, la previsione per il tempo futuro  $t=n+1$  viene ottenuta estrapolando la funzione, per cui  $\hat{T}_{n+1}=f(n+1)$

Per quanto riguarda la componente stagionale, avendo assunto l'ipotesi di stagionalità costante negli anni, l'effetto stimato della stagionalità del mese  $m$  ( $m=1, \dots, 12$ ) sarà data dal coefficiente di stagionalità di cui alla Tab. 5.

Infine, per la componente erratica, viene adottato, come previsione, il valore neutrale, che è pari a 1 nel caso di modello moltiplicativo.

La previsione  $F_{n+1}$  per il fenomeno in esame, si otterrà, dunque, ricomponendo le previsioni per le tre componenti:  $F_{n+1} = \hat{T}_{n+1} \cdot \hat{S}_m$ .

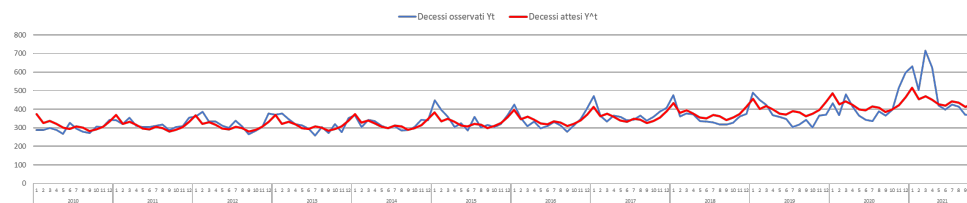
Nel caso in esame, si è proceduto a stimare il trend a partire dalla serie destagionalizzata che rappresenta la base di dati più idonea su cui basare l'individuazione della funzione analitica che descrive la componente trend, in quanto priva appunto dell'andamento stagionale. In questa serie, infatti, e in ipotesi di assenza di significative oscillazioni cicliche, l'unica componente sistematica presente nei dati è quella tendenziale di lungo periodo.

A partire dallo studio empirico della serie destagionalizzata, è stato stabilito di stimare la componente trend attraverso una forma quadratica del tipo  $T_t = \beta_0 + \beta_1 t + \beta_2 t^2$  da cui, applicando il metodo dei minimi quadrati, si è giunti alla seguente funzione analitica:  $\hat{T}_t = 317,74 - 0,5507 t + 0,0109 t^2$ , cui corrisponde un valore dell'indice di determinazione lineare  $R^2$  pari a 0,55.

Combinando la stima della componente trend così ottenuta con quella della componente stagionale, in ipotesi di effetto medio nullo della componente erratica, il modello di stima prescelto ( $\hat{Y}_t = \hat{T}_t \cdot \hat{S}_t$ ) sembra descrivere in maniera abbastanza fedele l'andamento dei decessi registrati nel comune di Bari, ad eccezione dell'ultima parte della serie, ovvero in corrispondenza del primo semestre 2021, in cui, come noto, hanno agito variabili esogene non prevedibili (seconda e terza

ondata di contagi da Coronavirus unitamente al tardivo avvio della campagna vaccinale, inefficacia delle misure di contenimento, ecc.) deviando la curva rispetto al suo andamento atteso.

**Figura 20.** *Decessi osservati e “teorici” dal 2010 al 2021*



Nonostante il non ottimale adattamento del modello nel periodo specificato, le misure di fitting sembrano confermare la discreta bontà del modello stimato nel descrivere il fenomeno oggetto di studio.

Indicando con  $r_t$  il residuo  $Y_t - \hat{Y}_t$ , si ottengono i seguenti valori per gli indici della bontà di adattamento:

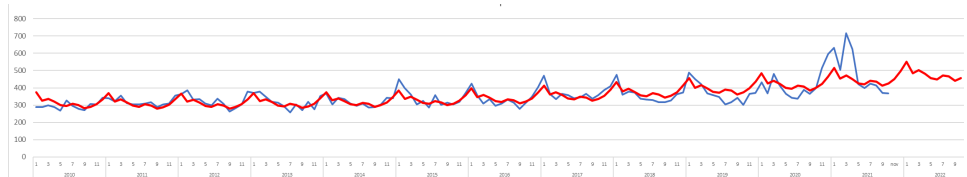
$$ME = \frac{1}{n} \sum r_t = 0,4 \quad MAPE = \frac{100}{n} \sum \frac{|r_t|}{y_t} = 7\%$$

Ipotizzando, in via del tutto teorica, che il modello stimato sui dati passati rimanga valido anche nel futuro, è possibile utilizzare la forma analitica ottenuta per proporre delle previsioni per i periodi futuri. Tuttavia, anche in considerazione della eccezionalità dell'evento pandemico recente e data la finalità meramente stilistica dell'applicazione, si ritiene opportuno limitare l'orizzonte temporale al breve periodo, pari a 12 mesi. Il risultato è illustrato nella tabella seguente:

**Tabella 1.** *Decessi mensili stimati con il modello*

| Periodo       | Stima decessi mensili | Periodo   | Stima decessi mensili |
|---------------|-----------------------|-----------|-----------------------|
| Novembre 2021 | 450                   | Maggio    | 455                   |
| Dicembre      | 496                   | Giugno    | 449                   |
| Gennaio 2022  | 551                   | Luglio    | 472                   |
| Febbraio      | 484                   | Agosto    | 465                   |
| Marzo         | 503                   | Settembre | 440                   |
| Aprile        | 482                   | Ottobre   | 456                   |

**Figura 21.** *Decessi osservati dal 2010 al 2021 e previsione 2022*



### **Riferimenti bibliografici**

- Biggeri, L., Bini, M., Coli, A., Grassini, L., Maltagliati, M. (2012). Statistica per le decisioni aziendali. Pearson. Milano.
- Dagum, E. B. (2001). Analisi delle serie storiche: modellistica, previsione e scomposizione. Springer Science & Business Media.
- Girone, G., Crocetta, C., Massari, A. (2019). Statistica. Cacucci. Bari.
- Hamilton, J. D. (1994). Time series analysis. Princeton University.
- Innocenti R. La funzione statistica nei comuni. Atti della XII Conferenza Nazionale di Statistica, giugno 2016.
- Istat, Linee guida per la qualità dei processi statistici che utilizzano dati amministrativi, 2016.
- Mochi Sismondi, C. (2021). I gap informativi nelle amministrazioni pubbliche, ISTAT, 14 Conferenza nazionale di Statistica.

## Appendice

**Tabella 1.** *Decessi osservati e media decessi negli ultimi 5 anni nel Comune di Bari da febbraio 2020 a maggio 2021*

| <i>Prima ondata</i>              |                          |                                    |           | <i>Seconda ondata</i>            |                          |                                    |           | <i>Terza ondata</i>              |                          |                                    |           |
|----------------------------------|--------------------------|------------------------------------|-----------|----------------------------------|--------------------------|------------------------------------|-----------|----------------------------------|--------------------------|------------------------------------|-----------|
| <i>Settimana</i>                 | <i>Decessi osservati</i> | <i>media decessi ultimi 5 anni</i> | <i>%</i>  | <i>Settimana</i>                 | <i>Decessi osservati</i> | <i>media decessi ultimi 5 anni</i> | <i>%</i>  | <i>Settimana</i>                 | <i>Decessi osservati</i> | <i>media decessi ultimi 5 anni</i> | <i>%</i>  |
| 24 feb-1 mar                     | 69                       | 58                                 | 19        | 19 ott 25 ott                    | 65                       | 48                                 | 35        | 01 mar-06 mar                    | 95                       | 60,2                               | 58        |
| 2 mar-8 mar                      | 69                       | 59                                 | 16        | 26 ott-1 nov                     | 79                       | 47                                 | 67        | 08 mar-14 mar                    | 105                      | 59,4                               | 77        |
| 9 mar-15 mar                     | 72                       | 50                                 | 45        | 2 nov-8 nov                      | 68                       | 53                                 | 28        | 14 mar-20 mar                    | 100                      | 51,6                               | 94        |
| 16 mar-22 mar                    | 61                       | 54                                 | 13        | 9 nov-15 nov                     | 66                       | 52                                 | 27        | 21 mar-27 mar                    | 133                      | 57,2                               | 133       |
| 23 mar-29 mar                    | 83                       | 51                                 | 63        | 16 nov-22 nov                    | 69                       | 55                                 | 26        | 28 mar-03 apr                    | 120                      | 55,4                               | 117       |
| 30 mar-5 apr                     | 79                       | 52                                 | 52        | 23 nov-29 nov                    | 78                       | 54                                 | 44        | 04 apr-10 apr                    | 89                       | 61,2                               | 45        |
| 6 apr-12 apr                     | 73                       | 55                                 | 32        | 30 nov-6 dic                     | 81                       | 51                                 | 59        | 11 apr-17 apr                    | 91                       | 52,6                               | 73        |
| 13 apr-19 apr                    | 58                       | 50                                 | 17        | 7 dic-13 dic                     | 83                       | 59                                 | 41        | 18 apr-24 apr                    | 83                       | 55,8                               | 49        |
| 20 apr-26 apr                    | 58                       | 51                                 | 14        | 14 dic-20 dic                    | 79                       | 53                                 | 48        | 25 apr-01 mag                    | 89                       | 54                                 | 65        |
| 27 apr-3 mag                     | 76                       | 53                                 | 43        | 21 dic-27 dic                    | 79                       | 58                                 | 36        | 2 mag-08 mag                     | 67                       | 55                                 | 22        |
| 4 mag-10 mag                     | 59                       | 50                                 | 19        | 28 dic-3 gen                     | 84                       | 61                                 | 37        | 9 mag-15 mag                     | 55                       | 52                                 | 6         |
| 11 mag-17 mag                    | 58                       | 46                                 | 27        | 4 gen – 10 gen                   | 96                       | 70                                 | 37        | 16 mag-22 mag                    | 76                       | 51                                 | 48        |
| Incremento medio mortalità/sett. |                          |                                    | <b>30</b> | Incremento medio mortalità/sett. |                          |                                    | <b>41</b> | Incremento medio mortalità/sett. |                          |                                    | <b>65</b> |

**Tabella 2.** *Decessi osservati e dato medio negli ultimi 5 anni dei residenti nel Comune di Bari nella prima ondata della pandemia per fasce d'età.*

| <i>Settimana</i>              | <i>Decessi nella prima ondata</i> |       |           |               |       |           |               |       |           |               |       |          |
|-------------------------------|-----------------------------------|-------|-----------|---------------|-------|-----------|---------------|-------|-----------|---------------|-------|----------|
|                               | <b>85+</b>                        |       |           | <b>65-84</b>  |       |           | <b>50-64</b>  |       |           | <b>&lt;50</b> |       |          |
|                               | 2020/21                           | media | Δ         | 2020/21       | media | Δ         | 2020/21       | media | Δ         | 2020/21       | media | Δ        |
| 24 feb - 1 mar                | 33                                | 26    | 7         | 26            | 26    | 0         | 9             | 5     | 4         | 1             | 1     | 0        |
| 2 mar - 8 mar                 | 39                                | 33    | 6         | 26            | 22    | 4         | 3             | 4     | -1        | 1             | 0     | 1        |
| 9 mar - 15 mar                | 26                                | 24    | 2         | 36            | 19    | 17        | 7             | 5     | 2         | 3             | 1     | 2        |
| 16 mar - 22 mar               | 25                                | 25    | 0         | 29            | 23    | 6         | 5             | 5     | 0         | 2             | 1     | 1        |
| 23 mar-29 mar                 | 40                                | 25    | 15        | 38            | 22    | 16        | 3             | 3     | 0         | 2             | 1     | 1        |
| 30 mar - 5 apr                | 40                                | 25    | 15        | 31            | 20    | 11        | 7             | 5     | 2         | 1             | 2     | -1       |
| 6 apr - 12 apr                | 31                                | 24    | 7         | 31            | 25    | 6         | 10            | 5     | 5         | 1             | 2     | -1       |
| 13 apr - 19 apr               | 33                                | 25    | 8         | 20            | 17    | 3         | 4             | 6     | -2        | 1             | 2     | -1       |
| 20 apr - 26 apr               | 23                                | 23    | 0         | 24            | 22    | 2         | 10            | 4     | 6         | 1             | 1     | 0        |
| 27 apr - 3 mag                | 30                                | 26    | 4         | 35            | 24    | 11        | 9             | 2     | 7         | 2             | 1     | 1        |
| 4 mag-10 mag                  | 29                                | 23    | 6         | 23            | 22    | 1         | 4             | 3     | 1         | 2             | 2     | 0        |
| 11 mag-17 mag                 | 27                                | 23    | 4         | 21            | 18    | 3         | 10            | 3     | 7         | 0             | 1     | -1       |
| Totale                        | 376                               | 303   | <b>73</b> | 340           | 260   | <b>80</b> | 81            | 49    | <b>32</b> | 17            | 16    | <b>1</b> |
| Var. ass. relativa alla media | <b>+24,1%</b>                     |       |           | <b>+30,8%</b> |       |           | <b>+65,3%</b> |       |           | <b>+6,2%</b>  |       |          |

**Tabella 3.** *Decessi osservati e dato medio negli ultimi 5 anni dei residenti nel Comune di Bari nella seconda e terza ondata della pandemia per fasce d'età.*

| Settimana                     | Decessi nella seconda ondata |       |     |         |       |     |         |       |    |         |       |    |
|-------------------------------|------------------------------|-------|-----|---------|-------|-----|---------|-------|----|---------|-------|----|
|                               | 85+                          |       |     | 65-84   |       |     | 50-64   |       |    | <50     |       |    |
|                               | 2020/21                      | media | Δ   | 2020/21 | media | Δ   | 2020/21 | media | Δ  | 2020/21 | media | Δ  |
| 19 ott - 25 ott               | 31                           | 26    | 5   | 27      | 19    | 8   | 6       | 3     | 3  | 1       | 0     | 1  |
| 26 ott - 1 nov                | 38                           | 23    | 15  | 30      | 18    | 12  | 6       | 5     | 1  | 5       | 1     | 4  |
| 2 nov - 8 nov                 | 36                           | 25    | 11  | 29      | 22    | 7   | 3       | 5     | -2 | 0       | 1     | -1 |
| 9 nov - 15 nov                | 31                           | 24    | 7   | 27      | 23    | 4   | 6       | 4     | 2  | 2       | 1     | 1  |
| 16 nov - 22 nov               | 39                           | 26    | 13  | 26      | 23    | 3   | 3       | 4     | -1 | 1       | 2     | -1 |
| 23 nov - 29 nov               | 39                           | 26    | 13  | 30      | 23    | 7   | 8       | 3     | 5  | 1       | 2     | -1 |
| 30 nov - 6 dic                | 31                           | 26    | 5   | 42      | 19    | 23  | 7       | 5     | 2  | 1       | 1     | 0  |
| 7 dic - 13 dic                | 36                           | 30    | 6   | 41      | 23    | 18  | 4       | 5     | -1 | 2       | 1     | 1  |
| 14 dic - 20 dic               | 42                           | 24    | 18  | 33      | 24    | 9   | 2       | 4     | -2 | 2       | 1     | 1  |
| 21 dic - 27 dic               | 39                           | 31    | 8   | 31      | 22    | 9   | 8       | 4     | 4  | 1       | 1     | 0  |
| 28 dic - 3 gen                | 39                           | 31    | 8   | 38      | 24    | 14  | 4       | 5     | -1 | 3       | 1     | 2  |
| 4 gen - 10 gen                | 49                           | 34    | 15  | 42      | 30    | 12  | 4       | 5     | -1 | 1       | 2     | -1 |
| Totale                        | 450                          | 324   | 126 | 396     | 270   | 126 | 61      | 52    | 9  | 20      | 15    | 5  |
| Var. ass. relativa alla media | +38,9%                       |       |     | +46,7%  |       |     | +17,3%  |       |    | +33,3%  |       |    |

| Settimana                     | Decessi nella terza ondata |       |     |         |       |     |         |       |    |         |       |    |
|-------------------------------|----------------------------|-------|-----|---------|-------|-----|---------|-------|----|---------|-------|----|
|                               | 85+                        |       |     | 65-84   |       |     | 50-64   |       |    | <50     |       |    |
|                               | 2020/21                    | media | Δ   | 2020/21 | media | Δ   | 2020/21 | media | Δ  | 2020/21 | media | Δ  |
| 01mar-07 mar                  | 37                         | 32    | 5   | 37      | 19    | 18  | 9       | 3     | 6  | 3       | 0     | 3  |
| 08 mar-14 mar                 | 46                         | 30    | 16  | 47      | 23    | 24  | 12      | 5     | 7  | 0       | 1     | -1 |
| 15 mar-21 mar                 | 40                         | 22    | 18  | 50      | 23    | 27  | 8       | 5     | 3  | 2       | 1     | 1  |
| 22 mar-28 mar                 | 54                         | 28    | 26  | 64      | 24    | 40  | 11      | 3     | 8  | 4       | 2     | 2  |
| 29 mar-04 apr                 | 48                         | 26    | 22  | 66      | 22    | 44  | 3       | 6     | -3 | 3       | 2     | 1  |
| 05 apr-11 apr                 | 30                         | 27    | 3   | 50      | 26    | 24  | 7       | 6     | 1  | 2       | 1     | 1  |
| 12 apr-18 apr                 | 41                         | 27    | 14  | 40      | 19    | 21  | 8       | 5     | 3  | 2       | 2     | 0  |
| 19 apr-25 apr                 | 34                         | 24    | 10  | 39      | 24    | 15  | 8       | 7     | 1  | 2       | 1     | 1  |
| 26 apr-02 mag                 | 31                         | 27    | 4   | 41      | 23    | 18  | 13      | 3     | 10 | 4       | 1     | 3  |
| 3 mag-9 mag                   | 25                         | 25    | 0   | 36      | 23    | 13  | 6       | 5     | 1  | 0       | 2     | -2 |
| 10 mag-16 mag                 | 16                         | 25    | -9  | 29      | 22    | 7   | 7       | 4     | 3  | 3       | 2     | 1  |
| 17 mag-23 mag                 | 32                         | 23    | 9   | 29      | 22    | 7   | 9       | 5     | 4  | 6       | 1     | 5  |
| Totale                        | 434                        | 318   | 116 | 528     | 271   | 257 | 101     | 56    | 45 | 31      | 15    | 16 |
| Var. ass. relativa alla media | +36,5%                     |       |     | +94,8%  |       |     | +80,4%  |       |    | +106,7% |       |    |

**Tabella 4.** *Decessi mensili rilevati nel Comune di Bari dal 2010 al 2021*

| Mese      | Anni  |       |       |       |       |       |       |       |       |       |       |      |
|-----------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|------|
|           | 2010  | 2011  | 2012  | 2013  | 2014  | 2015  | 2016  | 2017  | 2018  | 2019  | 2020  | 2021 |
| Gennaio   | 289   | 340   | 362   | 370   | 367   | 449   | 425   | 471   | 476   | 489   | 432   | 631  |
| Febbraio  | 289   | 321   | 386   | 377   | 305   | 398   | 355   | 363   | 361   | 450   | 367   | 504  |
| Marzo     | 299   | 354   | 333   | 347   | 342   | 360   | 309   | 334   | 378   | 422   | 481   | 715  |
| Aprile    | 288   | 311   | 334   | 320   | 335   | 305   | 334   | 365   | 372   | 369   | 413   | 624  |
| Maggio    | 267   | 304   | 310   | 314   | 310   | 324   | 297   | 359   | 337   | 358   | 366   | 421  |
| Giugno    | 327   | 305   | 299   | 294   | 297   | 285   | 309   | 341   | 333   | 348   | 343   | 398  |
| Luglio    | 296   | 310   | 338   | 258   | 308   | 358   | 332   | 344   | 329   | 303   | 337   | 425  |
| Agosto    | 278   | 318   | 307   | 303   | 286   | 302   | 313   | 366   | 317   | 317   | 388   | 413  |
| Settembre | 271   | 289   | 264   | 272   | 288   | 315   | 279   | 338   | 318   | 342   | 365   | 371  |
| Ottobre   | 307   | 303   | 284   | 319   | 303   | 303   | 313   | 358   | 326   | 301   | 398   | 369  |
| Novembre  | 304   | 309   | 303   | 276   | 343   | 319   | 347   | 388   | 362   | 366   | 514   | n.d. |
| Dicembre  | 342   | 355   | 377   | 353   | 340   | 369   | 404   | 408   | 374   | 370   | 596   | n.d. |
| Totale    | 3.557 | 3.819 | 3.897 | 3.803 | 3.824 | 4.087 | 4.017 | 4.435 | 4.283 | 4.435 | 5.000 | n.d. |

**Tabella 5.** *Coefficienti netti di stagionalità*

| Mese         | Coefficienti di stagionalità $\widehat{S}_m$ |
|--------------|----------------------------------------------|
| Gennaio      | 1,180                                        |
| Febbraio     | 1,031                                        |
| Marzo        | 1,065                                        |
| Aprile       | 1,014                                        |
| Maggio       | 0,952                                        |
| Giugno       | 0,935                                        |
| Luglio       | 0,977                                        |
| Agosto       | 0,957                                        |
| Settembre    | 0,901                                        |
| Ottobre      | 0,928                                        |
| Novembre     | 0,975                                        |
| Dicembre     | 1,069                                        |
| <b>Media</b> | <b>1,00</b>                                  |

# **Efficacia del Benraluzimab nel trattamento dell'asma severo eosinofilico. Identificazione di subfenotipi di risposta**

**Andrea Tanga<sup>1</sup>, Angela Maria D'Uggento<sup>2\*</sup>, Danilo Di Bona<sup>3</sup>**

<sup>1</sup>Dottore in Scienze statistiche, <sup>2</sup>Dipartimento di Economia e Finanza, Università degli Studi di Bari Aldo Moro <sup>3</sup>Dipartimento dell'Emergenza e dei Trapianti di organi (DETO), Università degli Studi di Bari Aldo Moro

**Riassunto:** Il presente contributo presenta i principali risultati dell'analisi statistica condotta sugli effetti di un farmaco biologico, il Benralizumab, utilizzato nella terapia dell'asma severo. Si tratta di un anticorpo monoclonale umanizzato di tipo IgG1 $\kappa$  in grado di legarsi in modo selettivo alla catena  $\alpha$  del recettore dell'interleuchina 5 (IL-5R) bloccando il pathway di attivazione di IL-5 / IL5R nonché delle cellule effettrici. Con riferimento alla popolazione di pazienti sottoposti a tale terapia, lo studio retrospettivo si pone l'obiettivo di individuare le caratteristiche dei pazienti che hanno mostrato una migliore risposta alla terapia con Benralizumab al fine di creare un modello predittivo, presupposto per l'attivazione di percorsi di "medicina personalizzata". Le analisi svolte hanno consentito di individuare 4 sub-fenotipi di pazienti associati a un differente grado di risposta alla terapia. Le caratteristiche dei pazienti che meglio rispondono alla terapia.

**Keywords:** Benralizumab; studio retrospettivo; regressione logistica, cluster analysis.

---

\* Autore corrispondente: [angelamaria.duggento@uniba.it](mailto:angelamaria.duggento@uniba.it).

Il lavoro qui descritto è frutto di un progetto comune, tuttavia Tanga ha scritto l'introduzione, il paragrafo 2 e curato le elaborazioni e il commento dei risultati nel paragrafo 3; D'Uggento ha curato il paragrafo 3.1 e contribuito alla individuazione delle metodologie e Di Bona ha contribuito alla redazione del paragrafo 3 e redatto il paragrafo 4.

## 1. Introduzione

L'ASMA è una malattia caratterizzata da difficoltà respiratoria dovuta a riduzione del calibro dei bronchi per infiammazione della parete bronchiale, contrazione della muscolatura liscia bronchiale e caratterizzata da un'infiammazione cronica delle vie aeree che determina un'ostruzione che può variare notevolmente da soggetto a soggetto.

L'etiologia, al momento, è ignota e questo fa sì che sia oggetto di studio da parte di molti ricercatori di ambito medico. Si distinguono una forma estrinseca, i cui fattori scatenanti sono gli allergeni inalatori, ed una intrinseca, cioè non allergica, i cui trigger sono aspecifici. Il trattamento di fondo prevede l'utilizzo di farmaci anti-infiammatori (corticosteroidi orale o inalatorio), frequentemente associati ad un farmaco che dilata i bronchi, entrambi somministrati per via inalatoria.

La diagnosi di asma viene fatta considerando la storia del paziente, le caratteristiche dei sintomi (sibili, respiro corto e tosse) e i risultati degli esami finalizzati ad accertare il livello di limitazione del flusso aereo. Quest'ultimo aspetto viene osservato attraverso la risposta che si ottiene somministrando un farmaco broncodilatatore. La risposta, ovvero la capacità di espellere l'aria dai polmoni, può essere valutata con specifiche prove di funzionalità respiratoria, tra cui la più importante è quella denominata FEV1. FEV1 è l'acronimo di Forced Expiratory Volume in the 1st second, procedura che misura il volume di aria che il paziente riesce ad espirare in 1 secondo.

Per quanto riguarda la severità della patologia, si distinguono forme lievi, moderate e severe. Si definisce asma severo quella condizione che necessita della terapia massimale inalatoria per ottenere il controllo (no esacerbazioni, no sintomi moderati/severi). L'Asma grave è una patologia con specificità che la rendono differente dalle più diffuse forme di asma e, in Italia, colpisce tra il 5 e il 10% della totalità degli asmatici per un totale di circa 300.000 persone tra adulti, adolescenti e bambini.

In Italia, come in altre nazioni europee, la prevalenza dell'asma bronchiale nella popolazione è del 4-5 % con un tasso di incidenza di 3-4 casi ogni 1000 abitanti all'anno. Per quanto riguarda l'età, l'asma insorge con notevole frequenza in età infantile rappresentando la più comune malattia cronica e la principale causa di ricoveri ospedalieri nell'infanzia. Si può poi osservare un'attenuazione dei sintomi asmatici, o anche la scomparsa, in un 20-30% dei casi, all'epoca della pubertà soprattutto nei soggetti di sesso maschile. Un secondo picco di alta frequenza

dell'asma bronchiale si riscontra in età adulta (prevalenza 2,5%, incidenza 2 casi ogni mille all'anno).

Il sesso maschile è nettamente più colpito nei casi di asma infantile (nel rapporto di 2 a 1 tra maschi e femmine) mentre nell'adolescenza e nell'età adulta non c'è differenza tra i due sessi. La familiarità sembra giocare un ruolo importante nella precocità di comparsa dell'affezione. Per quanto riguarda i fattori ambientali la prevalenza è maggiore nelle aree urbane e nelle nazioni industrializzate.

L'Asma grave condiziona notevolmente la vita quotidiana delle persone che ne sono affette: scuola, lavoro e vita sociale. Il processo di diagnosi e cura di questa patologia ha ancora molta strada da percorrere per migliorare il livello di cura ed assistenza da garantire ai pazienti.

La severità dell'asma viene valutata in base alle caratteristiche dei sintomi ed ai trattamenti necessari per controllarli, nonché alla presenza di riacutizzazioni. L'asma severo viene trattato con l'associazione di più farmaci a dosaggi massimali.

Con riferimento alla terapia, essa utilizza differenti tipologie di farmaci:

- farmaci di fondo, per il trattamento di mantenimento da seguire in maniera continuativa; riducono l'infiammazione delle vie aeree, controllano i sintomi e riducono i rischi futuri, come le riacutizzazioni e il declino della funzione polmonare;
- farmaci sintomatici (o di emergenza): offrono sollievo immediato dai sintomi e possono, quindi, essere utilizzati al bisogno nelle fasi di riacutizzazione o di peggioramento dell'asma. Sono consigliati anche per la prevenzione a breve termine della possibile chiusura dei bronchi (bronco-costrizione) indotta dall'attività fisica;
- farmaci aggiuntivi: vengono utilizzati soprattutto nell'asma severo, quando i sintomi persistono o si verificano riacutizzazioni nonostante il trattamento ottimizzato con alte dosi dei farmaci di fondo e sono stati attuati, quando possibile, i provvedimenti per intervenire sui fattori di rischio modificabili.

Non sempre la terapia standard è sufficiente per tenere sotto controllo la malattia. Laddove queste terapie non abbiano successo, si può ricorrere ai farmaci biologici, il cui target è uno specifico pathway infiammatorio. La nuova terapia in uso in pazienti adulti affetti da asma grave di tipo eosinofilo non controllato, in aggiunta ai corticosteroidi inalatori e beta-agonisti a lunga durata d'azione, è basata sull'utilizzo del Benralizumab, un anticorpo monoclonale umanizzato di tipo IgG1 $\kappa$ , in grado di legarsi in modo selettivo alla catena  $\alpha$  del recettore dell'interleuchina 5 (IL-5R) bloccando il pathway di attivazione di IL-5 / IL5R nonché delle cellule effettrici.

Il contributo è organizzato nel modo seguente: dopo l'introduzione, nel paragrafo 2 sono presentati i dati e metodi statistici utilizzati, nel paragrafo 3 vengono illustrati i principali risultati ottenuti e, infine, brevi conclusioni sono discusse nel paragrafo 4.

## **2. Materiali e metodi**

Il campione analizzato nello studio è composto da 160 pazienti affetti da asma severo, residenti in varie zone del sud Italia e in cura presso del Dipartimento dell'Emergenza e dei Trapianti di organi (DETO) del Policlinico di Bari. A ciascun paziente è stato somministrato il trattamento a base di Benralizumab ad un tempo T0 (considerato la baseline), al tempo T6 (dopo 6 mesi) per verificare quali siano state le risposte a tale trattamento in relazione alle caratteristiche dei pazienti di interesse per il clinico.

I dati sono stati rilevati dalle schede dei pazienti, gestite in formato elettronico dall'equipe medica di ricerca. Le variabili di interesse per l'analisi statistica sono state le seguenti:

- età
- genere
- BMI (body mass index)
- status di fumatore (classificato in tre categorie: sì, no, ex fumatore)
- età di insorgenza della malattia
- anni di presenza della malattia
- ereditarietà (atopia familiarità)
- numero di riacutizzazioni
- somministrazione corticosteroidi orali – OCS (dicotomica: sì/no)
- dose media prednisone
- esito dell'ACT - Asthma Control Test
- FEV1% (flusso massimo espiratorio massimo al 1° secondo)
- livello di eosinofili ematici
- livello di IgE totali

Ai fini del monitoraggio dell'efficacia della terapia, sono state utilizzate le seguenti variabili, rilevate rispettivamente ai tempi T0 e T6:

- numero di riacutizzazioni
- dose media prednisone
- ACT

- FEV1% (flusso massimo espiratorio massimo al 1° secondo)

La variabile numero di riacutizzazioni (esacerbazioni) è stata ricodificata sulla base del periodo di tempo in cui i sintomi hanno mostrato un peggioramento, con conseguente intensificarsi della terapia, espressa come variazione percentuale: se  $\Delta=100\%$  è stata attribuita l'etichetta "Super risposta", se  $50\% < \Delta < 100\%$  l'etichetta "Risposta" e se  $\Delta < 50\%$ , quella di "Non risposta".

La variabile dose media prednisone, che rappresenta la dose media di corticosteroide somministrato per via orale al paziente per combattere le infezioni causate dall'asma, è stata anch'essa oggetto di ricodifica col medesimo criterio della variazione percentuale usato per il numero di riacutizzazioni.

La variabile ACT riporta gli esiti del test per il controllo dell'asma ed esprime, attraverso una scala numerica (da 0=peggiore a 25=migliore), lo stato di malattia. Dal punteggio di 20 in su, la malattia è considerata controllata. Ai fini della presente analisi, la valutazione è effettuata attraverso una differenza tra T0 e T6 attribuendo le seguenti etichette: "Super risposta" se  $20 \leq \Delta \leq 25$ ; "Risposta" se  $3 \leq \Delta < 20$ ; "Non risposta" se  $\Delta < 3$ .

La variabile FEV1, espressa in litri o in percentuale rispetto alla norma, indica il grado di ostruzione delle vie aeree; è il parametro più oggettivo, ma non correla con la sintomatologia in modo chiaro e, dunque, viene considerato come una misura di beneficio surrogato, poiché, per il clinico è più importante controllare il sintomo, piuttosto che il miglioramento di un parametro strumentale. La variazione è stata valutata in termini percentuali con il seguente criterio: una variazione  $\Delta > 80\%$  è stata ricodificata come "Super risposta", se  $12\% \leq \Delta \leq 80\%$ , "Risposta", se  $\Delta < 12\%$  "Non risposta".

Una ulteriore variabile è stata ottenuta "incrociando" gli esiti "assenza esacerbazioni" e "assenza corticosteroide", (denominata ESA0andOCS0), ottenendo una variabile risposta politomica a tre categorie in cui:

- viene attribuita la categoria "super risposta" se le classificazioni delle variazioni percentuali attribuite alle variabili "numero riacutizzazioni" e "dose media di corticosteroide" per ciascun paziente sono entrambe "super risposta";
- se le precedenti classificazioni sono una "risposta" e l'altra "super risposta", la categoria sarà "risposta";
- se una delle due è "non rispondente", sarà attribuita la categoria "non risposta".

Prima di iniziare l'analisi è stata fatta una ulteriore ricodifica della variabile risposta ESA0andOCS0 trasformandola in binaria con i seguenti esiti: "Super risposta" e "Non super risposta".

Le analisi statistiche sono state realizzate con il software IBM SPSS Statistical Package for Social Sciences. In particolare, dopo il test di Shapiro-Wilk per la verifica della normalità delle distribuzioni, per capire quali variabili fossero più rilevanti per individuare le caratteristiche del paziente che ha sviluppato una super risposta al trattamento, sono state utilizzate la regressione logistica multivariata, l'analisi dei cluster. Queste ultime tecniche sono state usate per porre le basi della c.d. medicina personalizzata, che si pone l'obiettivo di dare al clinico gli strumenti conoscitivi per individuare la giusta dose di farmaco da somministrare al paziente, dati alcuni predittori.

### 3. Risultati

#### 3.1 *Analisi descrittiva delle caratteristiche dei pazienti*

I 160 pazienti coinvolti in questo studio preliminare sono di genere prevalentemente femminile (68,1%), hanno età media di 53,8 anni ( $\pm 13,9$  anni), che sale a 56,0 anni per le donne e si attesta sui 53 anni per gli uomini. La tabella 1 mostra i dati rilevati al tempo 0 sui pazienti, distinti per genere.

**Tabella 1.** *Principali caratteristiche dei pazienti arruolati nello studio secondo il genere (medie e percentuali)*

| Caratteristiche (medie)     | F      | M      | Caratteristiche (%)      | F    | M    |
|-----------------------------|--------|--------|--------------------------|------|------|
| Età (anni compiuti)         | 54     | 53     | Fumo                     |      |      |
| Peso (kg)                   | 67,12  | 81,81  | No                       | 76,2 | 70,6 |
| Altezza (cm)                | 159,63 | 172,71 | Si                       | 7,3  | 5,9  |
| BMI                         | 26,08  | 27,11  | ex fumatore              | 16,5 | 23,5 |
| Anni di malattia            | 21     | 20     | Familiarità atopia (=si) | 56,0 | 29,4 |
| Numero riscutizzazioni/anno | 5      | 5      | OCS (si)                 | 59,6 | 58,8 |
| Dose prednisone             | 7      | 8      | Poliposi nasale (=si)    | 52,7 | 67,4 |
| ACT                         | 13     | 14     | Bronchiectasie (=si)     | 22,9 | 19,6 |
| FEV1pre (litri)             | 1,64   | 2,24   |                          |      |      |
| FEV1pre (%)                 | 66,9   | 62,8   | Esito trattamento        |      |      |
| Eosinofili ematici          | 825    | 998    | NonSuperRisposta         | 32,2 | 40,5 |
| IgE totali                  | 407    | 423    | SuperRisposta            | 67,8 | 59,5 |

### 3.2 Analisi multivariate: modello logistico e analisi dei cluster

Per individuare le variabili che hanno maggior influenza sulla risposta del paziente alla terapia sperimentale è stato stimato un modello logistico multivariato, in cui la variabile risposta binaria è stata codificata attribuendo il valore 1 a tutti i soggetti che, dopo un determinato periodo dalla somministrazione del farmaco, hanno azzerato i propri livelli di ESA e OCS, etichettati come Super-rispondenti, e 0 nei restanti casi.

I predittori inseriti nel modello sono stati: età, età di insorgenza della malattia, anni di malattia, atopia, esinofili, Ige, FEV1%, numero di esacerbazioni, poliposi nasale e bronchiectasie.

Utilizzando la procedura stepwise backward con il metodo Wald, sono risultate significative (con  $\alpha=0,05$ ) solo l'età del paziente ( $p\text{-value}=0,020$ ), l'età di insorgenza della malattia ( $p\text{-value}=0,050$ ), la poliposi nasale ( $p\text{-value}=0,034$ ), e la somministrazione di azitromicina ( $p\text{-value}=0,002$ ). In particolare, l'età di insorgenza dell'asma e l'assunzione di azitromicina sono risultati fattori di protezione, come evidenziato dagli odds ratio ( $\text{Exp(B)}$ ) in Tabella 2.

**Tabella 2.** *Modello di regressione logistica binaria per i pazienti con super risposta*

| Predittori          | B      | S.E.  | Wald  | df | p-value | Exp(B) |
|---------------------|--------|-------|-------|----|---------|--------|
| Età                 | 0,039  | 0,017 | 5,426 | 1  | 0,001   | 1,040  |
| Età insorgenza asma | -0,440 | 0,022 | 3,854 | 1  | 0,049   | 0,957  |
| Azitromicina        | -1,818 | 0,597 | 9,276 | 1  | 0,002   | 0,162  |
| Poliposi nasale     | 0,858  | 0,404 | 4,519 | 1  | 0,034   | 2,359  |
| Costante            | 2,159  | 0,995 | 4,707 | 1  | 0,030   | 8,666  |

La cluster analysis è stata implementata considerando le medesime variabili usate come predittori nel modello logistico.

È stato effettuato un clustering gerarchico con il metodo Ward e la distanza euclidea al quadrato, ottenendo 4 gruppi: il primo composto da 23 pazienti, il secondo da 19, il terzo da 35, il quarto da 28. Il numero dei pazienti considerati nella analisi dei gruppi è sceso da 160 a 105 a causa di alcuni dati mancanti nel dataset.

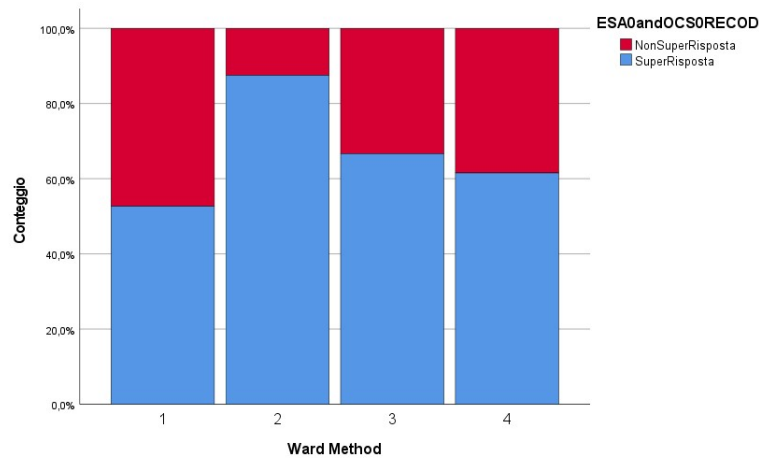
Ai fini della valutazione del clinico, il cluster più significativo è risultato essere il secondo, in cui si trova il maggior numero di super-rispondenti (Fig. 1).

Le caratteristiche dei cluster possono essere ricostruite attraverso i dati esposti nella Tabella 3. I pazienti che hanno risposto meglio alla terapia risultano essere coloro che hanno un'età media e un'età di insorgenza media inferiore agli altri, che

hanno un flusso espiratorio massimo al primo secondo in termini percentuali più alto, con un alto indice di massa corporea, a cui si associa solitamente uno stato infiammatorio più elevato, coloro che presentano un livello di eosinofili più alto, anch'essi sintomatici di livelli maggiori di infiammazione, che presentano poche bronchiectasie e un minor numero di polipi nasali.

In altri termini, la cluster analysis ha evidenziato che il paziente che ha una infiammazione più elevata, ma il cui flusso espiratorio è maggiormente conservato grazie alla terapia massimale inalatoria utilizzata, risponde meglio degli altri.

**Figura 1.** Distribuzione percentuale dei pazienti nei quattro cluster secondo la variabile super-risposta si/no.



Al contrario, la peggiore risposta alla terapia si riscontra per il cluster 1, caratterizzato da una più elevata presenza di patologia allergica come evidenziato dal prick test positivo in tutti i casi e dall'esecuzione in anamnesi di immunoterapia.

**Tabella 3.** Caratteristiche dei pazienti secondo i cluster

| Variabili                    | Cluster 1 | Cluster 2 | Cluster 3 | Cluster 4 |
|------------------------------|-----------|-----------|-----------|-----------|
| Età (media)                  | 58        | 35        | 55        | 62        |
| Età insorgenza asma (media)  | 46        | 24        | 27        | 45        |
| FEV1 pre (%)                 | 55,7      | 79,3      | 54,1      | 62,2      |
| Eosinofili ematici (medio)   | 613       | 1363      | 992       | 891       |
| BMI (medio)                  | 26,8      | 28,3      | 26,3      | 26,3      |
| Poliposi nasale (% si)       | 34,8      | 47,4      | 65,7      | 60,7      |
| Bronchiectasie               | 17,4      | 21,1      | 42,9      | 17,9      |
| Azitromicina (assunzione=si) | 13,0      | 0,0       | 22,9      | 21,4      |

#### 4. Conclusioni

Le tecniche statistiche trovano un impiego sempre crescente nelle analisi di dati clinici, soprattutto in quella che è la nuova sfida della “medicina personalizzata”.

Impiegate negli studi retrospettivi, consentono di individuare i fattori di rischio che hanno maggiormente influito sulla risposta ai protocolli sperimentali, orientando le scelte dei medici e rendendo le diagnosi più accurate.

Tenuto conto della variabilità dei pazienti, le analisi di cluster sembrano essere molto informative, poiché consentono di identificare diversi gruppi di pazienti simili tra loro per numerose caratteristiche. Questo ha consentito di indentificare 4 sottogruppi di pazienti all'interno del sottogruppo T2. Dunque, contrariamente a quanto ritenuto, cioè che il fenotipo T2 sia abbastanza omogeneo e, quindi, target ideale di terapia con il Benralizumab, in realtà si osserva una notevole eterogeneità tra i pazienti, che condiziona la risposta clinica.

La sperimentazione, condotta dall'equipe del Dipartimento dell'Emergenza e dei Trapianti di organi (DETO) del Policlinico di Bari, di cui fa parte il prof. Di Bona, proseguirà lo studio con l'integrazione di database multicentrici, rilevati sull'intero territorio nazionale. Ciò consentirà di disporre di popolazioni più numerose e database più ricchi di informazioni sui pazienti tali da giustificare in ricorso anche ad altre tecniche di analisi, tra cui quelle di machine learning.

Le due tecniche multivariate utilizzate, regressione logistica e analisi dei cluster, hanno consentito di delineare il profilo dei pazienti che meglio hanno risposto alla terapia con Benralizumab in base alla risposta infiammatoria dei biomarker e alle caratteristiche cliniche. Questi risultati sono utili per individuare percorsi terapeutici differenti e modulabili.

#### Riferimenti bibliografici

- Agache I, Song Y, Rocha C, et al. (2020). Efficacy and safety of treatment with dupilumab for severe asthma: a systematic review of the EAACI guidelines-Recommendations on the use of biologicals in severe asthma. *Allergy*. 75(5):1058-1068.
- Bleecker ER, FitzGerald JM, Chanez P, et al. (2016). SIROCCO study investigators. Efficacy and safety of benralizumab for patients with severe asthma uncontrolled with high-dosage inhaled corticosteroids and long-acting  $\beta_2$ -agonists (SIROCCO): a randomised, multicentre, placebo-controlled phase 3 trial. *Lancet*. 388(10056):2115-2127.

- Brown T, Jones T, Gove K, et al. (2018). Randomised controlled trials in severe asthma: selection by phenotype or stereotype. *Eur Respir J.* 52(6):1801444.
- Di Bona D., Crimi, C., D'Uggento, A. M. et al. (2021). Effectiveness of benralizumab in severe eosinophilic asthma: distinct sub-phenotypes of response identified by cluster analysis. *Clin Exp Allergy.* 00:1–12. DOI: 10.1111/cea.14026.
- Di Bona D, Minenna E, Albanesi M, Nettis E, Caiaffa MF, Macchia L. (2020). Benralizumab improves patient reported outcomes and functional parameters in difficult-to-treat patients with severe asthma: data from a real-life cohort. *Pulm Pharmacol Ther.* 64:101974.
- FitzGerald JM, Bleecker ER, Nair P, et al. (2016). CALIMA study investigators. Benralizumab, an anti-interleukin-5 receptor  $\alpha$  monoclonal antibody, as add-on treatment for patients with severe, uncontrolled, eosinophilic asthma (CALIMA): a randomised, double-blind, placebocontrolled phase 3 trial. *Lancet.* 388(10056):2128-2141.
- Fowlkes, E. B., Gnanadesikan, R., Kettinger, J.R. (1988). Variable selection in clustering, *Journal of Classification*, 5: 205-228.
- Friedman, J.H., Meulman, J.J. (2004). Clustering objects on subsets of attributes. *Journal of the Royal Statistical Society, series B*, 66 (4): 815:849.
- Global Initiative for Asthma. Global Strategy for Asthma Management and Prevention, 2020. Available from: [www.ginasthma.org](http://www.ginasthma.org).
- Kroes JA, Zielhuis SW, van Roon EN, Ten Brinke A. (2020). Prediction of response to biological treatment with monoclonal antibodies in severe asthma. *Biochem Pharmacol.* 179:113978.
- Price D, Bateman ED, Chisholm A, et al. (2014). Complementing the randomized controlled trial evidence base. evolution not revolution. *Ann Am Thorac Soc.* 11(Suppl. 2): S92-S98.
- Rabe K.F., Nair P, Brusselle G, et al. (2018). Efficacy and safety of dupilumab in glucocorticoid-dependent severe asthma. *N Engl J Med.* 2018;378(26):2475-2485.

# La valutazione delle azioni di sostenibilità ambientale delle imprese attraverso i censimenti permanenti

Agata Maria Madia Carucci<sup>1</sup>, Giovanni Vannella<sup>2\*</sup>.

<sup>1</sup>Istat, <sup>2</sup>Dipartimento di Economia e Finanza, Università degli Studi di Bari Aldo Moro

**Riassunto:** Con il presente lavoro si illustrano le potenzialità esplicative in riferimento alle valutazioni delle azioni di sostenibilità ambientale poste in essere dalle imprese, desumibili dai censimenti permanenti, unite ad una loro analisi statistica territoriale.

**Keywords:** censimenti permanenti; sostenibilità ambientale aziendale; analisi territoriali.

## 1. Introduzione

La sempre maggiore attenzione verso le forme di economia sostenibile passa anche attraverso una maggiore sensibilità ambientale da parte non solo delle famiglie ma altresì delle realtà imprenditoriali.

Le misurazioni delle connessioni economia – ambiente sono da tempo oggetto di rilevanti studi da parte dei ricercatori e già esistono approcci finalizzati all'analisi congiunta dei due aspetti, come, ad esempio, il SEEA (*System of Environmental – Economic Accounting*), 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*), etc.

Nel tempo si è assistito ad un notevole impegno finalizzato ad affiancare alle vecchie indagini sulle imprese, dicotomicamente tipicamente distinguibili in censuarie e campionarie, la sempre maggiore utilizzabilità di archivi integrati

---

\* Autore corrispondente: [giovanni.vannella@uniba.it](mailto:giovanni.vannella@uniba.it). Sebbene il presente articolo sia frutto del lavoro congiunto dei due autori, a A.M.M. Carucci sono attribuiti i par. 2 e 4, a G. Vannella i par. 1, 3, 5.

amministrativi, con riferimento ad esempio ai dati fiscali (Branchi, 2008) e statistici. Il frutto di tale impegno è visibile nella diffusione dei dati sulle statistiche strutturali delle imprese (Boselli, 2016; Seri, 2016), nelle innovazioni introdotte nella rilevazione sulle piccole e medie imprese e sull'esercizio di arti e professioni (Casciano, 2011) e, non da ultimo, dall'importante output dell'offerta di statistiche ufficiali fornito dalla produzione dei censimenti permanenti che stanno rendendo possibile l'estrapolazione di informazioni prima inimmaginabili.

Già in passato gli autori del presente lavoro (Carucci, 2017) avevano avuto modo di approfondire il confronto tra i dati forniti dai conti regionali dei flussi di materia (Istat, 2010; Tudini, 2015; Istat, 2016) con i dati provenienti dal frame-sbs e, date le interessanti risultanze emerse, considerando le recenti innovazioni fornite dai censimenti permanenti delle imprese che riescono da pochissimo ad "intercettare" informazioni circa le azioni di sostenibilità ambientale messe in opera dalle imprese, si è voluto in questa sede da un lato approfondire le potenzialità esplorative espresse da tale nuova offerta informativa, dall'altro tentare un approccio di analisi.

## **2. Le azioni di sostenibilità ambientale nei censimenti permanenti**

Il primo Censimento permanente delle imprese si è svolto nel 2019. L'obiettivo della rilevazione è quello di aggiornare il quadro sulla struttura e sulla competitività delle imprese per cogliere i cambiamenti del sistema produttivo nazionale. Le informazioni sono desunte dall'integrazione fra i registri statistici e le indagini economiche correnti, cui si aggiungono nuove rilevazioni dirette mirate all'analisi di aspetti emergenti di grande rilevanza per la competitività, la sostenibilità sociale e ambientale delle imprese, la crescita economica del Paese (cfr. Istat, 2020).

È possibile accedere direttamente al datawarehouse dei censimenti permanenti dal sito <https://dati.istat.it/> attraverso cui si possono consultare i dati in forma aggregata, la cui struttura è rappresentata analiticamente in Fig 1.

Da tale sezione dei dati dei censimenti permanenti delle imprese, è possibile accedere alle prime quattro tematiche che interessano in questa sede:

- azioni per la sostenibilità ambientale, responsabilità sociale e sicurezza;
- strategie per la sostenibilità ambientale;
- strategie per il benessere lavorativo e di interesse collettivo;
- strategie a beneficio del territorio e per la sicurezza.

**Fig. 1.** *Classificazione dei dati del censimento permanente delle imprese in tema di sostenibilità – parte 1*

|                                                                           |
|---------------------------------------------------------------------------|
| <b>■ CENSIMENTO DELLE IMPRESE</b>                                         |
| ▣ Azioni per sostenibilità ambientale, responsabilità sociale e sicurezza |
| ▣ Strategie per la sostenibilità ambientale                               |
| ▣ Strategie per il benessere lavorativo e di interesse collettivo         |
| ▣ Strategie a beneficio del territorio e per la sicurezza                 |
| ▣ Proprietà, controllo, gestione                                          |
| ▣ Acquisizione di risorse umane                                           |
| ▣ Formazione non obbligatoria                                             |
| ▣ Relazioni tra imprese                                                   |
| ▣ Competitività e mercato                                                 |
| ▣ Fonti di finanziamento                                                  |
| ▣ Internazionalizzazione produttiva                                       |
| ▣ Innovazione e piattaforme digitali                                      |
| ▣ Software gestionali, cloud, investimenti digitali                       |
| ▣ Formazione e competenze digitali                                        |

Fonte: <https://dati.istat.it>

L’offerta informativa relativa alla tematica “azioni per la sostenibilità ambientale, responsabilità e sicurezza”, utilizzata nel presente lavoro, è strutturata come rappresentato in Fig. 2 e permette di ottenere dati aggregati secondo i seguenti fattori di classificazione:

- imprese attive;
- imprese attive che riducono l’impatto ambientale delle proprie attività;
- imprese attive che migliorano il benessere lavorativo;
- imprese attive che sostengono o realizzano iniziative di interesse collettivo esterne all’impresa;
- imprese attive che incrementano i livelli di sicurezza all’interno dell’impresa o nel territorio in cui opera l’impresa;

- imprese attive che realizzano almeno un'azione di sostenibilità ambientale e/o di responsabilità sociale e/o di sicurezza.

**Fig. 2.** *Classificazione dei dati del censimento permanente delle imprese in tema di sostenibilità – parte 2*

|                                                                           |
|---------------------------------------------------------------------------|
| ■ CENSIMENTO DELLE IMPRESE                                                |
| ■ Azioni per sostenibilità ambientale, responsabilità sociale e sicurezza |
| ■ Macrosettori economici - prov                                           |
| ■ Sezioni Ateco, classe di addetti - reg.                                 |
| ■ Divisione di Ateco (2 digit), classe di addetti                         |
| ■ Classe di Ateco (4 digit)                                               |
| ■ Motivazioni                                                             |
| ■ Macrosettori economici - prov                                           |
| ■ Sezioni Ateco, classe di addetti - reg.                                 |
| ■ Divisione di Ateco (2 digit), classe di addetti                         |
| ■ Classe di Ateco (4 digit)                                               |
| ■ Strumenti finanziari                                                    |
| ■ Macrosettori economici - prov                                           |
| ■ Sezioni Ateco, classe di addetti - reg.                                 |
| ■ Divisione di Ateco (2 digit), classe di addetti                         |
| ■ Classe di Ateco (4 digit)                                               |
| ■ Modalità di finanziamento                                               |
| ■ Macrosettori economici - prov                                           |
| ■ Sezioni Ateco, classe di addetti - reg.                                 |
| ■ Divisione di Ateco (2 digit), classe di addetti                         |
| ■ Classe di Ateco (4 digit)                                               |

Fonte: <https://dati.istat.it>

La sezione relativa alle “strategie per la sostenibilità ambientale”, i cui dati saranno presentati di seguito, si articola come rappresentato in Fig. 3.

**Fig. 3.** *Classificazione dei dati del censimento permanente delle imprese in tema di sostenibilità ambientale*



Fonte: <https://dati.istat.it>

Tale sezione permette di ottenere la stima delle imprese attive con tre o più addetti che riducono l'impatto ambientale delle proprie attività articolate per le singole tipologie specifiche di azioni poste in essere per ridurre il consumo di risorse naturali e gestire in modo sostenibile rifiuti ed emissioni:

- contenimento dei prelievi e dei consumi di acqua;
- trattamento delle acque di scarico finalizzato al contenimento ed al controllo di inquinanti;
- riutilizzo e riciclo delle acque di scarico;
- risparmio del materiale utilizzato nei processi produttivi;

- utilizzo di materie prime seconde (scarti del processo produttivo recuperati e riammessi nella produzione);
- raccolta differenziata e riciclo dei rifiuti;
- gestione dei rifiuti finalizzata al contenimento e controllo degli inquinanti;
- contenimento delle emissioni atmosferiche;
- contenimento dell'inquinamento acustico e rumoroso;
- ricorso a fornitori che già adottavano processi finalizzati a ridurre l'impatto ambientale delle loro attività.

La sottosezione delle strategie di promozione permette di evidenziare quali siano le soluzioni adottate nell'ultimo triennio per promuovere iniziative di sostenibilità ambientale, ovvero:

- ridisegnare il processo produttivo e/o adottare nuovi modelli di produzione;
- acquisire certificazioni ambientali volontarie di prodotto o di processo;
- redigere bilanci e/o rendicontazioni ambientali e di sostenibilità;
- nominare un referente interno e/o istituire una struttura per la responsabilità ambientale;
- svolgere con personale interno o esterno iniziative di formazione per la protezione dell'ambiente e la gestione delle risorse naturali;
- altre soluzioni.

L'ultima sottosezione relativa alle tipologie di investimento premette di incrociare i dati relativi al tipo di investimento effettuato con l'aver usufruito o meno di incentivi.

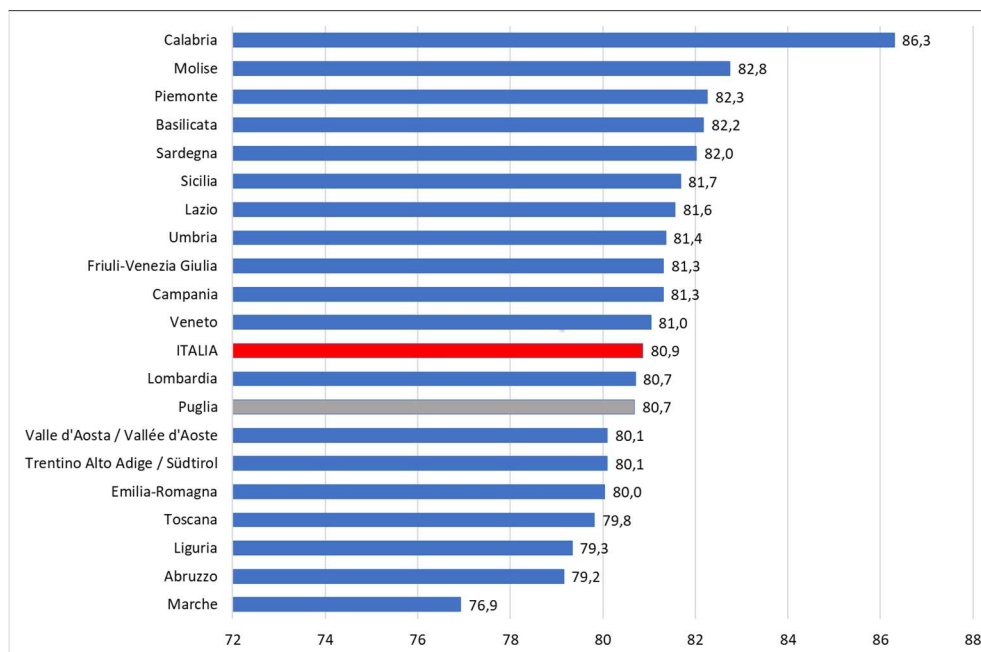
### **3. Caratteristiche delle imprese che attuano azioni di sostenibilità**

Alla luce di poter interpretare le potenzialità esplorative delle variabili ambientali evidenziate dai censimenti permanenti, si è analizzata la distribuzione regionale delle imprese che adottano almeno un'azione di sostenibilità (ambientale o di responsabilità sociale) rapportando, come primo step, il numero di imprese attive che realizzano almeno una azione di sostenibilità al numero totale di imprese attive (Graf. 1).

Le risultanze emerse evidenziano una forte attenzione da parte delle imprese alle azioni di sostenibilità, con una media a livello nazionale dell'80,9% e un valore minimo osservato (per le Marche) comunque piuttosto elevato (76,9%).

L'incidenza più alta di imprese che attuano almeno una azione di sostenibilità ambientale viene osservata in Calabria (86,3%), aspetto questo tuttavia scarsamente rappresentativo stante il ridotto range osservato (76,9% - 86,3%).

**Grafico 1.** *Imprese con almeno un'azione di sostenibilità ambientale, responsabilità sociale e sicurezza*

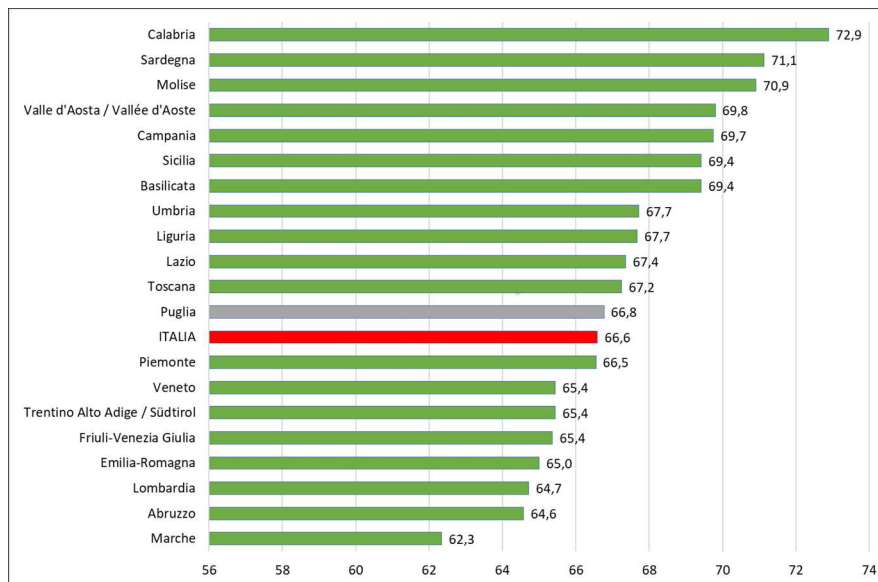


Fonte: Elaborazione su dati Istat

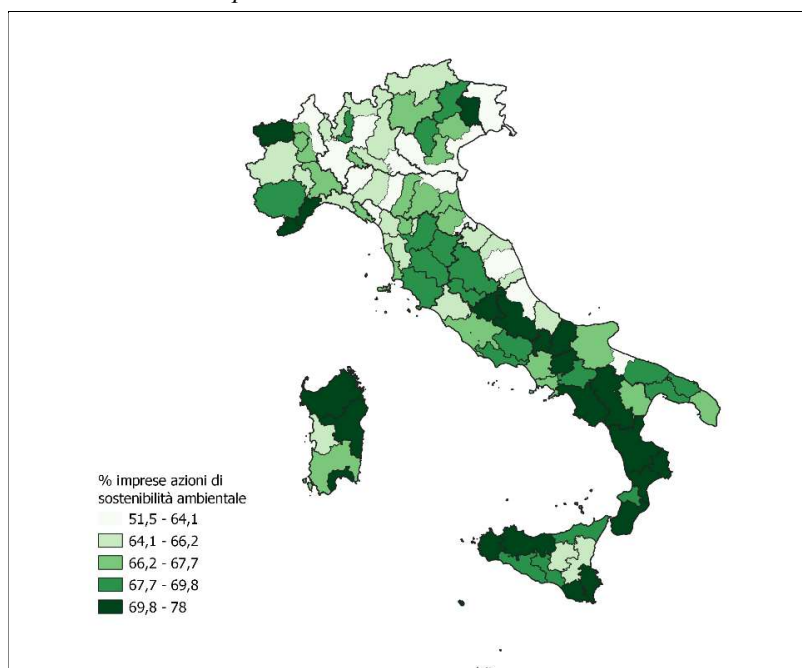
Si è ripetuta l'analisi valutando questa volta un dato più disaggregato, ovvero la percentuale di imprese che riducono l'impatto ambientale delle proprie attività, con riferimento alle aggregazioni regionali (Graf. 2) ed alle aggregazioni provinciali (Graf. 3).

I valori osservati sono ovviamente più bassi trattandosi della distribuzione di imprese che effettuano una azione di sostenibilità ambientale contro la precedente distribuzione in cui si valutavano oltre a tale variabile anche l'assolvimento di azioni di responsabilità sociale e sicurezza.

Ciò che emerge dall'analisi è come, a fronte di significative variabilità a livello intra-regionale, e di un valore medio nazionale pari al 66,6%, il dato calabrese sembra essere confermato pur se in presenza nuovamente di un range ridotto (62,3-72,9%).

**Grafico 2.** *Imprese con un'azione di sostenibilità ambientale*

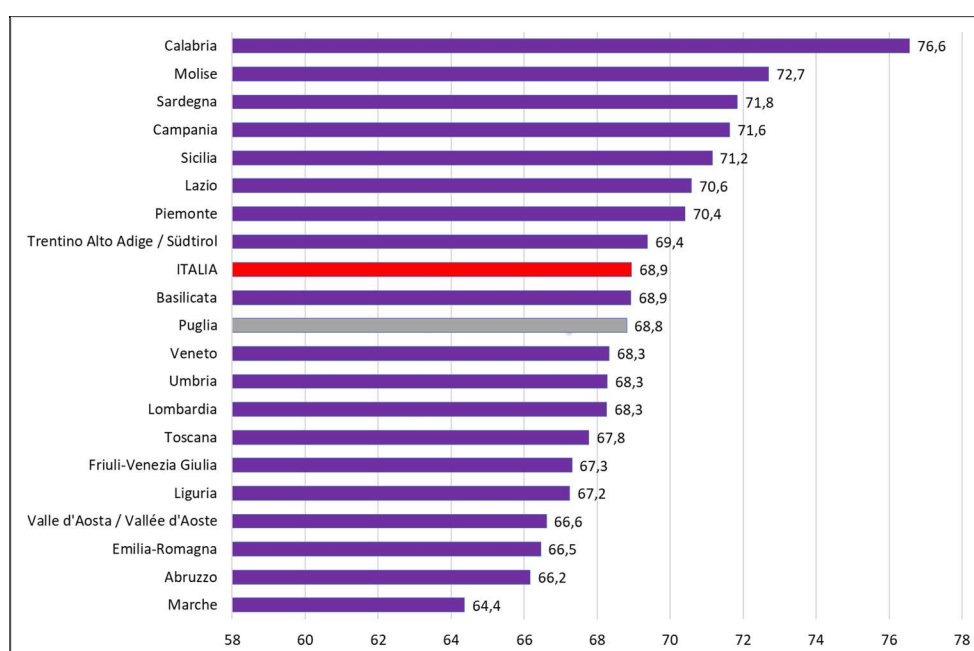
Fonte: Elaborazione su dati Istat

**Grafico 3.** *Imprese con un'azione di sostenibilità ambientale*

Fonte: Elaborazione su dati Istat

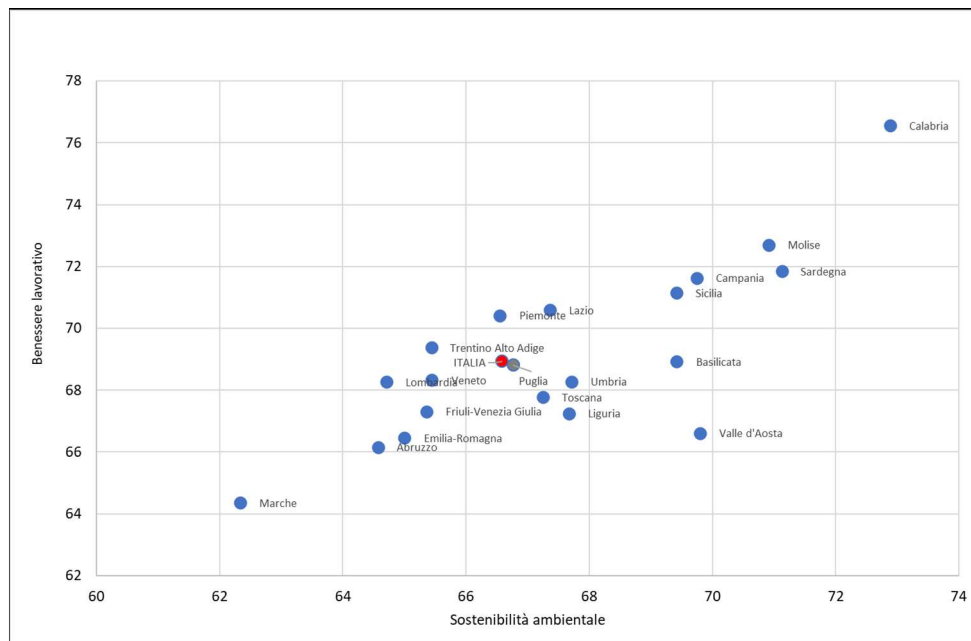
Ci si è quindi chiesto se le aree in cui le imprese avessero una relativamente forte attenzione ambientale fossero accomunate dall'altra manifestazione dell'attenzione sociale, ovvero il benessere lavorativo (Graf. 4), che appare ricalcare in grandi linee quanto precedentemente visto, al punto tale da avere reso opportuno condurre un confronto diretto delle due variabili (Graf. 5) che evidenzia come, al crescere dell'attenzione verso le tematiche ambientali, cresca quella verso il benessere lavorativo e viceversa.

**Grafico 4.** *Imprese con un'azione di benessere lavorativo*



Fonte: Elaborazione su dati Istat

**Grafico 5.** *Relazione tra incidenza di imprese con azioni di sostenibilità ambientale e benessere lavorativo.*

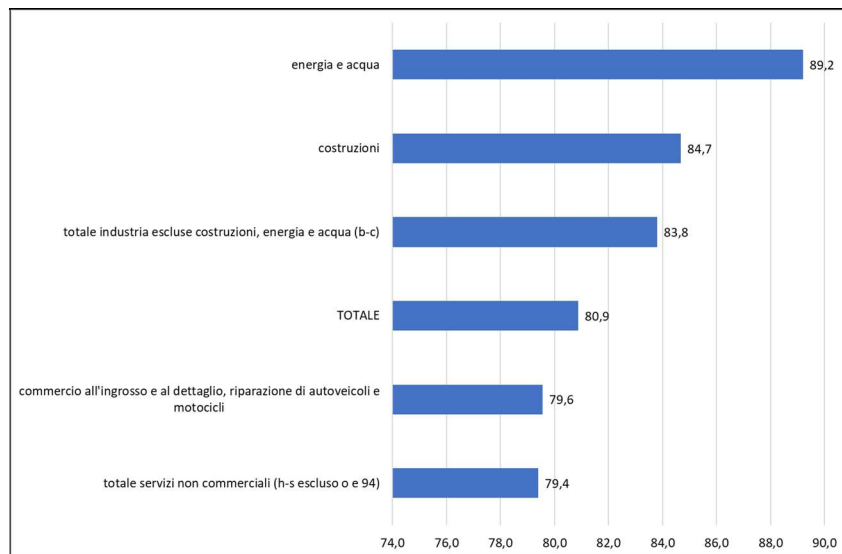


Fonte: Elaborazione su dati Istat

Esaminata una prima ipotesi di variabilità del fenomeno a livello territoriale, si è ritenuto opportuno verificare se la variabilità del fenomeno potesse avere una spiegazione in termini di attività economica specifica e/o di dimensionalità delle imprese e a tal fine si è condotta la prima analisi sulla variabile maggiormente aggregata, ovvero la generica presenza di almeno una azione di sostenibilità ambientale, responsabilità sociale e sicurezza.

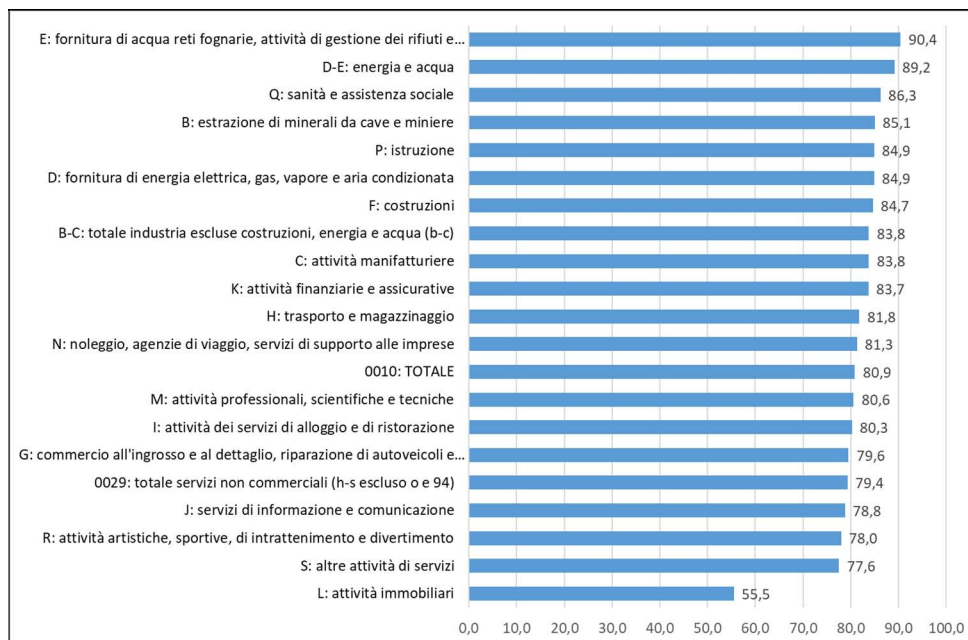
Come si vede dal Graf. 6, a fronte di una incidenza media dell'80,9% delle imprese che realizzano almeno una azione di sostenibilità (ambientale e benessere lavorativo), superano il livello medio il macrosettore "energia ed acqua" (89,2%), quello delle costruzioni (84,7%) e quello dell'industria (83,8%). Aspetto questo che trova ovvia conferma con riferimento alla stessa analisi condotta a dettaglio settoriale (Graf. 7) e a livello settoriale abbinato ad almeno una azione di sostenibilità ambientale (Graf. 8).

**Grafico 6.** *Imprese con almeno un'azione di sostenibilità per macrosettori di attività economica*

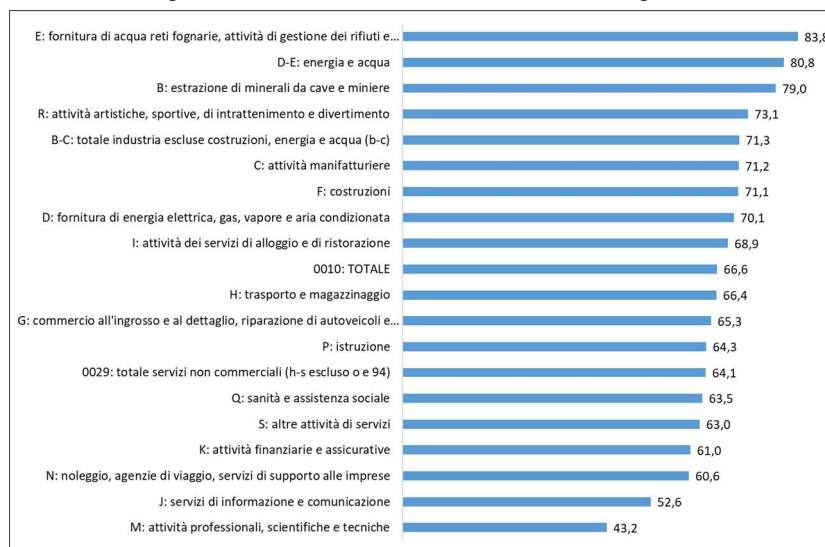


Fonte: Elaborazione su dati Istat

**Grafico 7.** *Imprese con almeno un'azione di sostenibilità ambientale per settori di attività economica - Italia*

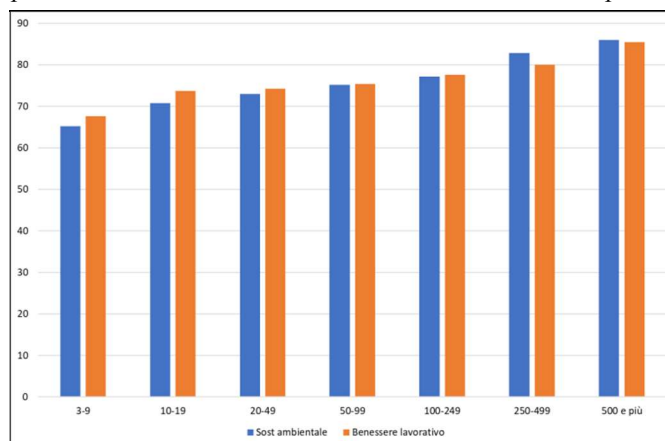


Fonte: Elaborazione su dati Istat

**Grafico 8.** *Imprese con azioni di sostenibilità ambientale per attività economica*

Fonte: Elaborazione su dati Istat

Passando poi all'altra ipotesi potenzialmente discriminante, la dimensione aziendale, considerando che il database fornito permette di suddividere le imprese in relazione alle classi di addetti, è stata condotta l'analoga analisi con riferimento a tale aspetto dimensionale (Graf. 9), da cui si evince come, pur in presenza di una maggiore attenzione su entrambe le variabili al crescere della dimensione aziendale, nelle piccole realtà l'attenzione al benessere lavorativo sia maggiore di quella ambientale.

**Grafico 9.** *Imprese con azioni di sostenibilità o benessere lavorativo per classe di addetti*

Fonte: Elaborazione su dati Istat

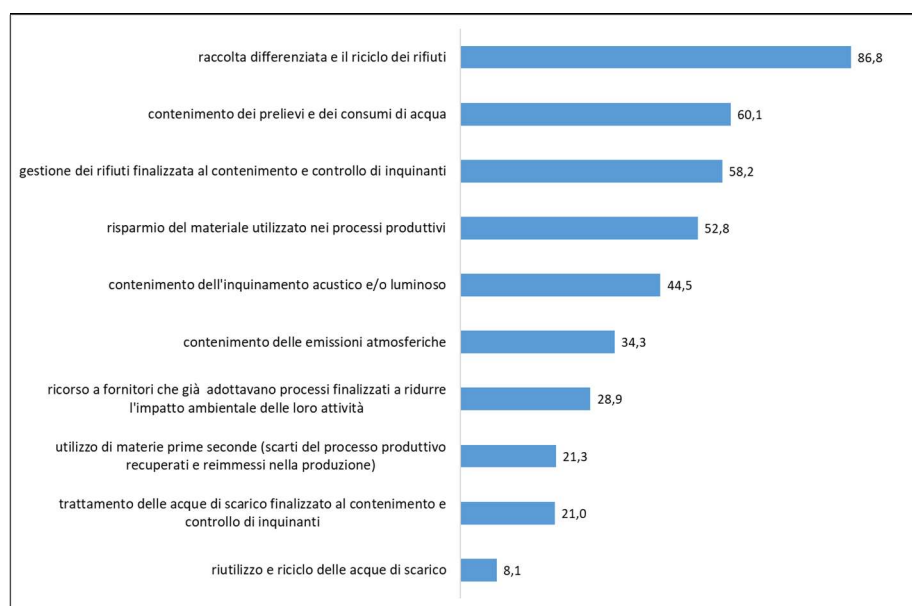
#### 4. Strategie delle imprese nelle azioni di sostenibilità ambientale

Evidenziata la stretta connessione tra attività economica dell'impresa e/o dimensione della stessa da un lato e presenza di azioni di sostenibilità ambientale e di benessere lavorativo dall'altro, nonché uno scarsamente rilevante fattore territoriale, si è voluto verificare in che specifica modalità venissero esplicate concretamente le azioni di sostenibilità ambientale e, a tal fine, si sono in primo luogo esaminate le distribuzioni delle azioni stesse indistintamente dal settore di attività economica.

Le risultanze (Graf. 10) evidenziano come oltre l'85% delle imprese italiane che riducono l'impatto ambientale delle proprie attività, si impegnino con la raccolta differenziata dei rifiuti, così come oltre il 50% delle imprese attuino azioni rivolte al contenimento dei prelievi e dei consumi di acqua, alla gestione dei rifiuti e al risparmio del materiale utilizzato.

Scarsa applicazione da parte delle imprese hanno invece le attività di riutilizzo e riciclo delle acque di scarico (8,1%), trattamento delle stesse (21,0%), riutilizzo degli scarti di produzione (21,3%).

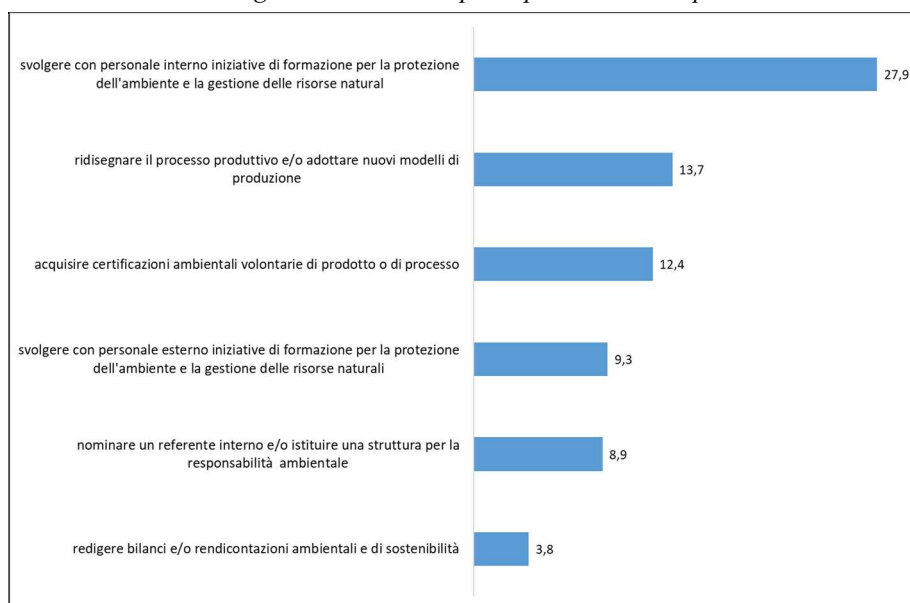
**Grafico 10.** Azioni di sostenibilità in tema di riduzione dei consumi



Fonte: Elaborazione su dati Istat

Con riferimento alle specifiche strategie attuate per ridurre l'impatto ambientale (Graf. 11), spicca da un lato la relativamente forte attenzione verso la formazione del personale (27,9% delle imprese), il ridisegnare il processo produttivo e l'acquisire certificazioni ambientali (13,7%), mentre d'altro lato si evidenzia una scarsa attenzione verso la comunicazione esterna ambientale, infatti solo il 3,8% delle imprese redige bilanci e rendicontazioni ambientali di sostenibilità.

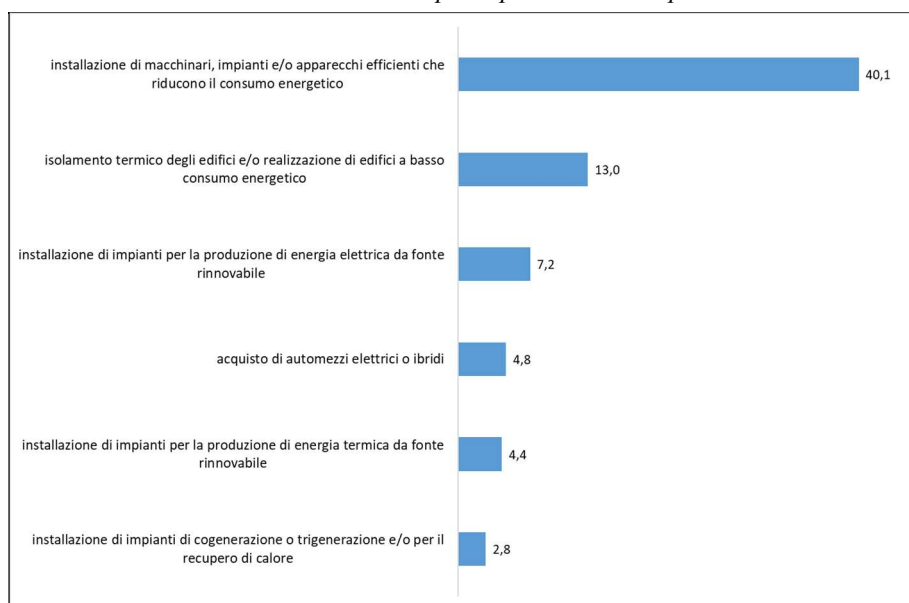
**Grafico 11.** Strategie attuate dalle imprese per ridurre l'impatto ambientale



Fonte: Elaborazione su dati Istat

Spostando invece l'attenzione sulle tipologie di investimenti poste in essere per ridurre l'impatto ambientale (Graf. 12), spiccano le installazioni dei macchinari, impianti e/o apparecchi efficienti energeticamente (40,1%) e le attività di isolamento termico degli edifici (13,0%), mentre scarse risultano essere le installazioni di impianti di produzione di energia elettrica da fonti rinnovabili (4,4%) e quelli per la cogenerazione e/o recupero del calore (2,8%).

**Grafico 12.** *Investimenti delle imprese per ridurre l'impatto ambientale*



*Fonte: Elaborazione su dati Istat*

Passando ora all'analisi condotta in modo disaggregato per macrosettore di attività economica (Figg. 4 - 8), si possono evidenziare azioni di sostenibilità, strategie ed investimenti sovrapponibili per tutti i macrosettori ed azioni invece fortemente caratterizzanti.

Per tutti i settori di attività economica, l'azione rilevante è la raccolta differenziata e il riciclo di rifiuti (oltre l'80% delle imprese); la principale strategia adottata per accrescere la sensibilità ambientale all'interno dell'impresa è sempre la formazione del personale ed il principale investimento è l'installazione di macchinari, impianti e/o apparecchi efficienti.

Per quanto concerne le azioni caratterizzanti nel settore dell'industria, il 67,6% delle imprese attua azioni di gestione dei rifiuti finalizzate al contenimento e controllo di inquinanti e il 67,2% al risparmio del materiale utilizzato nei processi produttivi. Per raggiungere gli obiettivi previsti in tema di sostenibilità ambientale, le imprese di questo macrosettore, in misura maggiore rispetto alle imprese impegnate in altre attività economiche, ridisegnano il processo produttivo (23,3% delle imprese) ed investono nell'isolamento termico degli edifici (14,6%).

Il trattamento delle acque di scarico finalizzato al contenimento e controllo di inquinanti caratterizza, necessariamente, il settore dell'energia ed acqua, che risulta essere il settore più attento all'ottenimento di certificazioni ambientali volontarie di

prodotto o di processo e che investe di più nell'installazione di impianti per la produzione di energia elettrica da fonte rinnovabile.

Il settore delle costruzioni, in termini di investimenti per ridurre l'impatto ambientale, ha decisamente un ruolo centrale: il 22,5% delle imprese investe nell'isolamento termico degli edifici ed oltre il 10% delle imprese nell'installazione di impianti per la produzione di energia elettrica o termica da fonte rinnovabile.

Sono i servizi commerciali e non commerciali ad avere la più alta incidenza di imprese attente al contenimento dei prelievi e dei consumi di acqua rispettivamente per il 59,9% e il 62,2% sebbene non presentino comportamenti caratterizzanti in termini di strategia ed investimenti.

**Fig. 4.** *Comportamenti delle imprese dell'industria in tema di sostenibilità ambientale.*

| AZIONI DI SOSTENIBILITA'                                                           | STRATEGIA                                                                                     | INVESTIMENTI                                                                                               |
|------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------|
| raccolta differenziata e il riciclo dei rifiuti (88,1%)                            | svolgere con personale interno iniziative di formazione per la protezione dell'ambiente (31%) | installazione di macchinari, impianti e/o apparecchi efficienti che riducono il consumo energetico (44,3%) |
| gestione dei rifiuti finalizzata al contenimento e controllo di inquinanti (67,6%) | ridisegnare il processo produttivo e/o adottare nuovi modelli di produzione (23,3%)           | isolamento termico degli edifici e/o realizzazione di edifici a basso consumo energetico (14,6%)           |
| risparmio del materiale utilizzato nei processi produttivi (67,2)                  |                                                                                               |                                                                                                            |

*Fonte: Elaborazione su dati Istat*

**Fig. 5.** *Comportamenti delle imprese dell'energia ed acqua in tema di sostenibilità ambientale.*

| AZIONI DI SOSTENIBILITA'                                                                         | STRATEGIA                                                                                       | INVESTIMENTI                                                                                                                                     |
|--------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------|
| trattamento delle acque di scarico finalizzato al contenimento e controllo di inquinanti (46,4%) | svolgere con personale interno iniziative di formazione per la protezione dell'ambiente (39,7%) | installazione di macchinari, impianti e/o apparecchi efficienti che riducono il consumo energetico (41,2%)                                       |
|                                                                                                  | acquisire certificazioni ambientali volontarie di prodotto o di processo (41%)                  | installazione di impianti per la produzione di energia elettrica da fonte rinnovabile (15,4%)<br>acquisto di automezzi elettrici o ibridi (9,4%) |

*Fonte: Elaborazione su dati Istat*

**Fig. 6.** *Comportamenti delle imprese delle costruzioni in tema di sostenibilità ambientale.*

| AZIONI DI SOSTENIBILITA'                                | STRATEGIA                                                                                       | INVESTIMENTI                                                                                                                                                                                      |
|---------------------------------------------------------|-------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| raccolta differenziata e il riciclo dei rifiuti (85,2%) | svolgere con personale interno iniziative di formazione per la protezione dell'ambiente (29,5%) | installazione di macchinari, impianti e/o apparecchi efficienti che riducono il consumo energetico (31,7%)                                                                                        |
|                                                         | acquisire certificazioni ambientali volontarie di prodotto o di processo (18,2%)                | isolamento termico degli edifici e/o realizzazione di edifici a basso consumo energetico (22,5%)<br>installazione di impianti per la produzione di energia elettrica da fonte rinnovabile (12,2%) |

*Fonte: Elaborazione su dati Istat*

**Fig. 7.** *Comportamenti delle imprese del commercio in tema di sostenibilità ambientale.*

| AZIONI DI SOSTENIBILITA'                                 | STRATEGIA                                                                                       | INVESTIMENTI                                                                                               |
|----------------------------------------------------------|-------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------|
| raccolta differenziata e il riciclo dei rifiuti (88,1%)  | svolgere con personale interno iniziative di formazione per la protezione dell'ambiente (26,7%) | installazione di macchinari, impianti e/o apparecchi efficienti che riducono il consumo energetico (38,2%) |
| contenimento dei prelievi e dei consumi di acqua (59,9%) |                                                                                                 | isolamento termico degli edifici e/o realizzazione di edifici a basso consumo energetico (10,2%)           |

*Fonte: Elaborazione su dati Istat*

**Fig. 8.** *Comportamenti delle imprese dei servizi non commerciali in tema di sostenibilità ambientale.*

| AZIONI DI SOSTENIBILITA'                                 | STRATEGIA                                                                                       | INVESTIMENTI                                                                                               |
|----------------------------------------------------------|-------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------|
| raccolta differenziata e il riciclo dei rifiuti (86,1%)  | svolgere con personale interno iniziative di formazione per la protezione dell'ambiente (26,5%) | installazione di macchinari, impianti e/o apparecchi efficienti che riducono il consumo energetico (41,4%) |
| contenimento dei prelievi e dei consumi di acqua (62,2%) |                                                                                                 | isolamento termico degli edifici e/o realizzazione di edifici a basso consumo energetico (11,3%)           |

*Fonte: Elaborazione su dati Istat*

## **5. Alcune considerazioni conclusive**

Le risultanze illustrate hanno evidenziato una interessante utilizzabilità esplorativa dei dati di rilevanza ambientale dichiarati dalle imprese diffusi nell'ambito dei nuovi censimenti permanenti delle imprese.

Se da un lato tale strumento non permette l'approfondimento di analisi tipico dei file dei microdati, d'altro canto esiste un innegabile valore aggiunto fornito dalla estrema semplicità di utilizzo dell'offerta statistica in oggetto, oltremodo utilizzabile da chiunque senza vincoli autorizzativi, facente parte del Sistema Statistico nazionale, quindi caratterizzato da notevole accuratezza, confrontabilità ed "autorevolezza" del dato.

Nel presente lavoro si è avuto modo di evidenziare come tale nuovo strumento informativo permetta di ottenere, volendo anche con dettaglio territoriale provinciale, mappature significative delle diverse "attenzioni" ambientali poste in essere da parte delle imprese.

Essendo inoltre l'offerta di tali dati disaggregabile anche per macrosettori economici, permette di agevolare non poco l'intervento degli stakeholders a diverso titolo interessati alla valutazione delle politiche ambientali poste in essere in un territorio e quindi anche le modalità potenziali con cui tali attenzioni possano essere migliorate nell'ottica di poter disporre di sistemi economici sempre più sostenibili da un punto di vista di sostenibilità socio-ambientale.

## **Riferimenti bibliografici**

Branchi F, Casciano M.C., Di Francescantonio T., Falorsi P.D., Filiberti S., Marsigliesi G., Sansone U., Santi E., Sanzo R., Siesto G., Zeli A. (2008). *Messa a regime dell'uso dei dati fiscali (modello Unico) per l'integrazione delle mancate risposte totali e la riduzione del numero delle unità campionarie della rilevazione PMI*, Documenti, Istat, n. 4

Istat (2010). *I flussi di materia del sistema socio-economico italiano, tavole dei dati*, 17 maggio 2010 (<https://www.istat.it/it/archivio/2243>)

Casciano M.C., Cirianni A., De Giorgi V., Di Francescantonio T., Mazzilli A., Luzi O., Oropallo F., Rinaldi M., Santi E., Seri G., Siesto G. (2011). *Utilizzo delle fonti amministrative nella rilevazione sulle piccole e medie imprese e sull'esercizio delle arti e professioni*, Istat working papers n. 7/2011

Tudini A., (2015). *Introduzione alla contabilità ambientale, Seminario sui conti economici ambientali*, Istat, Roma

- Istat (2016), *Conti economici regionali*. Comunicato stampa pubblicato il 12 dicembre 2016 (<https://www.istat.it/it/archivio/193916>)
- Boselli C., Brunetti S., Cammarota M., De Giorgi V., D'Urzo A., Ricci M., Pazzini R., Seri G., Siesto G., Virgili L. (2016). *Il processo di diffusione dei dati sulle statistiche strutturali delle imprese (FRAME-SBS): aspetti normativi e metodologici connessi all'ampliamento del dettaglio informativo*, Istat working papers n. 14/ 2016
- Seri G., Ichim D., Luchetti F., Costa S., Nurra A., Mastrostefano V., Salamone S., Pascucci C., Orsini F. (2016). *Integrazione del Frame con altre indagini e fonti amministrative ai fini della produzione di indicatori complessi*, Istat working papers n. 17/2016
- Carucci A.M.M., Fullone F., Vannella G. (2017). *Dai conti economici ai conti satelliti ambientali: Basilicata, un caso di studio*, Metodi ed analisi statistiche, Università degli studi di Bari Aldo Moro
- Istat (2020). *Censimenti permanenti delle imprese 2019: i primi risultati*, 7 febbraio 2020 (<https://www.istat.it/it/archivio/238337>)

# A generalization of the logistic function

Sabrina Diomedè<sup>1\*</sup>, Giovanni Taglialatela<sup>2</sup>

<sup>1</sup> *Dipartimento di Economia, Management e Diritto dell'Impresa*

<sup>2</sup> *Dipartimento di Economia e Finanza,  
Università degli Studi di Bari Aldo Moro*

**Abstract.** In this Note we introduce a generalization of the well-known logistic function, and we analyse some shape properties as well as some differential properties of such generalization.

**Keywords:** Logistic equation; sigmoid equation; generalized logistic function.

## 1. Introduction

The logistic function was introduced in the mid 1830s by the belgian mathematician Pierre Francois Verhulst aiming at modeling population growth; after the reading of an essay by Malthus on the Principle of Population, Verhulst made the hypothesis that population growth should obey a differential equation of the form:

$$(1) \quad P'(t) = \alpha P(t) \left(1 - \frac{P(t)}{K}\right)$$

where  $P(t)$  represents the population size,  $t$  is time,  $K$  stands for the highest value that  $P$  may reach, the so-called *carrying capacity*, and  $\alpha$  is the growth rate.

---

\* Autore corrispondente: [sabrina.diomedè@uniba.it](mailto:sabrina.diomedè@uniba.it)

The solution to (1) is the function

$$(2) \quad P(t) = \frac{K}{1 + \left(\frac{K - P_0}{P_0}\right) e^{-\alpha t}}$$

where  $P_0 = P(0)$ .

The main feature of (2) is that at first the growth of the population is nearly exponential, then it slows down and becomes approximatively linear, finally, as the competition among members of such population diminishes the growth rate,  $P(t)$  ceases to grow and asymptotically reaches  $K$ , which works as the “limiting value” of  $P(t)$ .

By a simple re-arranging of (2) namely by letting

$$\beta = \log\left(\frac{P_0}{K - P_0}\right)$$

it is possible to write (2) as

$$(3) \quad P(t) = \frac{K}{1 + e^{-(\alpha t + \beta)}}, \quad (\alpha > 0, \beta \in \mathbb{R}, t \in \mathbb{R}).$$

The graph of  $P$  is a sigmoig curve, strictly increasing from 0 to  $K$ , symmetric with respect to the point  $\left(t_0, \frac{K}{2}\right)$ , see Fig. 1.

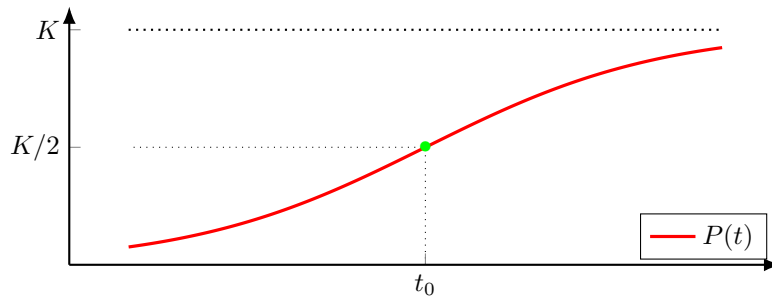


FIGURE 1. The graph of  $P(t)$ .

Functions as (3) are widely used in several fields of social and economic sciences, as well as in statistics, medicine, biology.

For instance, they have been used by Gompertz to describe the growth of tumors (Gompertz (1825)); moreover, a generalization of (3) referred to as the Richards growth-curve has been recently employed for modeling the Covid-19 infection trajectories.

In Economics the logistic functions are used to describe and forecast the diffusion of an innovation, of infrastructures, of energy sources (Ayres (1989), Grübler (1990), Perez (2002)).

In statistics, the logistic function is the distribution function of the logistic distributions, which resembles the normal distribution, though it has heavier tails.

In logistic regression, the logistic function models the way the probability  $\mathbb{P}$  of an event may depend on one or more explanatory variables.

In this Note we shall consider (3) taking into account a more general but suitable map  $g$  instead of an affine function.

We study many qualitative properties of such generalization as the asymptotic behavior, the critical and extrema points, the inflection point, but we also analyze the relationship between the shape properties of  $g$  and the analogous properties held by  $P(t)$ . In the last Section, we show how to obtain the generalized logistic function as a solution to a suitable differential equation which, in turn, generalizes (1).

Thus, let us consider a map  $g: I \rightarrow \mathbb{R}$ , where  $I = ]a, b[$  is some interval of  $\mathbb{R}$ , with  $a \in \{-\infty\} \cup \mathbb{R}$ ,  $b \in \mathbb{R} \cup \{+\infty\}$  and  $a < b$ ;  $g$  is assumed to be differentiable and to satisfy the following coercivity condition:

$$(4) \quad \lim_{x \searrow a} g(t) = -\infty \quad \lim_{x \nearrow b} g(t) = +\infty.$$

Hence, the generalized logistic function turns out to be

$$(5) \quad P_g(t) = \frac{K}{1 + e^{-g(t)}}, \quad (t \in I),$$

where  $K$  keeps the same meaning as in (2).

*Examples 1.* Some examples of maps behaving as required for  $g$  are

- (a)  $g(t) = t^{2k+1}$ ,  $k \in \mathbb{N}$ , with  $I = ]-\infty, +\infty[$ ;
- (b)  $g(t) = \begin{cases} \frac{|t|}{t} \log(1 + t^2) & \text{for } t \neq 0, \\ 0 & \text{for } t = 0, \end{cases}$  and  $I = ]-\infty, +\infty[$ ;
- (c)  $g(t) = t^3 - t$ ,  $I = ]-\infty, +\infty[$ ;
- (d)  $g(t) = t - \frac{1}{t}$ ,  $I = ]0, +\infty[$ ;
- (e)  $g(t) = t e^{t^2}$ ,  $I = ]-\infty, +\infty[$ .

In (Mininni et al. 2021), the authors have approximated the Call function of the Black and Scholes model by means of a function  $P_g$ , where  $g$  is of type (d).

In Figure 2 we present some of the maps  $g$  listed above, together with their corresponding function  $P_g$ .

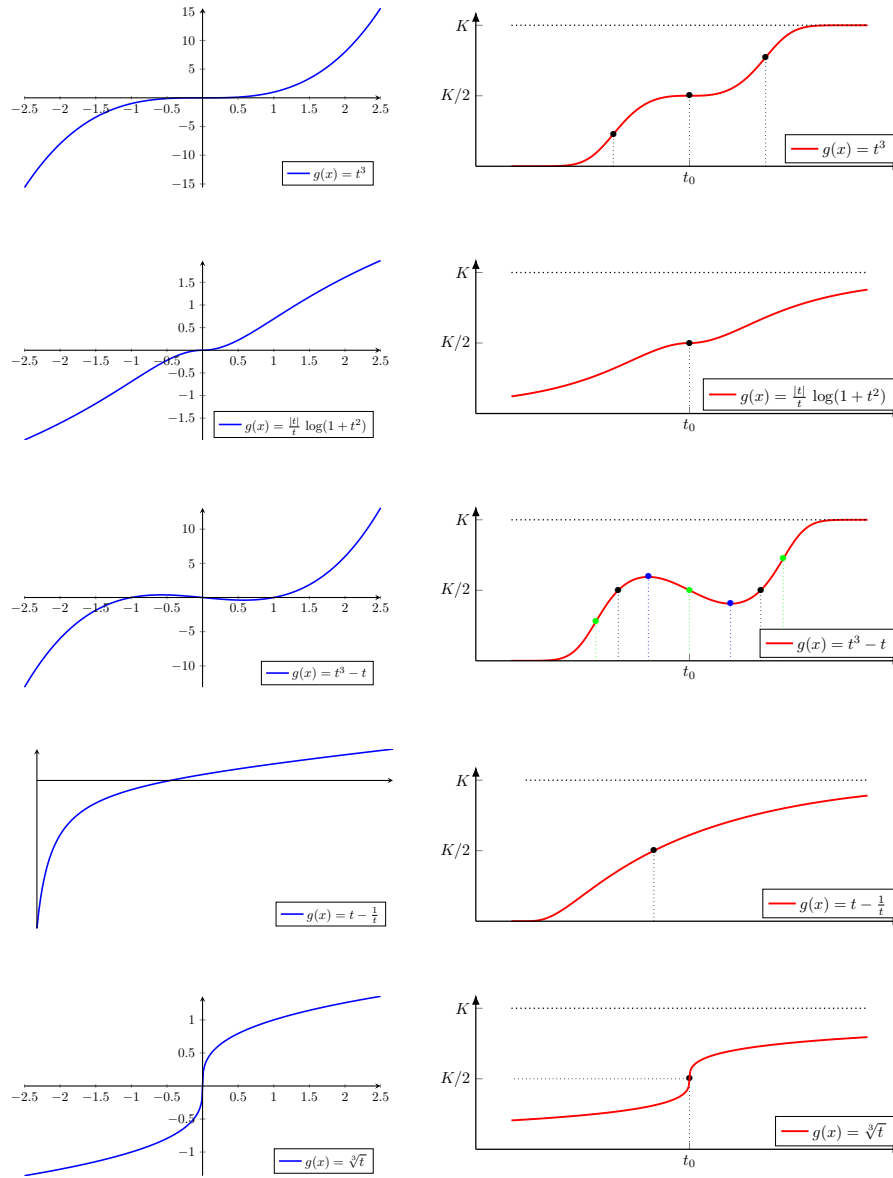


FIGURE 2.

## 2. Main properties

In what follows we discuss several properties of  $P_g$ .

### 2.1. Asymptotic behavior

**Proposition 1.** *Under hypothesis (4) the function  $P_g$  verifies the following*

$$\lim_{t \rightarrow a^-} P_g(t) = 0 \quad \lim_{t \rightarrow b^+} P_g(t) = K .$$

### 2.2. Symmetry

Under a suitable assumption on  $g$  it turns out that the graph of  $P_g$  is symmetric w.r.t. to some point  $\left(t_0, \frac{K}{2}\right)$ .

**Proposition 2.** *Suppose  $g: \mathbb{R} \rightarrow \mathbb{R}$  satisfies*

$$g(t_0 + t) = -g(t_0 - t)$$

*for some  $t_0 \in \mathbb{R}$ , then*

$$(6) \quad P_g(t_0 - t) = K - P_g(t_0 + t) .$$

*Proof.* We have

$$\begin{aligned} P_g(t_0 - t) &= \frac{K}{1 + e^{-g(t_0 - t)}} = \frac{K}{1 + e^{g(t_0 + t)}} \\ &= \frac{K e^{-g(t_0 + t)}}{1 + e^{-g(t_0 + t)}} = K \left[ 1 - \frac{1}{1 + e^{-g(t_0 + t)}} \right] \\ &= K - P_g(t_0 + t) , \end{aligned}$$

which proves (6). □

Note that, by setting  $t = 0$  in (6), we get  $P(t_0) = \frac{K}{2}$ .

### 2.3. Critical points

By differentiation, we may derive some useful qualitative properties for  $P_g$ .

As

$$(7) \quad P'_g(t) = -\frac{K D(e^{-g(t)})}{(1 + e^{-g(t)})^2} = \frac{K e^{-g(t)} g'(t)}{(1 + e^{-g(t)})^2}, \quad (t \in I)$$

We deduce a critical-point preservation property as stated below

**Proposition 3.** *The following statements are equivalent:*

- (a)  $t_0$  is a critical point for  $g$ ;
- (b)  $t_0$  is a critical point for  $P_g$ .

The proof directly follows from (7).

Observe that, by (7), the sign of  $P'_g$  depends on that of  $g'$ ; more precisely, the intervals upon which  $g'$  is positive (resp. negative) are the same intervals in which  $P'_g$  is positive (resp. negative).

Thus we may infer the following

**Proposition 4.** *Assume  $g$  is continuously differentiable, then the following statements are equivalent:*

- (a)  $t_0$  is a strict local minimum (resp. maximum) point for  $g$ ;
- (b)  $t_0$  is a strict local minimum (resp. maximum) point for  $P_g$ .

#### 2.4. Elasticity

We now pass to compute the elasticity function  $\mathcal{E}_g$  for  $P_g$ , assuming that  $g$  is continuously differentiable.

By definition, we have

$$\mathcal{E}_g(t) = \frac{t P'_g(t)}{P_g(t)} = \frac{t \frac{K e^{-g(t)} g'(t)}{(1 + e^{-g(t)})^2}}{\frac{K}{1 + e^{-g(t)}}} = \frac{t e^{-g(t)} g'(t)}{1 + e^{-g(t)}},$$

which may also be written as

$$\mathcal{E}_g(t) = -\frac{t D(e^{-g(t)})}{1 + e^{-g(t)}}.$$

Under a suitable assumption on  $g$ , the elasticity has a prescribed asymptotic behavior as shown in the following

**Proposition 5.** *Let  $g$  be continuously differentiable and assume that it satisfies (4). If there exists  $C > 0$  such that*

$$(8) \quad t |g'(t)| \leq C g(t)$$

for  $t$  large enough, then

$$\lim_{t \rightarrow +\infty} \mathcal{E}_g(t) = 0$$

*Proof.* As

$$|\mathcal{E}_g(t)| \leq t e^{-g(t)} |g'(t)|,$$

because of (8), we get

$$|\mathcal{E}_g(t)| \leq C \frac{g(t)}{e^{g(t)}},$$

and the conclusion follows.  $\square$

*Remark 1.* If  $g$  is a polynomial, then (8) occurs.

## 2.5. Inflection points

We notice that if a suitable condition occurs for  $g$ , then the inflection points are transferred from the function  $g$  to the function  $P_g$ .

Preliminarily we compute the second and third-order derivatives of  $P_g$ :

$$(9) \quad P_g''(t) = \frac{K e^{-g(t)}}{(1 + e^{-g(t)})^3} \left[ (1 + e^{-g(t)}) g''(t) - (1 - e^{-g(t)}) [g'(t)]^2 \right]$$

$$(10) \quad P_g'''(t) = \frac{K e^{-g(t)}}{(1 + e^{-g(t)})^4} \left[ (1 + e^{-g(t)})^2 g'''(t) + (1 - 4e^{-g(t)} + e^{-2g(t)}) [g'(t)]^2 - 3(1 - e^{-2g(t)}) g'(t) g''(t) \right].$$

hence we obtain the following

**Proposition 6.** *Assume  $g$  is three times differentiable at some point  $t_0$ , and*

$$(11) \quad g(t_0) = g'(t_0) = 0.$$

*Then the following statements are equivalent:*

- (a)  $g''(t_0) = 0$  and  $g'''(t_0) \neq 0$ ;
- (b)  $P_g''(t_0) = 0$  and  $P_g'''(t_0) \neq 0$ .

*Proof.* It's clear that if (11) holds true, then (9) and (10) reduces to

$$(9') \quad P_g''(t) = \frac{K e^{-g(t)}}{(1 + e^{-g(t)})^2} g''(t)$$

$$(10') \quad P_g'''(t) = \frac{K e^{-g(t)}}{(1 + e^{-g(t)})^2} g'''(t)$$

and the conclusion follows.  $\square$

## 2.6. Higher order derivatives

In what follows we establish a relationship between the higher order derivatives of  $g$  and those of  $P_g$ .

**Proposition 7.** *Assume that  $g$  is  $n$ -times differentiable at  $t_0$ ; then the following statements are equivalent:*

- (a)  $g^{(j)}(t_0) = 0$  for  $j = 1, \dots, n-1$  and  $g^{(n)}(t_0) > 0$  (resp.  $g^{(n)}(t_0) < 0$ );
- (b)  $P_g^{(j)}(t_0) = 0$  for  $j = 1, \dots, n-1$  and  $P_g^{(n)}(t_0) > 0$  (resp.  $P_g^{(n)}(t_0) < 0$ ).

*Proof.* As  $P_g(t) = P(g(t))$ , in order to compute the higher order derivatives of  $P_g$  we recall the Faa di Bruno's formula: let  $f$  and  $g$  be  $n$ -times differentiable functions, then

$$(12) \quad D^j[f(g(t))] = \sum_{\nu=1}^j (D^\nu f)(g(t)) \sum_{h_1+\dots+h_\nu=j} D^{h_1}g(t) \cdots D^{h_\nu}g(t),$$

for  $j = 1, \dots, n$ , where we use the notation

$$D^j\varphi(t) = \frac{1}{j!} \frac{d^j}{dt^j} \varphi(t).$$

Now, if all the derivatives of  $g$  of order  $\leq n-1$  vanish at  $t_0$ , then (12) gives

$$D^n P_g(g)(t_0) = (DP)(g(t_0)) D^n g(t_0),$$

and the conclusion holds true.

We prove by induction the converse statement. The case  $n = 1$  is straightforward (cf. Proposition 3). Assuming that the derivatives of order  $< j$  of  $P_g$  vanish at  $t_0$ , by induction hypothesis the derivatives of order  $< j$  of  $g$  vanish at  $t_0$ , hence, by Faa di Bruno's formula, we have

$$D^j[P(g(t))] = P'(g(t)) D^j g(t).$$

As  $P'$  never vanishes, we deduce that if  $D^j[P(g(t_0))] = 0$  (resp.  $> 0$ , resp.  $< 0$ ) then  $D^j g(t_0) = 0$  (resp.  $> 0$ , resp.  $< 0$ ), and induction follows.  $\square$

### 3. Differential equation

It is possible to derive  $P_g(t)$  as a solution to the following differential equation

$$(13) \quad y'(t) = r(t) y(t) \left(1 - \frac{y(t)}{K}\right),$$

where  $r(t) = g'(t)$ .

*Remark 2.* We underline that in the differential equation related to the classical logistic function (3) the term  $r(t)$  is a constant function; more precisely it represents the constant growth parameter.

To solve (13) we remark at first that

$$(14) \quad y(t) \equiv 0 \quad \text{and} \quad y(t) \equiv K$$

are the (only) equilibria solutions. Thanks to the Cauchy-Lipschitz existence and uniqueness Theorem, any other solution of (13) verifies one and only one of the condition

$$(15) \quad y(t) < 0 \quad \text{for any } t \in \mathbb{R},$$

$$(16) \quad 0 < y(t) < K \quad \text{for any } t \in \mathbb{R},$$

$$(17) \quad y(t) > K \quad \text{for any } t \in \mathbb{R}.$$

Now we use the method of separation of variables to find the other solutions: moving all terms containing  $y$  to the left-hand side of the equation we get

$$\frac{y'(t)}{y(t) \left(1 - \frac{y(t)}{K}\right)} = r(t).$$

Now, as

$$\frac{1}{y\left(1 - \frac{y}{K}\right)} = \frac{1}{y} + \frac{1}{K - y}$$

we get

$$\begin{aligned} \int \frac{1}{y\left(1 - \frac{y}{K}\right)} dy &= \int \left[ \frac{1}{y} + \frac{1}{K - y} \right] dy \\ &= \left[ \log |y| - \log |K - y| \right] + \text{const.} \\ &= \log \left| \frac{y}{K - y} \right| + \text{const.} \end{aligned}$$

Equation (13) becomes:

$$\frac{d}{dt} \log \left| \frac{y(t)}{K - y(t)} \right| = g'(t)$$

hence

$$(18) \quad \log \left| \frac{y(t)}{K - y(t)} \right| = g(t) + C,$$

where  $C$  is some constant which can be determined by the initial conditions.

In order to keep an empirical significance, we just consider solutions with initial data  $y(0) = y_0 \in ]0, K[$ . As the solution cannot intersect the equilibria solutions (cf. (14)) it should verify (16), and we can drop the modulo in (18). Thus we obtain

$$\log \left[ \frac{y(t)}{K - y(t)} \right] = g(t) + C,$$

from which

$$y(t) = \frac{K e^{g(t)+C}}{1 + e^{g(t)+C}} = \frac{K}{1 + e^{-g(t)-C}}.$$

Hence  $P_g$  is a solution to (13) with

$$C = \log \left( \frac{y_0}{K - y_0} \right) - g(0),$$

or, equivalently, with the initial condition

$$(19) \quad y(0) = y_0 \equiv \frac{K}{1 + e^{-g(0)}}.$$

## References

- Ayres, R.U. (1990). Technological transformations and long waves. Part I, *Technological Forecasting and Social Change*, 37(1), 1–37, (<https://www.sciencedirect.com/science/article/pii/0040162590900573>)
- Gompertz, B. (1825). On the nature of the function expressive of the law of human mortality, and on a new mode of determining the value of life contingencies. In a letter to Francis Baily. *Philosophical Transactions of the Royal Society of London*, 115: 513–585. 115513-583 (<http://doi.org/10.1098/rstl.1825.0026>).
- Grübler, A. (1990). *The Rise and Fall of Infrastructures: Dynamics of Evolution and Technological Change in Transport*. Heidelberg: Physica-Verlag. ISBN 3-7908-0479-7 (<http://pure.iiasa.ac.at/id/eprint/3351>).
- Lee S.Y.; Lei, B.; Mallick, B. (2020). Estimation of COVID-19 spread curves integrating global data and borrowing information. *PLoS ONE* 15(7): e0236860. (<https://doi.org/10.1371/journal.pone.0236860>).
- Marchetti, C. (1988). Kondratiev Revisited – After One Kondratiev Cycle. International Conference on Regularity of Scientific-Technical Progress and Long-Term Tendencies of Economic Development, Novosibirsk, Akademgorodok, USSR, March 14-19, 1988 (<https://www.cesaremarchetti.org/archive/scan/MARCHETTI-037.pdf>).
- Marchetti, C. (1996). Pervasive Long Waves: Is Society Cyclotymic. Conference “Offensiv zu Arbeitsplätzen: Weltmärkte 2010”, Cologne, September 14-15, 1996 (<https://www.cesaremarchetti.org/archive/scan/MARCHETTI-038.pdf>).
- Mininni, M., Orlando, G. and Taglialatela, G. (2021). Challenges in approximating the Black and Scholes call formula with hyperbolic tangents. *Decisions Econ Finan* 44, 73–100.
- Perez, C. (2002). *Technological Revolutions and Financial Capital: The Dynamics of Bubbles and Golden Ages*. UK: Edward Elgar Publishing Limited.

- Verhulst, P. F. (1838). Notice sur la loi que la population suit dans son accroissement. *Correspondance mathématique et physique*, 10: 113–121.
- Verhulst, P. F. (1841). *Traité élémentaire des fonctions elliptiques : ouvrage destiné à faire suite aux traités élémentaires de calcul intégral*. Bruxelles, Hayez.
- Verhulst, P. F. (1845). Recherches mathématiques sur la loi d'accroissement de la population. *Nouveaux Mémoires de l'Académie Royale des Sciences et Belles-Lettres de Bruxelles*, 18: 1–42.
- Verhulst, P. F. (1847). Deuxième mémoire sur la loi d'accroissement de la population. *Mémoires de l'Académie Royale des Sciences, des Lettres et des Beaux-Arts de Belgique*, 20: 1–32.

# Crisi finanziaria e politiche di bilancio. Alcune evidenze empiriche

Anna Lucia Muserra<sup>1</sup>, Francesco Grimaldi<sup>1\*</sup>

<sup>1</sup>Dipartimento di Economia e Finanza, Università degli Studi di Bari Aldo Moro

**Riassunto:** L'obiettivo della ricerca è quello di analizzare la relazione tra la crisi finanziaria del 2008 e le politiche di bilancio (o politiche di manipolazione del reddito o *earnings management*) delle imprese. In particolare, questo lavoro si propone di indagare se, nel contesto italiano, le condizioni macroeconomiche conseguenti alla crisi finanziaria del 2008 abbiano rappresentato o meno un incentivo a porre in essere politiche di manipolazione del reddito. La ricerca viene condotta utilizzando due proxies del Modello di Beneish (2013), per la sua capacità di identificare, sebbene su base probabilistica, le società che potenzialmente adottano politiche di bilancio. I risultati conseguiti consentono una migliore conoscenza del fenomeno che si estende dal periodo pre-crisi a quello post-crisi e suggeriscono una presenza complessivamente bassa di società a rischio di manipolazione per tutto il periodo in esame.

**Keywords:** politiche di bilancio; crisi finanziaria; società quotate Italiane; modello di Beneish.

## 1. Introduzione

I numerosi scandali societari conseguenti all'opportunismo manageriale e/o alla realizzazione di frodi contabili da parte dei redattori del bilancio verificatisi in numerosi paesi negli ultimi decenni, hanno riportato al centro dell'attenzione di accademici, professionisti, organi di regolamentazione e *policy makers* il ruolo dell'informativa esterna d'impresa, con particolare riferimento al grado di qualità (*earnings quality*) della stessa informativa (Grimaldi e Muserra, 2017).

---

\* Autore corrispondente: francesco.grimaldi1@uniba.it

La “ritrovata” centralità dell’informativa esterna d’impresa è riconducibile alle sue caratteristiche, ovvero, essere allo stesso tempo l’input per la pianificazione e la programmazione delle risorse e l’output funzionale all’analisi della situazione economico-finanziaria dell’azienda da parte di numerosi soggetti portatori di interessi (Ferramosca, 2018).

Tuttavia, benché le imprese dispongano di molteplici canali attraverso i quali comunicare l’informativa esterna, non esiste allo stato attuale, nell’ambito della vasta letteratura in tema di *disclosure*, una definizione univoca di qualità della comunicazione. Di traslato, pertanto, non esiste una modalità di misurazione universale (Grimaldi, 2015). Difatti, la letteratura si è sviluppata attorno ad un’accezione ristretta di qualità della comunicazione, avente ad oggetto la capacità dei valori contabili di fornire informazioni funzionali ai fini decisionali, in virtù della capacità degli stessi di evidenziare sia la performance conseguita sia quella prospettica (Francis et al., 2003).

Tale filone di studi incentra l’attenzione sulla cosiddetta *earnings quality*, ossia sulla qualità del risultato aziendale, che costituisce un riferimento per i fruitori dell’informativa di bilancio. Difatti, poiché i destinatari dell’informativa sono numerosi e molteplici sono le aspettative e le modalità di utilizzo della stessa, secondo tale prospettiva di ricerca, la qualità degli utili è definita sulla base della sussistenza contemporanea di due dimensioni, basate su attributi *market - based* (*value relevance*, tempestività e *conservatism*) e di tipo contabile (qualità degli *accruals*, sostenibilità, prevedibilità, variabilità).

La rafforzata rilevanza della relazione tra la qualità dell’informativa di bilancio e le politiche di bilancio è stata evidenziata, nell’ultimo ventennio, anche dal verificarsi di fenomeni di carattere politico-istituzionale e macroeconomico, non ultima la deflagrazione della crisi finanziaria del “2008”.

La crescente letteratura non è concorde nell’individuare la direzione della relazione che lega la crisi finanziaria del 2008 e il ruolo, l’attendibilità e la rilevanza dell’informativa di bilancio.

Negli ultimi decenni, numerosi studiosi hanno analizzato le determinanti e le conseguenze delle politiche di bilancio (si veda per tutti, Ronen e Yaari, 2008) ma solo con la deflagrazione della crisi del “2008”, un numero crescente di studi (Balasumbayyan et al., 2013; Bornemann et al., 2012; Cimini, 2015; Filip e Raffournier, 2014; Mari et al., 2016; Paolone et al., 2015) hanno esaminato come può modificarsi la qualità degli utili in un periodo di instabilità economica e/o finanziaria e, soprattutto, la misura in cui l’instabilità, una volta in essere, può influenzare le scelte contabili dei redattori del bilancio, con particolare riferimento alle politi-

che di manipolazione del reddito (Trombetta e Imperatori, 2014). Difatti, sino agli studi appena citati, l'impostazione prevalente della letteratura era concorde nel ritenere che le condizioni macroeconomiche non rappresentassero un incentivo a porre in essere politiche di manipolazione del reddito (Filip e Raffournier, 2014).

L'obiettivo di questa ricerca è quello di indagare sulla relazione tra crisi finanziarie e politiche di *bilancio* e, in particolare, di esaminare se, nel contesto italiano, con particolare riferimento alle società quotate non finanziarie, la particolare situazione macroeconomica conseguente alla deflagrazione della crisi finanziaria innestatasi nel quarto trimestre 2007 negli USA, abbia rappresentato o meno un incentivo ad attuare politiche di *earnings management* (E.M.).

Nell'ambito della corposa letteratura di matrice accademica e professionale sul tema, molteplici e multiformi sono le definizioni proposte in merito al concetto dell'E.M.. Considerando alcune delle definizioni più utilizzate è possibile rilevare il loro comune denominatore nel ruolo discrezionale dei manager esercitabile, secondo differenti stili (Dechow e Sinner, 2000), nel processo di redazione dell'informativa di bilancio e nei connessi processi di comunicazione.

In tale direzione, pare utile osservare che l'*earnings management*, quale fenomeno che incide sulla qualità del reddito d'esercizio, tende ad assumere una connotazione negativa. Da tale prospettiva discende che i termini *earnings quality* ed *earnings management* vengano sovente accomunati, essendo quest'ultimo considerato una misura inversa del primo (Grimaldi, 2015).

Diversamente da quanto osservato nel processo di definizione del concetto stesso di E.M., la letteratura (Prencipe, 2006) appare sostanzialmente concorde nel ritenere che gli obiettivi, conseguibili in via disgiunta o congiunta, sottostanti alle politiche di *earnings management*, sono riconducibili: a) alla conferma o all'ottenimento di determinati livelli di reddito; b) al soddisfacimento delle aspettative degli investitori esterni e/o degli analisti; c) all'ottenimento di benefici economici privati; d) all'ottenimento di determinati livelli di utile e i connessi benefici economici privati.

Le politiche di bilancio possono essere realizzate sia con modalità connesse alla rilevazione, alla valutazione e all'iscrizione in bilancio degli accadimenti aziendali, c.d. politiche di manipolazione valutative o indirette (McVay, 2006), sia con modalità connesse direttamente alle operazioni aziendali, c.d. politiche di manipolazione gestionali o dirette (Roychowdhury, 2006).

Parimenti rilevanti nel processo di determinazione del livello di attendibilità dell'informativa di bilancio appaiono le variabili di contesto. Difatti, in letteratura sono state sostanzialmente individuate quattro macro-variabili di contesto in grado

di favorire o di ostacolare l'implementazione di politiche di bilancio, e in particolare: a) il sistema normativo/regolamentare e di vigilanza (Leuz et al., 2003); b) il mercato dei capitali (Beatty et al., 1999); c) il sistema di corporate governance (Beasley, 1996) e dei controlli (Hirst, 1994); d) il fattore etico e culturale (Pini, 1991).

La letteratura (Dechow et al., 1995; Healy e Wahlen, 1999) appare concorde nella classificazione degli incentivi che inducono all'attuazione di politiche di manipolazione in due macro-tipi: incentivi di mercato (*capital market incentive*) e incentivi contrattuali (*contractual incentive*).

Tuttavia, nonostante i diversi ambiti di ricerca che si sono sviluppati intorno ai temi delle politiche di bilancio, dall'analisi della letteratura sono emersi solo alcuni studi che, con risultati contrastanti, hanno esaminato - in contesti antecedenti rispetto alla crisi finanziaria del 2008 - la relazione tra ciclo economico e qualità del reddito (Johnson, 1999; Jenkins et. al., 2009).

Difatti, la "centralità" delle politiche di E.M. è cresciuta con l'inizio della crisi finanziaria, che ha portato ad un ulteriore calo della fiducia nei confronti del bilancio come strumento atto a rappresentare la rilevanza del valore dei dati finanziari riportati e ha attirato l'attenzione degli investitori sull'affidabilità dei valori di bilancio e sulle potenziali politiche di manipolazione del reddito. In tale direzione, come osservato da Filip e Raffournier (2014) è presumibile che i cambiamenti macroeconomici, indotti dalla crisi finanziaria, abbiano un impatto sulle politiche di manipolazione del reddito in termini di direzione e di intensità. Come osservato da Trombetta e Imperatore (2014) c'è ancora molto da apprendere e comprendere circa gli effetti dei fattori macroeconomici sulle scelte che implicano discrezionalità contabile, e la crisi finanziaria del 2008 può certamente rappresentare uno di tali fattori.

Pertanto, considerato che i risultati delle analisi condotte sulla relazione tra la crisi finanziaria del 2008 e le politiche di manipolazione del reddito mostrano alternativamente, da un lato, una tendenza all'aumento delle politiche contabili (Bornemann et al., 2012; Balasumbrayanyan et al., 2013; Mari et al., 2016) e dall'altro, una tendenza alla riduzione delle politiche di manipolazione del reddito in periodi di crisi (Filip e Raffournier, 2014; Cimini, 2015; Paolone et al., 2015), il presente contributo si propone di indagare se, nel contesto italiano, le precarie condizioni macroeconomiche/finanziarie e le conseguenti difficoltà subite dalle società quotate abbiano costituito o meno un incentivo ad attuare politiche di bilancio. In altri termini, è stata effettuata un'analisi non direzionale per verificare se il numero di aziende classificate come manipolatrici riscontrate nel periodo di crisi differisca

da quanto registrato nel periodo pre-crisi. Inoltre, a differenza di quanto osservato in precedenti ricerche (Franceschetti, 2018), oltre al periodo precedente e concomitante alla crisi finanziaria, la suddetta verifica è estesa anche al periodo post-crisi.

Sulla base di considerazioni precedenti, sono state sviluppate le seguenti ipotesi alternative non direzionali:

*la crisi finanziaria non influenza la politica di bilancio, in termini di numero di società classificate come manipolatrici;*

*o,*

*la crisi finanziaria influenza la politica di bilancio, in termini di maggiore/minore numero di società classificate come manipolatrici.*

Per misurare l'impatto della crisi finanziaria sulla manipolazione contabile, è stato analizzato un campione di 75 società italiane quotate non finanziarie operative nell'arco temporale 2005-2016, che consente una visione olistica del fenomeno che si estende dal periodo precedente (2005-2008) a quello successivo alla crisi (2013-2016). Tra i diversi metodi di misurazione potenzialmente utilizzabili è stato scelto, seppur con due diverse proxies, il modello Beneish (Beneish, 1999; 2001; 2013; Beneish e Vargus, 2002; Beneish et al., 2013), in quanto ben noto e diffuso sia nel mondo accademico che in quello aziendale, per la sua capacità di identificare, sebbene su base probabilistica, le aziende che potenzialmente adottano una politica di manipolazione del reddito.

I risultati dello studio suggeriscono che vi è una presenza complessivamente bassa di società a rischio di manipolazione durante tutto il periodo in esame (2005-2016); tuttavia, il numero più consistente si registra nel periodo pre-crisi (2005-2008).

Il presente contributo è organizzato come segue: nel secondo paragrafo viene sviluppato l'impianto metodologico, nell'ambito del quale viene spiegato il processo di selezione del campione, di raccolta dei dati e il modello di riferimento utilizzato; nel terzo paragrafo si presentano e si discutono i risultati empirici dell'analisi condotta. Il quarto paragrafo, infine, espone le conclusioni della ricerca segnalando i limiti della stessa nonché le possibili direzioni per la ricerca futura.

## 2. Materiali e metodi

Il presente paragrafo è dedicato alla rappresentazione dei materiali e dei metodi di ricerca. L'analisi è stata condotta su un campione di società quotate sul mercato borsistico italiano con l'obiettivo di verificare, alla luce di quanto evidenziato dalla letteratura e alla luce delle problematiche indagate, la sussistenza delle ipotesi poste alla base della ricerca.

Pertanto, in relazione a quanto premesso, si procede alla definizione e all'esposizione delle scelte metodologiche effettuate, con specifico riguardo: a) alla identificazione del modello statistico prescelto, secondo due differenti *proxies* del Modello di Beneish, per l'individuazione di potenziali politiche manipolatrici del reddito; b) alla selezione e alla definizione del campione oggetto di analisi e alle fonti utilizzate per il reperimento dei dati funzionali per il calcolo delle variabili contemplate nelle due differenti *proxies* del modello statistico prescelto; c) alla definizione del periodo considerato e del *cut-off* (punto di inizio della crisi finanziaria).

### 2.1 Metodi di analisi utilizzati

Per stimare la probabilità di manipolazione nelle imprese del campione, vengono utilizzati il modello non ponderato di Beneish (1999), come rivisto nel 2013 (Beneish et al., 2013) e il riadattamento del modello di Beneish alla realtà italiana (Giunta et al., 2014).

Il modello di Beneish del 1999 nasce da un'analisi condotta dallo stesso studioso che, analizzando i dati di un campione di imprese note alla Securities and Exchange Commission (SEC) come trasgressori dei corretti principi contabili (Mafrolla, 2009) in quanto avevano chiaramente manipolato i propri bilanci, ne ha prima identificato le caratteristiche e successivamente ha cercato le differenze quantitative all'interno di un campione di aziende non manipolatrici.

Sulla base di questo confronto, Beneish ha sviluppato un modello probabilistico (*Manipulation Score* o *M-Score*) che adotta diverse metriche per identificare l'entità di una manipolazione o, quanto meno, le precondizioni che potrebbero spingere le aziende a impegnarsi in questa attività, scegliendo variabili che: a) misurano le prestazioni economico-finanziarie delle imprese; b) esprimono incentivi/possibilità di violare i principi contabili da parte del management.

Il risultato è stato un modello non ponderato e probabilistico in grado di stimare la probabilità tendenziale di manipolazione dei risultati d'esercizio nel periodo che include otto variabili: *Days Sales in Receivables Index* (DSRI); *Gross Margin In-*

*dex* (GMI); *Asset Quality Index* (AQI); *Sales Growth Index* (SGI); *Depreciation Index* (DEPI); *Sales, General and Administrative Expenses Index* (SGAI); *Leverage Index* (LVGI); *Total Accruals to Total Asset* (TATA).

La formula del *Manipulation Score*, frutto delle stime di Beneish, deriva dalla seguente relazione lineare:

$$M\text{-Score}(8) = -4,840 + 0,920 \cdot DSRI + 0,528 \cdot GMI + 0,404 \cdot AQI + 0,892 \cdot SGI + 0,115 \cdot DEPI - 0,172 \cdot SGAI - 0,327 \cdot LVGI + 4,679 \cdot TATA \quad [1]$$

La composizione e il significato di ciascuna variabile sono indicati come segue:

1. *Indice del tempo medio di incasso* (DSRI) è dato dal rapporto tra il tempo medio di incasso nell'anno (n) e la misura corrispondente nell'anno precedente (n-1). Il tempo medio di incasso è calcolato come rapporto tra crediti e vendite. Difatti, un aumento sproporzionato del tempo medio di incasso può essere indicativo di inflazione dei ricavi o può riflettere un cambiamento nella politica commerciale a supporto delle vendite. In altri termini, un notevole aumento del tempo medio di incasso è associato ad una maggiore probabilità di manipolazione degli utili.

2. *Indice del margine lordo* (GMI) è dato dal rapporto tra il margine lordo dell'anno precedente (n-1) e il margine lordo dell'anno in corso (n). Il margine lordo è dato dal rapporto tra la differenza delle vendite e il costo del venduto e le vendite (al denominatore). Un GMI maggiore di 1 indica un peggioramento del margine lordo e rappresenta un segnale negativo sulle prospettive delle imprese.

3. *Indice della qualità degli investimenti* (AQI) è dato dal rapporto tra le attività non correnti (diverse dalle immobilizzazioni materiali) e il totale dell'attivo dell'anno in corso (n) e le attività non correnti (diverse dalle immobilizzazioni materiali: DPI) e il totale dell'attivo dell'anno precedente (n-1). Questo indice mette in evidenza la proporzione delle attività totali per le quali i benefici futuri sono potenzialmente meno sicuri. Difatti, un AQI maggiore di 1 indica che l'azienda ha probabilmente aumentato, ad esempio, il volume di capitalizzazione delle attività intangibili che potrebbero incidere sulla qualità del totale degli asset.

4. *Indice di Crescita delle Vendite* (SGI) è dato dal rapporto tra le vendite dell'anno (n) e quelle dell'anno precedente (n-1). Un SGI maggiore di 1 indica che le vendite sono in crescita rispetto all'anno precedente. Tuttavia, seppur un aumento delle vendite non implica che vi sia una potenziale manipolazione, le imprese in crescita sono considerate maggiormente a rischio di commettere operazioni fraudolente funzionali al raggiungimento di determinati target di mercato.

5. *Indice di deprezzamento* (DEPI) è dato dal rapporto tra il tasso di deprezzamento dell'anno precedente (n-1) rispetto a quello in corso (n), dove tali tassi sono calcolati come i rapporti tra gli ammortamenti e la somma tra gli ammortamenti e le attività tecniche immobilizzate. Un DEPI maggiore di 1 indica che il tasso di svalutazione dei beni è diminuito, il che aumenta la possibilità che la società abbia alzato le stime sulla vita utile dei beni pluriennali o abbia adottato un nuovo e più confacente metodo di valutazione.

6. *Indice delle spese di vendita, generali ed amministrative* (SGAI) è dato dal rapporto tra le spese di vendita, generali e amministrative e le vendite dell'anno corrente (n), rispetto alla corrispondente misura dell'anno precedente (n-1). Un SGAJ maggiore di 1 indica che le spese di vendita, generali ed amministrative aumentano più che proporzionalmente rispetto alle vendite e rappresenta un rallentamento dell'attività d'impresa che rischia di peggiorare le prospettive future.

7. *Indice di leva finanziaria* (LVGI) è dato dal rapporto tra il totale dei debiti di un'impresa e il totale attivo dell'anno corrente (n) rispetto all'anno precedente (n-1) e misura la variazione dell'indebitamento da un anno all'altro. Un LVGI maggiore di 1 equivale ad un aumento della leva finanziaria, ovvero un aumento della componente di debito nella struttura finanziaria dell'impresa e potrebbe essere conseguente agli incentivi legati ai contratti di debito che possono indurre a porre in essere politiche di manipolazione del reddito.

8. *Accruals totali rispetto al Totale Attivo* (TATA) è dato dal rapporto tra la differenza tra le attività correnti (escluse le liquidità immediate), le passività a breve e gli ammortamenti e il totale dell'attivo (al denominatore). Pertanto, non rappresentando gli *accruals* una componente materiale, un indice TATA elevato potrebbe essere conseguente ad utili d'esercizio basati su *accruals*. All'opposto, invece, l'indice TATA può essere negativo quando il cash flow operativo supera di gran misura gli utili netti, ovvero, gli *accruals* sono negativi.

In definitiva, si può osservare che DSRI, AQI, DEPI e TATA sono utili per catturare entrambe le distorsioni dovute a manipolazioni degli utili mentre GMI, SGI, SGAJ e LVGI sono funzionali a impegnarsi nella manipolazione degli utili a causa del deterioramento delle condizioni economiche (Beneish et al., 2013).

La tabella 1 riassume i coefficienti stimati delle variabili e della costante, la loro significatività statistica (t test) e i valori medi osservati dalle imprese classificate da Beneish, nel suo campione, come potenzialmente manipolatrici e non potenzialmente manipolatrici delle variabili.

**Tabella 1.** *Coefficienti, significatività e valori medi delle otto variabili del Modello di Beneish (1999)*

| <i>Variabili</i> | <i>Coefficienti</i> | <i>t</i> | <i>Imprese<br/>potenzialmente<br/>manipolatrici</i> | <i>Imprese<br/>potenzialmente<br/>non manipola-<br/>trici</i> |
|------------------|---------------------|----------|-----------------------------------------------------|---------------------------------------------------------------|
| Costante         | -4,840              | -11,01   | -                                                   | -                                                             |
| DSRI             | 0,920               | 6,02     | 1,465                                               | 1,031                                                         |
| GMI              | 0,528               | 2,20     | 1,193                                               | 1,014                                                         |
| AQI              | 0,404               | 3,20     | 1,254                                               | 1,039                                                         |
| SGI              | 0,892               | 5,39     | 1,607                                               | 1,134                                                         |
| DEPI             | 0,115               | 0,70     | 1,077                                               | 1,001                                                         |
| SGAI             | -0,172              | -0,71    | 1,041                                               | 1,054                                                         |
| LVGI             | -0,327              | -1,22    | 1,111                                               | 1,037                                                         |
| TATA             | 4,679               | 3,73     | 0,031                                               | 0,018                                                         |

Per la società sulla quale si intende indagare la potenziale manipolazione è possibile calcolare un valore *M-Score* (8) dopo aver sostituito, per ciascuna variabile del modello, gli indici calcolati con gli specifici valori di bilancio della società in esame. L'interpretazione dei risultati ottenuti dal modello Beneish del 1999 è la seguente:

se *M-Score* (8)  $> -1,78$  c'è un'alta probabilità di manipolazione;

se *M-Score* (8)  $< -1,78$  c'è una bassa probabilità di manipolazione.

Questa soglia è il valore più utilizzato e accettato nella letteratura che ha adottato questo modello, ma non è l'unico proposto, poiché, per discriminare le aziende manipolatrici, è fondamentale scegliere il valore soglia (*cut-off*). Questa scelta dipende dal costo e dalla consistenza degli errori probabilistici di tipo I (considerare non alterato un bilancio che è alterato) e II (considerare alterato un bilancio che non è alterato) errori del modello: abbassare il valore di cut-off riduce gli errori di tipo I e aumenta gli errori di tipo II e certamente per gli investitori è più costoso considerare come non alterato un bilancio che lo sia (Beneish et al., 2013; Giunta et al., 2014).

In questa analisi, quindi, dopo aver considerato i limiti insiti nel modello Beneish (1999; Beneish et al., 2013) ed evidenziati in varie ricerche (Giunta et al., 2014), come già fatto in altri studi che hanno preso in esame aziende operanti in diversi contesti paese, abbiamo adottato un valore soglia (cut-off) per *M-score* (8) di -2,22 (Muntari, 2015). In altre parole, se *M-Score* (8)  $> -2,22$  c'è un'alta probabi-

lità di manipolazione, mentre se  $M\text{-Score } (8) < -2,22$  c'è una bassa probabilità di manipolazione.

Inoltre, poiché il modello di Beneish appena rappresentato è basato sulla stima di variabili parametrize al contesto economico-istituzionale statunitense, l'applicazione dello stesso modello su imprese operanti in contesti paese profondamente differenti, come quello italiano, potrebbe "inficiare" l'analisi. Pertanto, ai fini della presente ricerca, è stato utilizzato un secondo modello di analisi maggiormente adottato al contesto italiano, caratterizzato da piccole e medie imprese che redigono il bilancio secondo le prescrizioni del Codice civile, integrato dai principi contabili nazionali statuiti dall'Organismo Italiano di contabilità (OIC).

Un adattamento del modello al contesto italiano è stato proposto da Giunta, Bini e Dainelli (2014), che hanno adottato lo stesso metodo di Beneish per stimare i coefficienti degli indici al fine di ottenere la formula del *Manipulation Score* per il contesto italiano. Non essendo stati inseriti gli indici SGI, SGAI e TATA in quanto non significativi, la formula è quindi composta da cinque variabili:

$$M\text{-Score } 5 (IT) = -6,2273 + 0,488 \cdot DSRI + 0,1871 \cdot GMI + 0,2001 \cdot AQI + 0,2819 \cdot DEPI + 0,6288 \cdot LVGI \quad [2]$$

In questa applicazione, il valore soglia è pari a  $-4,14$ . Pertanto, se  $M\text{-Score } 5 (IT)$  calcolato per ciascuna società è maggiore di  $-4,14$ , i dati contabili potrebbero essere stati alterati. In altri termini, è prudente considerare manipolati quei bilanci i cui valori contabili, elaborati secondo la formula [2], restituiscono un valore del  $M\text{-SCORE } 5 (IT)$  superiore a  $-4,14$ . Inoltre, adottando tale soglia, Giunta et al. (2014), stimano che il suddetto modello identifica correttamente oltre il 92% delle imprese manipolatrici del campione, con errori del tipo I al 7,14% e del tipo II al 21,60% (Giunta et al., 2014).

## 2.2 Selezione e definizione del campione oggetto di analisi

L'analisi è condotta su un campione di società italiane non finanziarie quotate alla Borsa di Milano attive nel periodo 2005-2016, al fine di limitare i possibili effetti contraddittori derivanti da un'analisi non focalizzata su un unico contesto paese (Trombetta e Imperatore, 2014) e di migliorare la conoscenza sulla relazione in esame, con specifico riferimento al contesto italiano.

Il campione iniziale comprendeva aziende che coprono tutti i settori, ma quello finale riguarda 75 aziende. Nello specifico, al fine di garantire una maggiore omogeneità e rappresentatività dei fenomeni indagati, sono state escluse le società finanziarie, bancarie, assicurative (Klein, 2002; Peasnell et al., 2005) e di servizi

pubblici (Arkan, 2015), in quanto, la loro precipua caratterizzazione in termini gestionali, derivante da un'articolata disciplina legale e regolamentare, avrebbe inevitabilmente "influenzato" l'analisi. Sono state escluse, inoltre, anche le società straniere quotate nella borsa valori italiana e le società per le quali non sono disponibili, per l'intero arco temporale di riferimento, le informazioni relative alle variabili contabili. Infine, dal campione sono anche escluse, quelle società i cui valori di riferimento si discostano notevolmente dalla media delle altre osservazioni (*outliers*). I dati economico-finanziari sono stati raccolti da DataStream (ultimo accesso, 7 ottobre 2021) mentre i dati macroeconomici sono stati raccolti dal Dataset della Banca Mondiale (ultimo accesso, 3 settembre 2021).

### ***2.3 Scelta del periodo di indagine e del punto di partenza della crisi finanziaria***

La scelta del periodo di indagine e del *cut-off point* di partenza della crisi finanziaria rappresentano aspetti molto critici, non privi di conseguenze in termini di influenza sui risultati e sui conseguenti risvolti teorici e pratici (Franceschetti, 2018).

Difatti, anche con specifico riferimento agli studi sul contesto europeo, è possibile osservare scelte diverse circa il periodo di indagine: 2005-2011 (Iatridis e Dimitras (2013), 2005-2012 (Vladu, 2013) e 2008-2011 (Kousenidis et al, 2013) nonché del punto di partenza (*cut-off point*) della crisi finanziaria: 2008 (Vladu, 2013), 2009 (Iatridis e Dimitras, 2013), 2010 (Kousenindis et al., 2013).

Pertanto, per testare le ipotesi alternative non direzionali precedentemente formulate, il presente studio ha individuato il 2009, coerentemente con Iatridis e Dimitras (2013), come anno di inizio di un periodo di crisi. Infatti, osservando la crescita annua dell'Italia, simile a quella dei paesi di riferimento (Portogallo, Irlanda, Italia, Grecia e Spagna) delle società quotate analizzate da Iatridis e Dimitras (2013), si registra una contrazione dell'1% nel 2008 rispetto al 2007, mentre nel 2009 si registra una contrazione del 5,3%, rispetto al 2008, che denota un forte segnale di crisi.

La particolarità della presente analisi consiste nel fatto che, in via differente rispetto ad altre analisi similari, si estende il periodo di indagine a tre diversi sottoperiodi, ovvero: periodo pre-crisi, periodo di crisi e periodo post-crisi. (Habib et al. 2013; Iatridis e Dimitras, 2013; Kousenidis et al., 2013; Rusmin et al., 2013; Vladu, 2013).

La scelta del 2005 come primo anno di osservazione è riconducibile, coerentemente con quanto rilevato da Iatridis e Dimitras (2013) e Vladu (2013), all'obbligo imposto alle società quotate italiane, così come a tutte le società quotate europee, di

redigere il bilancio secondo i dettami dei principi contabili internazionali (IAS-IFRS) proprio da quell'esercizio contabile.

Inoltre, al fine di definire su base oggettiva la delimitazione temporale di ciascun sottoperiodo di analisi, è stato osservato l'andamento del rapporto tra gli utili e le attività complessive delle società oggetto di analisi nel periodo dal 2005 al 2019 (ultimo anno di disponibilità dei dati). Successivamente, assumendo come riferimento il 2005, si è proceduto al calcolo delle variazioni del suddetto rapporto.

Di conseguenza, dall'analisi congiunta dell'andamento del rapporto tra utili e totale attivo e variazioni annue, dato che il 2005 è l'anno di inizio dell'analisi, e che il periodo di crisi è iniziato nel 2009, è stato possibile definire un tempo omogeneo (quattro anni) per ogni sottoperiodo: 2005-2008 (periodo pre-crisi); 2009-2012 (periodo di crisi) e 2013-2016 (periodo post-crisi).

### 3. Risultati

Coerentemente con quanto definito nell'ambito della sezione metodologica, è stato utilizzato il modello *M-Score* (8) e il *modello M-Score IT* (5).

#### 3.1 Risultati *M-Score* (8) e *cut-off* -2.22

Per ciascuna società e per ciascuno dei 12 anni osservati sono stati calcolati i valori medi di ciascuna variabile ricompresa nel *M-Score* (8). Tuttavia, per esigenze espositive, nella Tab. 2 tali valori sono riportati per ciascun sottoperiodo.

**Tabella 2.** Valori medi del modello Beneish e del campione nei tre periodi considerati

|             | Manipolatrici<br>(valori medi<br>di Beneish) | Non<br>Manipolatrici<br>(valori medi<br>di Beneish) | Periodo pre-<br>crisi<br>(2005-2008) | Periodo di<br>crisi<br>(2009-2012) | Periodo<br>post-crisi<br>(2013-2016) |
|-------------|----------------------------------------------|-----------------------------------------------------|--------------------------------------|------------------------------------|--------------------------------------|
| <i>DSRI</i> | 1,465                                        | 1,031                                               | 1,0076                               | 1,0764                             | 1,0401                               |
| <i>GMI</i>  | 1,193                                        | 1,014                                               | 1,0267                               | 0,9480                             | 0,9406                               |
| <i>AQI</i>  | 1,254                                        | 1,039                                               | 1,1381                               | 1,0797                             | 1,1112                               |
| <i>SGI</i>  | 1,607                                        | 1,134                                               | 1,1044                               | 0,9933                             | 1,0199                               |
| <i>DEPI</i> | 1,077                                        | 1,001                                               | 1,1549                               | 1,0039                             | 1,0668                               |
| <i>SGAI</i> | 1,041                                        | 1,054                                               | 1,0490                               | 1,0238                             | 1,0411                               |
| <i>LVGI</i> | 1,111                                        | 1,037                                               | 1,0941                               | 1,0560                             | 1,0606                               |
| <i>TATA</i> | 0,031                                        | 0,018                                               | -0,0321                              | -0,0511                            | -0,0425                              |

Diventa interessante esaminare le singole variabili per capire quali specifici parametri, confrontati con i valori stimati da Beneish (1999), determinano il rischio di potenziale manipolazione.

Partendo dal DSRI, notiamo che questo valore aumenta in tempo di crisi e diminuisce leggermente nel periodo successivo. Va evidenziato che un aumento di tale indice è associato ad una maggiore probabilità di manipolazione; i valori calcolati di DSRI sono più vicini al valore medio dell'indice di Beneish (1999) per le società non manipolatrici (1,031). Pertanto, non è possibile capire se si tratta di una sospetta inflazione dei ricavi o di un cambiamento nella politica commerciale delle imprese.

Con riferimento al GMI, si può osservare che questo indice è più deteriorato nel periodo pre-crisi e questo mette sotto pressione il management e quindi aumenta il rischio di manipolazione. In effetti, in questo periodo il valore del GMI (1,0267) è superiore al valore medio dell'indice delle imprese non manipolatrici (1,014), ma è comunque lontano dal valore medio delle imprese manipolatrici (1,193).

L'AQI è superiore al valore 1 in tutti i periodi. La qualità delle attività risente in modo particolare del periodo pre-crisi, dove il valore medio dell'indice (1,1381) si attesta comunque su un livello inferiore a quello delle imprese manipolatrici (1,254).

SGI mostra valori medi più vicini a quelli delle società non manipolative in tutti i periodi considerati. Pertanto, considerato che tale indice misura l'evoluzione dei ricavi, si ritiene che, durante l'intero periodo osservato, non vi siano state particolari pressioni all'incremento dei ricavi.

Il DEPI mostra un valore medio particolarmente significativo nei periodi pre-crisi e post-crisi: in particolare, il valore medio pre-crisi (1,1549) è superiore al valore medio delle imprese manipolatrici (1,077); ciò significa che i redattori del bilancio sembrano aver sfruttato il tasso di deprezzamento per aumentare il reddito d'impresa.

SGAI mostra un valore maggiore di 1 in tutti i periodi, ma è particolarmente significativo nei periodi pre-crisi (1,0490) e post-crisi (1,0411).

I valori LVGI sono superiori all'unità in tutti i sottoperiodi. Tuttavia, mentre nel periodo pre-crisi (2005-2008) l'indice raggiunge 1,0941, media leggermente inferiore a quella delle società potenzialmente manipolatrici (pari a 1,111), nei due periodi successivi mostra valori più vicini alla media delle non manipolatrici (1,037), seppur leggermente superiori, 1,0560 (crisi-periodo) e 1,0606 nel post-crisi. La particolare contrazione di questo indice nel periodo di crisi riflette la difficoltà di ac-

cesso al credito per le imprese; pertanto, gli incentivi legati ai contratti di debito, che possono portare alla manipolazione degli utili sono limitati.

La TATA è sempre negativa nei tre periodi sotto osservazione; ciò accade quando il flusso di cassa operativo supera l'utile netto. La componente dei ratei è negativa, quindi il rischio che gli utili siano stati manipolati è relativamente basso.

Sulla base di quanto finora emerso, è evidente una maggiore propensione verso politiche di gestione degli utili nel periodo pre-crisi.

Tutti i valori delle variabili vengono poi utilizzati per calcolare l'*M-Score*, ovvero il valore che indica se la probabilità di manipolazione è alta o bassa. Per testare le ipotesi bidirezionali è stato esplorato *M-Score* con otto variabili e sono stati utilizzati i coefficienti stimati da Beneish (1999) e un valore di cut-off di -2,22 (Muntari, 2015).

Per evidenziare in modo specifico il potenziale rischio di manipolazione, nella Tab. 3 viene riportato il numero e il valore percentuale delle imprese potenzialmente manipolatrici ( $M\text{-Score} \geq -2,22$ ) e non manipolatrici ( $M\text{-Score} \leq -2,22$ ) per sottoperiodo, ottenuto calcolando il valore del *M-Score* (8) per ciascuna società del campione e per ciascuno dei 12 anni.

La Tab. 3 riporta, per ciascun sottoperiodo, il numero di società classificate come potenzialmente manipolatrici e non manipolatrici.

**Tabella 3.** Numero delle imprese potenzialmente manipolatrici e non manipolatrici, suddivisi per sottoperiodi, utilizzando *M-Score* (8) e cut-off -2,22

|                                    | Periodo<br>(2005-2016) | Periodo<br>pre-crisi<br>(2005-2008) | Periodo di crisi<br>(2009-2012) | Periodo post-<br>crisi<br>(2013-2016) |
|------------------------------------|------------------------|-------------------------------------|---------------------------------|---------------------------------------|
| <i>Punteggio</i><br>$M \leq -2,22$ | 698                    | 217                                 | 251                             | 230                                   |
|                                    | 77,56%                 | 72,33%                              | 83,67%                          | 76,67%                                |
| <i>Punteggio</i><br>$M \geq -2,22$ | 202                    | 83                                  | 49                              | 70                                    |
|                                    | 22,44%                 | 27,67%                              | 16,33%                          | 23,33%                                |
| <i>N</i>                           | 900                    | 300                                 | 300                             | 300                                   |

In primo luogo, durante tutto il periodo di indagine, si riscontra una bassa presenza di aziende ad alto rischio di manipolazione (22,44%) rispetto ad un'elevata presenza di aziende a basso rischio (77,56%).

In secondo luogo, l'indice *M-Score* (8) ha classificato 83 aziende come potenzialmente manipolatrici nel 2005-2008. Nel periodo pre-crisi, il numero di aziende

classificate come potenzialmente manipolatrici era superiore a quello del periodo di crisi (49 aziende) e a quello del periodo post-crisi (70 aziende). In altre parole, in base all'indice *M-Score* (8) si rileva che nel contesto italiano le precarie condizioni macroeconomiche e le conseguenti difficoltà subite dalle società quotate non hanno costituito un incentivo ad attuare politiche di manipolazione dei redditi durante il periodo di crisi (2009-2012).

### 3.2 Risultati *M-Score IT* (5) e *cut-off* -4,14

Per testare le ipotesi bidirezionali è stato calcolato l'indice *M-Score IT* con cinque variabili (sono escluse, per le ragioni precedentemente esposte, le variabili SGI, SGAI e TATA) ed è stato utilizzato valore di *cut-off* di -4,14 (Giunta et al., 2014). Per evidenziare in modo specifico il potenziale rischio di manipolazione, nella Tab. 4 viene riportato il numero e il valore percentuale delle aziende potenzialmente manipolatrici (*M-Score*  $\geq$  -4,14) e non manipolatrici (*M-Score*  $\leq$  -4,14) per ciascun sottoperiodo, ottenuto calcolando il valore del *M-Score IT* (5) per ciascuna società del campione e per ognuno dei 12 anni.

**Tabella 4.** Numero delle imprese potenzialmente manipolatrici e non manipolatrici, per sottoperiodi, utilizzando *M-Score IT* (5) e *cut-off* -4,14

|                                           | Periodo<br>(2005-2016) | Periodo<br>pre-crisi<br>(2005-2008) | Periodo di crisi<br>(2009-2012) | Periodo post-<br>crisi<br>(2013-2016) |
|-------------------------------------------|------------------------|-------------------------------------|---------------------------------|---------------------------------------|
| <i>Punteggio</i><br><i>M</i> $\leq$ -4,14 | 771                    | 248                                 | 267                             | 256                                   |
|                                           | 85,67%                 | 82,67%                              | 89,00%                          | 85,33%                                |
| <i>Punteggio</i><br><i>M</i> $\geq$ -4,14 | 129                    | 52                                  | 33                              | 44                                    |
|                                           | 14,33%                 | 17,33%                              | 11,00%                          | 14,67%                                |
| <i>N</i>                                  | 900                    | 300                                 | 300                             | 300                                   |

In primo luogo, durante tutto il periodo di indagine, si riscontra una bassissima presenza di società con potenziale rischio di manipolazione del bilancio (14,33%) rispetto ad un'elevata presenza di società a basso rischio (85,67%).

In secondo luogo, l'indice *M-Score IT* (5) ha classificato 33 aziende, pari all'11%, come potenzialmente manipolatrici nel periodo di crisi (2009-2012). Nel periodo pre-crisi, il numero di aziende classificate come potenzialmente manipolatrici (52) era superiore a quello del periodo di crisi e a quello del periodo post-crisi (44 aziende). In altri termini, in base all'indice *M-Score IT* (5) si rileva che nel con-

testo italiano le precarie condizioni macroeconomiche, e le conseguenti difficoltà subite dalle società quotate, non hanno costituito un incentivo ad attuare politiche di manipolazione dei redditi durante il periodo di crisi (2009-2012).

In definitiva, i risultati derivanti dall'applicazione dei due differenti indici *M-Score*, ovvero *M-Score (8)* con valore di cut-off pari a -2,22, e *M-Score IT (5)* con valore di cut-off pari a -4,14, sono sostanzialmente concordi, seppur con diversa intensità, ad evidenziare che nel contesto italiano le precarie condizioni macroeconomiche, e le conseguenti difficoltà subite dalle società quotate, non hanno costituito un incentivo ad attuare politiche di manipolazione dei redditi durante il periodo di crisi (2009-2012).

Le ragioni sottostanti a tale comportamento potrebbero essere ricondotte a differenti motivazioni. In primo luogo, il maggior livello di attenzione indotto dalla congiuntura economica sfavorevole potrebbe aver dissuaso i redattori del bilancio ad attuare politiche di manipolazione del reddito a causa di una minore tolleranza del mercato per performance scadenti.

In secondo luogo, è possibile che l'attività di controllo dei revisori dei conti contribuisca ad aumentare la qualità dell'informativa di bilancio, che comporta un basso livello di ricorso a pratiche di manipolazione del reddito (Chia et al., 2007; LaFond e Watts, 2008; Francis et al., 2013; Iatridis e Dimitras, 2013).

Infine, è possibile che si sia verificata una diminuzione nelle pratiche di manipolazione del reddito dopo il 2008, probabilmente in presenza di incentivi comuni, soprattutto durante il periodo di crisi (2009-2012), per attrarre potenziali investitori attraverso una informativa finanziaria di qualità (Kousenidis et al., 2013; Cimini, 2015).

#### **4. Conclusioni**

La presente ricerca ha indagato la relazione tra crisi finanziaria e gestione degli utili nel periodo 2005-2016 nel contesto italiano, che è caratterizzato da una bassa presenza di società quotate, da un'elevata concentrazione proprietaria e dalla prevalenza di imprese familiari (Magli e Nobolo, 2014).

Difatti, lo scoppio della crisi finanziaria ha comportato un calo della fiducia nei confronti del bilancio, specie da parte degli investitori, quale strumento idoneo a rappresentare la rilevanza e l'affidabilità del valore dei dati economico-finanziari in esso riportati.

La ricerca è stata motivata da una carenza di studi, con riferimento al contesto italiano, sul rapporto tra crisi finanziaria e politiche di manipolazione del reddito e si caratterizza per la scomposizione del periodo di osservazione in tre sottoperiodi: 2005-2008 (periodo pre-crisi); 2009-2012 (periodo di crisi); 2013-2016 (periodo post-crisi) e per l'utilizzo di due differenti metriche di misurazione e individuazione di imprese potenzialmente manipolatrici del bilancio, quali l'indice *M-Score* (8) nella prospettiva di Muntari (2015) e l'indice *M-Score IT* (5) elaborato da Giunta et al., 2014).

Per ciascuna delle 75 società del campione, è stato calcolato il cd *Manipulation Score*, utilizzando le suddette metriche, che rappresentano due proxies del Modello di Beneish.

In termini generali è emersa una presenza complessivamente bassa di imprese a rischio di manipolazione durante tutto il periodo in esame. Nello specifico, invece, si osserva che il numero più consistente di imprese potenzialmente manipolatrici è riscontrato nel periodo pre-crisi (2005-2008). Difatti, durante l'intensa crisi finanziaria (2009-2012), il numero di società potenzialmente manipolative diminuisce. Infine, solo negli ultimi anni del periodo post-crisi, si registra nuovamente un aumento delle società potenzialmente manipolatrici, senza tuttavia raggiungere la stessa consistenza numerica osservata nel periodo pre-crisi.

Questi risultati hanno diverse interpretazioni e implicazioni per gli accademici, i professionisti, i regolatori e i policy makers che possono apprendere come le crisi finanziarie siano in grado di influenzare positivamente la qualità dell'informativa di bilancio, attraverso il limitato ricorso a pratiche di manipolazione del reddito.

Purtuttavia, lo studio presenta diversi limiti che potrebbero rappresentare un punto di partenza per ricerche future. In primo luogo, ci si è concentrati su un solo paese (l'Italia) e, quindi, si potrebbero ottenere risultati diversi considerando altri contesti con caratteristiche diverse in termini di sistema di governo societario, regolamentazione, tipo di economia e cultura (Franceschetti, 2018). In secondo luogo, lo studio ha considerato due varianti dello stesso modello di analisi. Pertanto, l'utilizzo di un modello alternativo, rispetto a quello utilizzato in questa ricerca, potrebbe essere funzionale ad evitare il condizionamento derivante dalle variabili che caratterizzano il modello Beneish, a testare l'attendibilità dei risultati ottenuti e a confermare o smentire le ipotesi di questa ricerca.

Sebbene questa analisi contribuisca a colmare parzialmente il divario osservato in letteratura, la ricerca futura è altamente incoraggiata ad arricchire fruttuosamente tale dibattito e a migliorare ulteriormente tale ambito. La ricerca futura potrebbe essere indirizzata verso: 1) l'utilizzo di differenti metodi di misurazione delle poli-

tiche di manipolazione del reddito, al fine di confrontare i risultati che emergono dall'utilizzo dei vari metodi; 2) il confronto di campioni di imprese operanti in diversi sistemi paese ed operanti sotto diversi paradigmi di governo societario, al fine di comprendere come l'influenza del contesto legale e regolamentare e della forza di mercato possa influenzare il comportamento contabile delle imprese in termini di politiche di bilancio in condizioni macroeconomiche critiche.

### Riferimenti bibliografici

- Arkan, T. (2015). The effects of earning management techniques, net income and cash flow on stock price. *Finanse, Rynki Finansowe, Ubezpieczenia*, 74, 2: 245-258.
- Balasubramanyan, L.; Zaman, S.; Thomson, J. B. (2013). Are banks forward-looking in their loan loss provisioning? Evidence from the Senior Loan Officer. *Opinion Survey (SLOOS)* (FRB of Cleveland Working Paper No. 13-13).
- Beasley, M. (1996). An empirical analysis of the relation between the board of director composition and the financial statement fraud. *The Accounting Review*, 71, 4: 443-465.
- Beatty, A.L.; Harris, D.G. (1999). The effects of taxes, agency costs and information asymmetry on earnings management: a comparison of public and private firms. *Review of Accounting Studies*, 4: 299-326.
- Beneish, M. D. (1999). The detection of earnings manipulation. *Financial Analysts Journal*, 55, 5: 24-36.
- Beneish, M. D. (2001). Earnings management: A perspective. *Managerial Finance*, 27, 12: 3-17.
- Beneish, M. D.; Vargus M. E. (2002). Insider trading, earnings quality, and accrual mispricing. *The Accounting Review*, 77, 4: 755-791.
- Beneish, M. D.; Lee, C; M.Nichols, D. C. (2013). Earnings manipulation and expected returns, *Financial Analysts Journal*, 69, 2: 57-82.
- Bornemann, S.; Kick, T.; Memmel, C. (2012). Are banks using hidden reserves to beat earnings benchmarks? Evidence from Germany. *Journal of Banking & Finance*, 36, 8: 2403-2415.
- Chia, Y. M.; Lapsley, I.; Lee, H. (2007). Choice of auditors and earnings management during the Asian financial crisis. *Managerial Auditing Journal*, 22, 2: 177-196.
- Cimini, R. (2015). How has the financial crisis affected earnings management? A European study. *Applied Economics*, 47, 3: 302-317.
- Dechow, P.; Skinner, D. (2000). Earnings management: Reconciling the views of

- accounting academics, practitioners, and regulators. *American Accounting Association Accounting Horizons*, 14, 2: 235-250.
- Dechow, P.; Sloan, M. R.; Sweeney, A. P. (1995). Detecting earnings management. *The Accounting Review*, 70, 2: 193-225.
- Ferramosca, S. (2018). *Politiche di bilancio e corporate governance. Relazioni ed effetti economico-aziendali*. Giappichelli, Torino.
- Filip, A.; Raffournier, B. (2014). Financial crisis and earnings management: The European evidence. *The International Journal of Accounting*, 49, 4: 455-478.
- Franceschetti, B. M. (2018). *Financial crises and earnings management behavior. Arguments and evidence against causality*. Springer International Publishing, Cham.
- Francis, B.; Hasan, I.; Wu, Q. (2013). The benefits of conservative accounting to shareholders: Evidence from the financial crisis. *Accounting Horizons*, 27, 2: 319-346.
- Giunta, F.; Bini, L.; Dainelli, F. (2014). Verifica della base informativa per l'analisi di bilancio: le azioni di manipolazione contabile. *Controllo di Gestione*, 2, 11: 5-17.
- Grimaldi, F. (2015). Dalle politiche di bilancio all'earnings management: lineamenti teorici ed evidenze empiriche. In: S. Lippolis, F. Grimaldi (eds.), *Corporate governance e politiche di earnings management*. FrancoAngeli, Milano: 76-103.
- Grimaldi, F.; Muserra, A. L. (2017). The effect of the ownership concentration on earnings management. Empirical evidence from the Italian context. *Corporate Ownership & Control*, 14, 3-1: 236-248.
- Habib, A.; Uddin Bhuiyan, B.; Islam, A. (2013). Financial distress, earnings management and market pricing of accruals during the global financial crisis. *Managerial Finance*, 39, 2: 155-180.
- Healy, P. M.; Wahlen, J. M. (1999). A review of the earnings management literature and its implications for standard setting. *Accounting Horizons*, 13, 4: 365-383.
- Hirst, D.E. (1994). Auditor sensitivity to earnings management, *Contemporary Accounting Research*, 11, I-II: 405-423.
- Iatridis, G.; Dimitras, A. I. (2013). Financial crisis and accounting quality: evidence from five European countries. *Advances in Accounting*, 29,1: 154-160.
- Jenkins, D. S.; Kane, G. D.; Velury, U. (2009). Earnings conservatism and value relevance across the business cycle. *Journal of Business Finance & Accounting*, 36, 9-10: 1041-1058.
- Johnson, M. F. (1999). Business cycles and the relation between security returns and earnings. *Review Accounting Studies*, 4: 93-117.
- Klein, A. (2002). Audit committee, board of director characteristics and earnings

- management. *Journal of Accounting and Economics*, 33, 3: 375-400.
- Kousenidis, D. V.; Ladas, A. C.; Negakis, C. I. (2013). The effects of the European debt crisis on earnings quality. *International Review of Financial Analysis*, 30: 351-362.
- LaFond, R.; Watts, R. L. (2008). The information role of conservatism. *The Accounting Review*, 83, 2: 447-478.
- Leuz, C.; Nanda, D.J.; Wysocki P.D. (2003). Earnings Management and Investor Protection: An International Comparison, *Journal of Financial Economics*, 69: 505-527.
- Mafrolla E. (2009). *Le scelte contabili. Elementi definitivi, studi teorici e ricerche empiriche*. Giappichelli: Torino.
- Magli, F., Nobolo, A. (2014). Corporate governance and control in Italian family business: An empirical analysis. In *EBES - Eurasia Business and Economics Society 2013 Anthology*, 71-91.
- Mari, L. M.; Terzani, S., Agnello, M.; Iorio, S. (2016). Il rischio di manipolazione dei bilanci in tempi di crisi: Analisi di un campione di imprese. *Rivista Italiana di Ragioneria e di Economia Aziendale*, 1, 1-4: 23-39.
- McVay, S. E. (2006). Earnings Management Using Classification Shifting: an Examination of Core Earnings and Special Items. *The Accounting Review*, 81, 3: 501-531.
- Muntari, M. (2015). Detecting corporate fraud and financial distress using the Altman and Beneish models. The case of Enron Corp. *International Journal of Economics, Commerce and Management*, 3, 1: 1-18.
- Paolone, F.; De Luca F.; Prater-Kinsey, J. (2015). The impact of the financial crisis on earnings management: Empirical evidence from the top 5,000 non-listed stock Italian companies. *Ratio Mathematica*, 28: 45-64.
- Peasnell, K. V.; Pope, P. F.; Young, S. (2005). Boards monitoring and earnings management: Do outside directors influence abnormal accruals? *Journal of Business, Finance and Accounting*, 32, 7-8: 1311-1346.
- Pini M. (1991). *Politiche di bilancio e direzione aziendale*, Etas Libri: Milano.
- Prencipe, A. (2006). Earnings quality. Analisi della qualità degli earnings in una prospettiva internazionale. Pearson: Milano.
- Ronen, J.; Yaari, V. (2008). Earnings management: Emerging insights in theory, practice, and research. Springer: New York.
- Roychowdhury, S. (2006). Earnings management through real activities manipulation, *Journal of Accounting and Economics*, 42: 335-370.
- Rusmin, R.; Scully, G.; Tower, G. (2013). Income smoothing behaviour by Asian transportation firms. *Managerial Auditing Journal*, 28, 1: 23-44.
- Trombetta, M., Imperatore, C. (2014). The dynamic of financial crises and its non-monotonic effects on earnings quality. *Journal of Accounting and Public Poli-*

cy, 33, 3: 205-232.

Vladu, A. B. (2013). Smoothing behavior of firms in times of crisis: Empirical evidence from the Spanish economic environment. *Annales Universitatis Apulensis Series Oeconomica*, 15, 2: 473-480.







UNIVERSITÀ  
DEGLI STUDI DI BARI  
ALDO MORO

DIPARTIMENTO DI  
ECONOMIA E FINANZA

PDF finito di comporre il 21 dicembre 2021



