COMPOSITE INDICES FOR MEASURING THE "COMPLEXITY OF DATA COLLECTION" IN ITALIAN MUNICIPALITIES[1]

Katia Bontempi, Samanta Pietropaoli

Abstract. Data collection is an important and strategic phase of every statistical survey. Its organization can be very complex and require considerable effort, especially when different players are included. In particular, one of the most challenging surveys is the Italian Permanent Population and Housing Census. Its data collection time frame is brief and the great variety and volume of the field operations can be quite burdensome. In this context, the municipalities play a key role in the data collection process. We propose an index to evaluate the level of difficulty that the municipalities face during the Census, with the aim of classifying them according to the "effort" required for the data collection activities. Our work focuses on the municipalities participating in the Population Census "List" survey, with most of them being involved in the Census every year, resulting in a significant effort. This effort can be measured by several indicators, combined together to represent a proxy of the phenomenon. Due to the multiple dimensions of the selected indicators, we have applied methodologies known as composite indices.

1. Introduction

Understanding the complexity of data collection processes is crucial for researchers, policymakers and organizations in order to plan and allocate resources efficiently and to optimize data collection strategies.

In the case of the Italian municipalities, the Permanent Population and Housing Census is one of the most challenging surveys because it requires a greater effort in terms of data collection operations compared to other surveys. Furthermore, most of the municipalities are included in Census sample every year; therefore, their effort is stronger and continuous over time and can vary between different areas. Indeed, the municipalities involved may have budgets and personnel which are too limited to meet the data collection requirements.

[1] The article expresses exclusively the authors' opinions. It is the result of the combined work of the authors: K. Bontempi has written sections 1 and 2; S. Pietropaoli has written sections 3, 4 and 5.

Coordinating various activities during field operations and engaging in gathering data can be challenging, as it requires building trust and cooperation with citizens to access valuable information.

In this paper, we propose a composite index capable of providing a measurement of the difficulty level faced by the Italian municipalities involved in the 2022 edition of the Permanent Population Census. The aim is to classify the municipalities according to the effort required by the data collection activities.

The term "complexity" of data collection is employed to identify the main difficulties experienced by the municipalities during their field operations. For some, these operations can be very burdensome due to various factors, including the characteristics of the socio-demographic environment, the extent of the territory and the economic context. All these aspects may influence the population's responsiveness and contribute to the challenges in the data collection process.

Moreover, data collection operations can be time-consuming and costly. Therefore, it is important to plan the data collection activities to meet, in an efficient way, deadlines and budget constraints, especially when conducting extensive surveys such as the Census.

As is well known, this phenomenon cannot be represented by means of a single aspect; it is necessary to use the "combination" of different dimensions, considered together as components of the phenomenon (Mazziotta and Pareto, 2013). This combination can be achieved by applying methodologies known as composite indices (Salzman, 2003; Mazziotta and Pareto, 2011; Diamantopoulos and Riefler, 2008).

We aim to provide a robust and standardized metric that can assist managers and researchers in the assessment of the complexity associated with the data collection activities for each municipality. In fact, the findings of this research could have implications for management decisions, such as resource allocation, in an ethical and responsible manner.

Moreover, fairness, transparency, and an equitable distribution of resources should be taken into account to maintain good relations with the municipalities.

The paper is structured as follows: Section 2 describes the data which provide the reference for the simple indicators; the use of the composite index methodology is described in Section 3; Section 4 discusses the main results obtained using the synthetic indicator chosen (the Mazziotta-Pareto Index); and finally, in Section 5 some conclusions are drawn.

2. Selection of indicators and data sources

We considered all the 1,188 municipalities involved in the "List" survey of the 2022 Population Census and decided to merge a set of indicators from two sources of information.

The first archive considered is "IstatData"[2], the new access platform for aggregating data published by the Italian National Institute of Statistics (Istat) at the end of 2022, which will gradually replace the old database I.Stat. To complete the information for this study, we added some monitoring information from the archive of the Census monitoring system[3].

We selected nine dimensions to reveal and describe the specificities of each municipality, focusing on the individual and territorial characteristics that link the material conditions (labour and territorial extension), socio-demographic aspects (elderly, foreign population, population variation) and quality of life (education). We also took into account the influence of the field activities carried out by the municipalities during the data collection, such as cleaning the list related to the off-target and no-contact units and data collected through the self-completion of the questionnaire. The choice of indicators was driven by the desire to have non-substitutable and highly informative dimensions, which are not compensable, i.e. a deficit in one indicator cannot be balanced by a surplus in another. Indeed, the imbalance between the various dimensions is crucial for an understanding of the complexity of the data collection. Finally, they were chosen according to their relevance, analytical soundness, timeliness and availability. A description of the elementary indicators chosen is provided below:

(a) *Ageing indicator*. The ratio of the population aged 65 years and over to the population aged 0-14 years (percentage - 2021).
(b) *Education indicator*. The ratio of the population with at least a post-secondary school certificate to the total population on December 31st (percentage - 2021).
(c) *Off-target units*. Monitoring indicator of the Census - list units with issues of over-coverage, not belonging to the target population (percentage – 2022).

[2] IstatData is the new platform to disseminate Istat aggregate data. The platform uses the open source tools "Data Browser" and "Meta & Data Manager" developed by Istat (https://sdmxistattoolkit.github.io) following the international standard SDMX (Statistical Data and Metadata eXchange) for the exchange and sharing of data and statistical metadata. It is available at the following link https://esploradati.istat.it/databrowser/#/en .

[3] The monitoring system is called SGI- "Sistema di Gestione Indagine". It is available only for internal use and allows a control of the survey procedures.

(d) *Non-contact units.* Monitoring indicator of the Census - list units for which contact could not be established (percentage – 2022).
(e) *Questionnaires filled in by means of the CAWI technique.* Monitoring indicator of the Census – questionnaires self-filled by users, without any intervention by the municipality (percentage – 2022).
(f) *Foreign population indicator.* The ratio of the foreign population to the total population on December 31st (percentage – 2021).
(g) *Territorial extension.* The ratio of the municipality's surface area to the total Italian surface area (percentage – 2021).
(h) *Employment rate.* The ratio of people employed, aged 15 to 89, to the active population (workforce) (percentage – 2019).
(i) *Population change.* The demographic variation of the population between the years 2001 and 2021 (percentage -2021).

The selection was also inspired by the research undertaken for the Post-Enumeration Survey (Grossi and Mazziotta, 2012; Bernardini et al., 2014), where one of the post-stratification variables was the Hard To Count index (HTC). The purpose of including the HTC as a post-stratification variable was to identify and detect homogeneous areas based on the difficulty faced by a specific subpopulation during the enumeration process. Just as this grouping allowed for a more accurate representation of the population in the final survey estimates, we believe the complexity index can categorize the municipalities in an efficient way.

In this case study, we chose indicators with high data quality in terms of clarity, comparability, completeness and accuracy in a deterministic way. Furthermore, the discussions with other experts, including researchers and managers involved in data collection activities, gave us the confidence to trust the reliability of the results.

3. Complexity and its measurement: methodological aspects

Choosing the right composite index is fundamental for data treatment. Indeed, a "composite index is a mathematical combination (or aggregation as it is termed) of a set of individual indicators (or variables) that represent the different components of a multidimensional phenomenon to be measured (e.g., development, well-being or quality of life). Therefore, the composite indices are used for measuring concepts that cannot be captured by a single indicator" (Mazziotta and Pareto, 2018).

To synthesize the individual indicators into a single measure for each municipality, we used the Mazziotta-Pareto Index (MPI). This decision was driven by a recognition of the method's applicability in aggregating non-substitutable indicators and was also made with careful consideration of the specific 'users'

targeted in this work. It represents an aggregation approach that lends itself to easy interpretation.

Building a composite index is a complex task, involving challenges like data availability, the choice of individual indicators, data treatment, normalization, standardization, and the assigning of appropriate weights.

In our analysis, we employed a formative measurement model, where the level of correlation between the basic indicators is not relevant. This approach allows for independent polarities and correlations and the basic indicators can have positive or negative correlations or may have no correlations (Maggino, 2009).

Normalization is essential to make the individual indicators comparable. We standardized (or transformed into z-scores) the indicators based on the mean and variance of the reference time to convert them to the same dimensionless scale, with an average of 100 and mean square error of 10, resulting in values roughly within the range of 70-130.

The MPI computation is a non-compensative approach. In fact, it introduces a penalty coefficient based on the coefficient of variation, penalizing units with greater imbalances between the individual indicators despite having the same average. This rewards units that exhibit a greater balance between the indicator values (Mazziotta and Pareto, 2020).

We also considered the polarity of indicators in relation to the "complexity" being measured. Some indicators had a positive polarity, such as ageing, foreign population, off-target units, territorial extension and population variation, while others had a negative polarity.

For the system of weights, we opted for an equal weighting, assigning the same weight to all the components.

4. Results

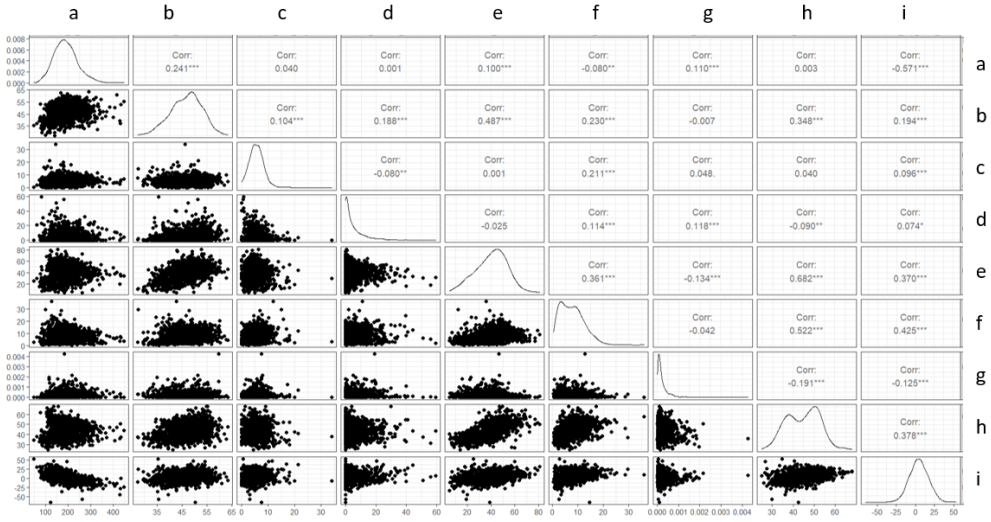
The "complexity index" was assessed using COMiC[4] (*COMposite Index Creator*), a free software designed to compute composite indices using various aggregation methods, based on the SAS programming language.

Figure 1 displays the correlation matrix between the nine indicators chosen in the upper right panel. In general, there are no strong correlations among the indicators, which supports the validity of our indicator selection. As demonstrated by the scatter plots in the lower left panel of the figure, each indicator represents a piece of valuable information. Consequently, it cannot be substituted by any other, affirming our hypothesis regarding the non-substitutability of the selected indicators. The highest

[4] https://www.istat.it/en/methods-and-tools/methods-and-it-tools/analyse/analysis-tools/comic

correlation (0.69) is observed between "Questionnaires filled in with CAWI technique" (e) and "Employment rate" (h). This can be explained in terms of the limited free time available to a person in employment, who usually prefers to fill in the questionnaire independently. The highest negative correlation (-0.57) is found between "Ageing" (a) and "Population change" (i). As expected, municipalities with a high ageing indicator are primarily located in the Liguria Region, characterized by its elderly population. Therefore, a high ageing indicator is often associated with a low birth rate and, consequently, has a negative impact on natural population changes.

Figure 1 – *Correlations plot, scatters plot and densities plot matrix between selected indicators. Year 2022.*



Among the 1,188 municipalities studied, 663 (56.4%) had a complexity index greater than 100. Table 1 and Table 2 display the first 10 and the last 10 Italian municipalities, respectively, sorted by the complexity index.

The first 10 municipalities with a high complexity index value are mainly located in the South of the country, except for Rome and Fiumicino, which are situated in the Central area. Conversely, the last 10 municipalities are all in the Northen region.

It is also worth noting that Rome holds the highest complexity index value among the municipalities, which is not unexpected given its large population and its territorial extension.

Table 1– *The first 10 Italian municipalities sorted by the complexity index. Year 2022.*

Municipality	Province	Region	MPI
Rome	Rome	Lazio	132.32
Acate	Ragusa	Sicily	118.43
Pompei	Naples	Campania	117.59
Cerignola	Foggia	Apulia	114.08
Corigliano-Rossano	Cosenza	Calabria	113.68
Amantea	Cosenza	Calabria	113.22
Fiumicino	Rome	Lazio	112.67
Melito di Napoli	Naples	Campania	112.40
Caltagirone	Catania	Sicily	112.11
Monreale	Palermo	Sicily	111.78

Table 2 – *The last 10 Italian municipalities sorted by the complexity index. Year 2022.*

Municipality	Province	Region	MPI
Buccinasco	Milan	Lombardy	91.11
Valdaora/Olang	Bolzano/Bozen	Trentino South Tyrol/Südtirol	91.90
Gais/Gais	Bolzano/Bozen	Trentino South Tyrol/Südtirol	92.83
Chiusa/Klausen	Bolzano/Bozen	Trentino South Tyrol/Südtirol	92.94
Cusano Milanino	Milan	Lombardy	93.10
Colle Umberto	Treviso	Veneto	93.13
Velturno/Feldthurns	Bolzano/Bozen	Trentino South Tyrol/Südtirol	93.15
Eupilio	Como	Lombardy	93.60
Monticello Conte Otto	Vicenza	Veneto	93.62
Valle Aurina/Ahrntal	Bolzano/Bozen	Trentino South Tyrol/Südtirol	93.70

Figure 2 displays the geographical distribution of the municipalities with a complexity index greater than 100, using a colour scale. Dark green represents the highest complexity levels, while light green indicates the lowest levels. The scale in this figure, as well as the scale in the subsequent Figure 3, is defined by the quartiles of the index distribution. Municipalities with a high complexity are mostly concentrated in the Central and Southern regions. Notably, the Sicily region seems to present widespread criticalities, as does the Lazio region, especially in the metropolitan area of Rome and its neighbouring municipalities. These municipalities differ most significantly in terms of the percentage of Non-contacts (CV 121) and the Territorial extension (CV 112), while they are more similar in terms of the Education indicator (CV 14). On average, they have an Education indicator of 45%, a CAWI completion rate of 35%, and an Employment rate of 42%. The distributions of the Education indicator and the percentage of CAWI completions exhibit positive skewness indices, indicating the presence of a greater number of values closer to the minimum. In fact, the mean value is lower than the median value for these two indicators.

Figure 2 – *Italian municipalities with a complexity index greater than 100. Year 2022.*

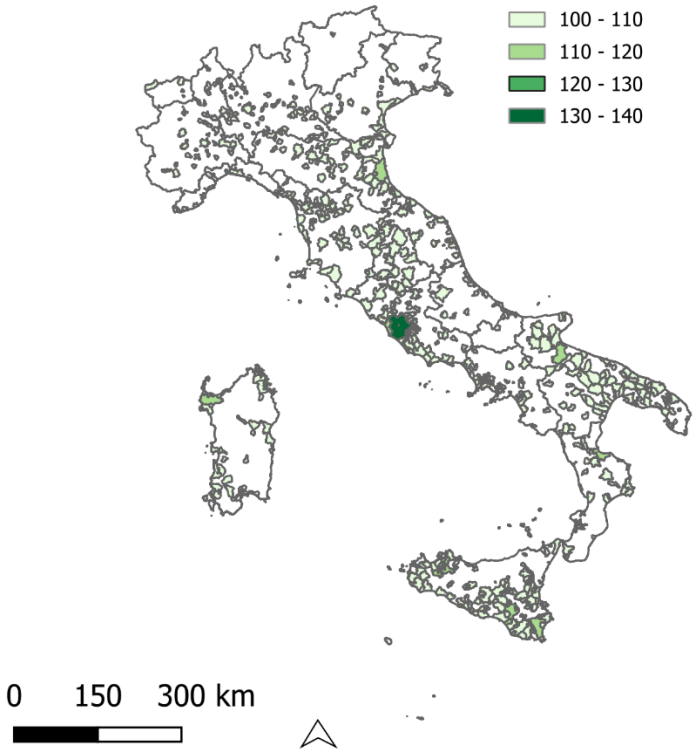
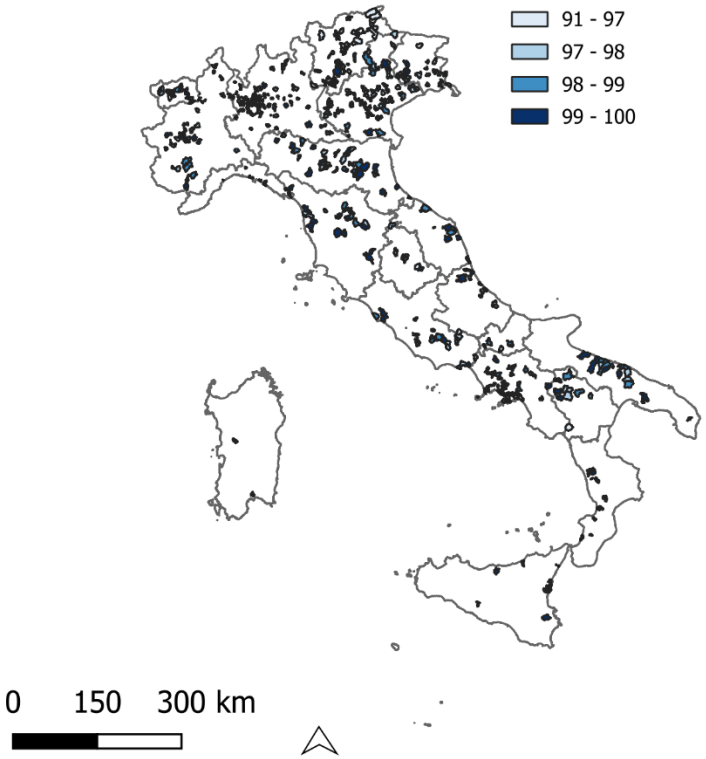


Figure 3 shows the geographical distribution of municipalities with a complexity index lower than 100. The lowest levels of complexity are represented in light blue, while the highest levels are shown in dark blue. Municipalities with a low complexity are concentrated in the North, especially in the North-East area. These municipalities differ most significantly in terms of Population change (CV 193) and the percentage of Non-contact units (CV 140), while they are more similar in terms of the Education indicator (CV 10). On average, they have an Education indicator of 49%, a CAWI completion rate of 47%, and an Employment rate of 48%. The distributions of the Ageing indicator, the percentage of Non-contact units, the Foreign population indicator and the Territorial extension exhibit positive skewness indices, highlighting the presence of a greater number of values closer to the maximum value. This is also evident in the comparison between the mean and median values, as the former is higher than the latter for all these indicators. The distributions of these indicators cannot be assimilated to a normal distribution

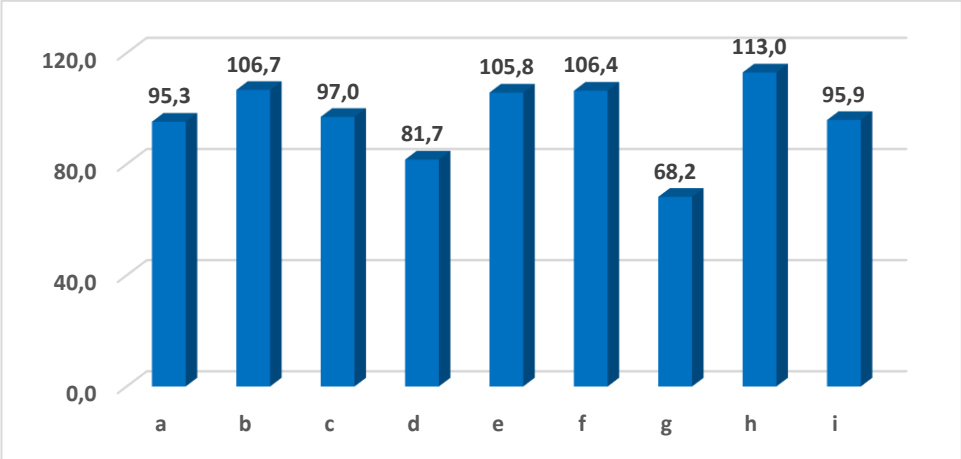
because the kurtosis index suggests that they have sharper or more peaked curves than a normal distribution.

Figure 3 – *Italian municipalities with a complexity index lower than 100. Year 2022.*



Finally, we carried out an influence analysis for the impact of each indicator on the composite index. This analysis identifies, on average, how many positions the municipality's ranking shifts when each indicator is eliminated, one at a time (Figure 4). The most influential indicator is "Employment rate" (h), followed by the "Education" (b) and "Foreign population" (f). On the other hand, the least influential indicators are "Territorial extension" (g) and "Non-contact units rate" (d). Nevertheless, when considering the output of the influence analysis as a whole, it becomes evident that the selected indicators, due to their weak correlations with each other, collectively have a similar influence on the results and are highly informative.

Figure 4 - *Influence Analysis of the Basic Indicators for the MPI Ranking Construction***.**



5. Conclusions

The composite index condenses numerous data into a single value, enhancing the understanding and communication of complex information. It provides a concise and intuitive measure that is easily interpretable and comparable. By combining multiple indicators or variables, it also provides a comprehensive assessment of the concept of data collection complexity, accounting for different dimensions and presenting a holistic perspective on the subject.

The results of this research may have implications for decision-makers involved in data collection processes. Our composite index, serving as a standardized metric for assessing data collection complexity, provides a valuable tool for evaluating and benchmarking the municipalities' data collection operation. For example, by synthesizing the values derived from the indicators considered for each municipality, it becomes feasible to make informed decisions regarding the allocation of economic resources for data collection activities. The use of an impartial and standardized metric is recommended since it can assist managers in resources allocation, ensuring an ethical and responsible distribution. In future research, it may be beneficial to incorporate specific performance indicators related to municipality work into the index computation. These indicators could be used to evaluate the effectiveness, efficiency and quality of the data collection operations conducted by the municipalities.

Istat has been actively involved in evaluating the quality of its censuses for many years. This assessment encompasses both traditional censuses, performed through the Post-Enumeration Survey that include the calculation of the Hard to Count, and permanent censuses. Measuring non-sampling error is a key strategic objective with the aim of continually enhancing the quality of these censuses.

This paper represents only the latest attempt to employ a statistical methodology to classify both the territory and the respondents based on the complexity of data collection. Notably, for the first time in the existing literature, composite indices have been utilized to enhance progressively the quality of the Permanent Population Census over time.

References

BERNARDINI, A., FASULO, A., TERRIBILI, M.D. 2014. A Model Based Categorisation of the Italian Municipalities Based on Non-Response Propensity in the 2011 Census, *Rivista Italiana di Economia Demografia e Statistica*, Vol. 68 No. 3, pp. 79-86.

DIAMANTOPOULOS A., RIEFLER P. 2008. Advancing Formative Measurement Models, *Journal of Business Research*, Vol. 61, pp.1203-1218.

GROSSI, P., MAZZIOTTA, M. 2012. Qualità del 15° Censimento Generale della Popolazione e delle Abitazioni attraverso una Indagine di Controllo che Misuri il Livello di Copertura. *Istat Working Papers*, Roma , Istituto Nazionale di Statistica.

MAGGINO F. 2009. La Misurazione dei Fenomeni Sociali attraverso Indicatori Statistici. Aspetti Metodologici. *Working Papers*, Università degli Studi di Firenze.

MAZZIOTTA M., PARETO A. 2020. *Gli indici sintetici*. Torino: Giappichelli.

MAZZIOTTA M., PARETO A. 2018. Measuring Well-Being Over Time: The Adjusted Mazziotta-Pareto Index Versus Other Non-Compensatory Indices, *Social Indicators Research*, Vol. 136, pp. 967-976.

MAZZIOTTA M., PARETO A. 2016. On a Generalized Non-compensatory Composite Index for Measuring Socio-economic Phenomena, *Social Indicators Research*, Vol. 127, pp. 983-1003.

MAZZIOTTA M., PARETO A. 2013. Methods for Constructing Composite Indices: One for All or All for One, *Rivista Italiana di Economia Demografia e Statistica*, Vol. 67, No. 2, pp. 67-80.

MAZZIOTTA M., PARETO A. 2011. Un Indice Sintetico Non Compensativo per la Misura della Dotazione Infrastrutturale: Un'Applicazione in Ambito Sanitario, *Rivista di Statistica Ufficiale*, Vol. 1, pp.63-79.

SALZMAN J. 2003. Methodological Choices Encountered in the Construction of Composite Indices of Economic and Social Well-Being. *Technical Report*, Center for the Study of Living Standards, Ottawa.

Katia BONTEMPI, Istat, katia.bontempi@istat.it
Samanta PIETROPAOLI, Istat, pietropaoli@istat.it