VOLUME LXXVIII – N. 1 GENNAIO – MARZO 2024

RIVISTA ITALIANA DI ECONOMIA DEMOGRAFIA E STATISTICA

SIEDS
SOCIETÀ ITALIANA
DI ECONOMIA DEMOGRAFIA E STATISTICA

SEDE LEGALE:
C/O Studio Associato Cadoni, Via Ravenna n. 34 – 00161 ROMA

info@sieds.it

rivista@sieds.it

VOLUME FUORI COMMERCIO – DISTRIBUITO GRATUITAMENTE AI SOCI

INDICE

DEVELOPMENT, TERRITORY, SUSTAINABILITY: SOME REFLECTIONS ON THE ROLE OF URBANIZATION AND DEMOGRAPHIC DYNAMICS[1]

Federico Benassi, Luca Salvati

Abstract. Urban concentration played an important role in economic growth over the whole 20th century, being more recently less and less associated with the rate of population growth, suggesting the growing importance of other forces acting on a local scale. Large metropolitan regions, however, seem to escape from this general model, adhering to even more individual growth paths. Regional peculiarities also impact this framework, suggesting how cities growth is mostly unpredictable and largely volatile. Building models of urban growth means to take seriously into account the active constraints – basically land availability and spatial planning. These factors have the indirect objective of envisaging more sustainable urban models, allowing cities to approach sustainability objectives, and reducing environmental, economic, and social risks for the resident population. Moving from a sort of a structural crisis – characteristic of Southern Europe since decades – the present work reflects on a vast portfolio of theoretical approaches and empirical examples contributing to shift toward a resilience discourse in urban affairs. Focusing on both morphological and functional issues, these approaches may provide the appropriate vision to interpret metropolitan complexity in an upcoming urban world. Within this context, resilience of metropolitan regions can be understood as the ability to adapt to economic, technological, and political changes, affecting evolutionary dynamics and trajectories pursued by regional economies.

1. Introduction

A comprehensive understanding of (apparent and latent) mechanism underlying metropolitan growth, provides a more general contribution to the clarification of economic and social development processes on a local and regional scale, placing the territory at the centre of the debate on sustainable, spatially balanced, socially cohesive, and environmentally friendly model of urbanization (Drake and Vafeidis, 2004). Within this framework, cities are no longer places of tradition and history; over time, they have become regions of increasing complexity in all the pillars of sustainability (James, 2014).

[1] The Authors have equally contributed to the conceptualization and to the realization of the article.

Rapidly expanding cities have been seen as internally articulated and fragmented units, difficult to manage and plan, with a dominant position with respect to the surrounding territories, also due to the continuous extraction of natural resources from the nearest rural district. The most recent international reports on urban growth highlighted the crucial role of cities in the increasingly pervasive global transformations persistently observed on our planet – from climate change to water scarcity, from poverty to environmental migration (Iosifides and Politidis, 2005).

The inherent difficulty in analysing how urban systems are articulated, and consolidate, gradually growing or declining, highlights the complexity of this issue. Research tools should, therefore, adhere to a multidisciplinary vision that integrates different methodologies and empirical approaches. Quantitative analysis, exploiting the power of 'big data' available free of charge on a large scale, allows to draw increasingly updated maps of urban growth at an individual level, also through a retrospective analysis of sufficiently long periods, to capture the different socioeconomic dynamics that have shaped metropolitan evolution in the last century (Balk *et al.*, 2006; Quaas *et al.*, 2007; Balsa-Barreiro *et al.*, 2022).

Based on a large collection of indicators, statistical analysis allows for the investigation of integrated dynamics of growth and development at different observation scales, highlighting common growth paths and peculiar individual behaviours (Pisani-Ferry, 2005). These can be traced back to specific historical phases, regional social contexts, and economic cycles, thus contributing to the understanding of how urban development can adhere to long-term sustainability principles. The contribution presents some reflections, theoretically and empirically framed, about on the role played by urbanization processes and demographic dynamics on sustainability considering development needs and environmental constraints. Empirical data are referred to the context of Europe and in particular to the Southern (or Mediterranean) Europe an area of extreme interest given its peculiarity in terms of demographic growth, urban structure, and socio-economic fragility (Heidenreich, 2022).

The structure of the paper is as follows: next section provides an overview of urbanization process as the engine of socio-economic change; section 3 describes the nexus between settlement models and population change; section 4 is devoted to the southern European urban contexts; section 5 holds some reflections on the future city and legacy of the (recent) crises; last section, 6, is devoted to summary considerations and conclusions.

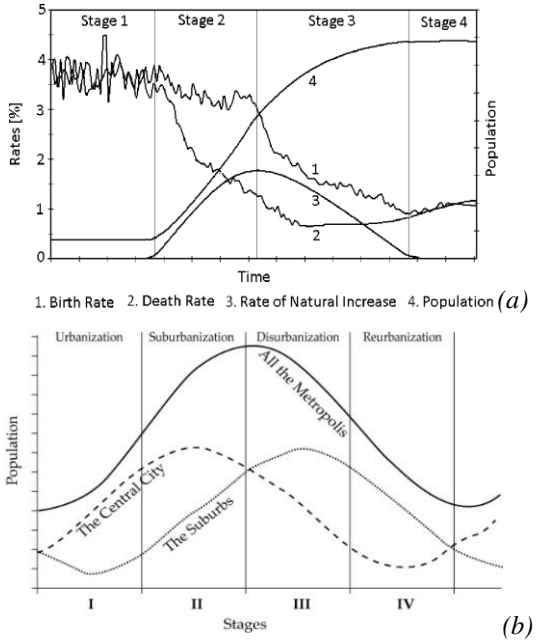
2. Urbanization: the engine of socio-economic change

The joint analysis of specific exploratory approaches, enables to derive an interpretative model of the life cycle of a city (Figure 1), based on progressive phases responding to different local contexts (Zuindeau, 2007). Currently, in Europe, we

have reached the end of a cycle, approaching a new urban experience, likely to be very different from the one experienced in the last century. To better prefigure this new development path, an empirical modelling of the individual growth paths can be proposed, following three interpretative axes:

(i) the temporal one (identifying active change factors on medium-short time scales and longer times),
(ii) the spatial one (identifying active change factors on local, regional, and wider spatial scales), and
(iii) the sectoral one (highlighting the most significant dimensions of analysis).

Figure 1 – *A summary representation of the first demographic transition in Europe representing the evolution over time of basic population indicators (a); a graphical illustration of the City Life Cycle, distinguishing four development waves in metropolitan regions of Southern Europe (b).*



Sources: our elaboration on Kirk (1996) and Roberts (1991).

Despite the wealth of information, the issue of what future metropolitan models will be in Europe remains and requires further investigation. In consideration of the increasing consumption of land and natural resources, a combined study of the progressive demographic decline in many urban areas and the slower but still sustained settlement growth compared to the past, leads to potentially less sustainable urban models from an environmental perspective. Results should be,

then, more clearly demonstrated, from an economic and social viewpoint. Indeed, at the level of individual cities, there is a great heterogeneity in growth paths.

At the same time, the factors of urban concentration - which have played a crucial role in the last century - are less and less connected to the rate of urban growth, leading to an increasing importance of other forces acting at a local scale. Capital cities also seem to deviate from this general model, adhering to even more individualistic growth paths. The knowledge acquired so far, will contribute to building (deterministic and stochastic) models of urban growth, considering the active constraints – basically land availability and spatial planning. All these factors have the indirect objective of envisaging more sustainable urban models, allowing cities to approach those sustainability objectives long pursued by international organizations and local governments, and reducing environmental, economic, and social risks for the resident population (Samways 2022). Regional growth, and the consequent urban expansion, are considered as expression of cultural and intellectual characteristics of human society.

Within this context, the complex issue of quality of life was considered a pillar of urban sustainability and metropolitan resilience. The practical implementation of quality of life as a basic dimension of sustainable development allows to focus on the importance of composite dimensions of change when describing complex social, economic, and demographic phenomena underlying the regional sustainable development. Assuming quality of life as a relevant dimension of a truly sustainable development path, the empirical findings of future studies may support specific policies in both wealthy and economically depressed regions.

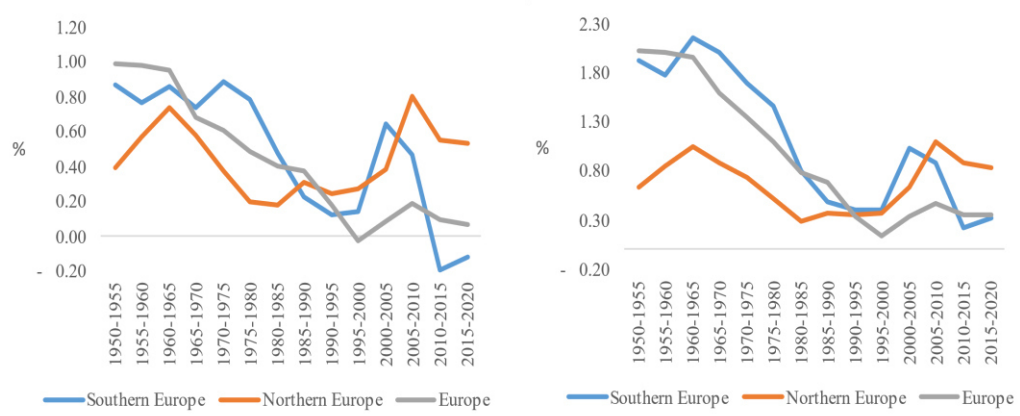
3. Settlement models and regional population change

Settlement models and urban growth were considered strictly interconnected issues since long time. These are two dimensions of regional development that have strong implications for a sustainable land management. Spatial planning, integrated with multifaceted policy dimensions characterizing Mediterranean regions (e.g. social, economic, demographic, cultural, financial, and institutional issues) seems to be an appropriate approach to urban sustainability. Permanent assessment of these factors allows for the implementation of different development scenarios and contributes to systemic and multi-scale strategies of metropolitan growth. The pursuit for comprehensive urban policies achieving an integrated management of human landscapes is finally discussed in the present context of urban crisis in Southern Europe (Figure 2). As a result of social transformations, investigating processes of change in regional spatial structures represents a relevant issue in the identification of monocentric, polycentric, and scattered urban models.

Recent literature has broadly documented how the shift from monocentric spatial organizations to more dispersed structures has determined an increase of

morphological entropy and fractal dimension of land parcels, with a declining importance of the distance from downtown as a factor of urbanization (Polinesi *et al.*, 2020; Benassi and Salvati, 2020; Salvati, 2022). Changes in urban population is a key indicator for understanding settlement models of a given socio-economic system especially if we consider together with the changes in total population. To this aim we can consider Southern Europe and Northern Europe as two regional contexts belong to the same socio-economic and even demographic system (i.e., the European one). These two regional contexts are very different in terms of demographic profiles and dynamics, socio-economic structures and labour market, urban hierarchy so that they represent an interesting case of study (Rees *et al*., 2012; Potančoková *et al.*, 2021).

Figure 2 – *Average annual rate of change of the total population (left panel) and of the urban population (right panel), 1950-205 (per cent). Southern Europe, Northern Europe, and Europe as a whole.*



Author's elaboration on United Nations, Department of Economic and Social Affairs, Population Division (2018). World Urbanization Prospects: The 2018 Revision, Online Edition .

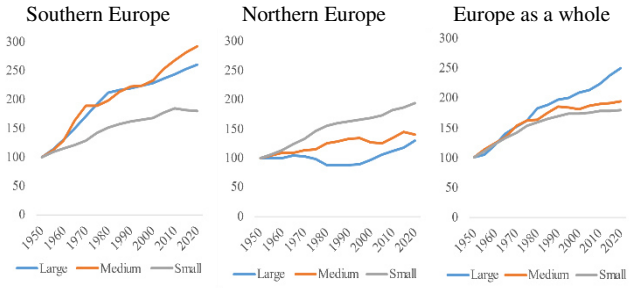
Empirical evidence about long period trends of the annual rate of change for total and urban population tell us a crispy tale. It is quite clear that the trend of urban population rate of change is divergent between Northern and Southern Europe with and inverse tendencies. Staring from 2000-2005 this path is even more clear. The idea is that in Northern Europe the population growth pass by the growth of urban population. While the opposite holds for Southern Europe. This pattern can be linked, in a certain way, to the better life conditions that normally characterise norther cities that are greener, smarter and with a higher level of well-being compared to the ones of the Southern Europe. An indicator of that is the growing level of spatial segregation of foreign population residing in Southern cities,

compared to the Northern European ones, and the growing level of spatial inequalities (Benassi *et al.*, 2020, 2023). Processes that act as detrimental to social cohesion and that have strongly negative effects on the social sustainability of the host societies (Cassiers and Kesteloot, 2012).

4. 'Southern urbanities' from global to local

Recent impulses toward sprawl have increased economic inequality and socio-spatial disparities contributing to a spatially unbalanced distribution of natural amenities with higher consumption of high-quality land. Urban settlements have globally expanded into rural land. Considering conservation of peri-urban biodiversity and local traditions at the fringe of mega-city regions, the role of a typical Mediterranean landscape dominated by olive groves was crucial in urban containment. Having a great cultural, culinary, and aesthetic importance, olive groves characterized Mediterranean peri-urban landscapes in a distinctive way. This contribution identifies individual processes of urban expansion and changes, proposing a new vision for sustainable land management in metropolitan contexts under quick socioeconomic transformations.

Figure 3 – *Population in cities classified by size class of urban settlement, 1950-2020. Index number to a fixed base (1950 = 100). Southern Europe, Northern Europe, and Europe*[(a)].



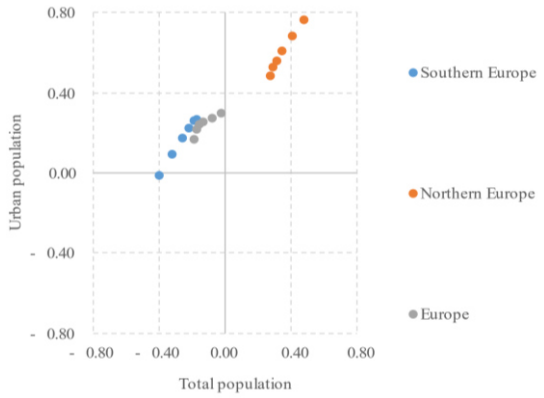
Author's elaboration on United Nations, Department of Economic and Social Affairs, Population Division (2018). World Urbanization Prospects: The 2018 Revision, Online Edition .[(a)]*Small = less than 300,000 residents; Medium = from 300,000 to 1 million residents; Large = major than 1 million and more residents.*

We report here follows some features about Southern Europe, Northern Europe, and Europe (Figure 3). In particular, we consider the variation over the last 70 years of the population in cities classified by size of urban settlement.

The different paths of long-time evolution (in terms of relative variations relating to the initial year) of the "urban" populations are clear. In Southern Europe

urban population residing in medium and large urban settlement recorded the major variation and are more intense in the first type of urban settlement starting from the 2000s. In the Northern Europe the paths is the opposite: here the most intense variations are recorded by population resident in small urban settlement. So, it seems that here a process of population redistribution is happened (and still happening). On the contrary in Southern Europe a process of relative concentration seems to emerge. The causes of these processes can lead to the different level of accessibility and spatial interconnections that characterize the urban settlements of the North and the South of Europe (Figure 4). Finally, if we look to Europe as a whole, we can clearly appreciate how the most intense variations are the ones of the population that reside in Large urban settlements, indicating a process of spatial concentration (Martí-Henneberg, 2005).

Figure 4 – *Scatterplot, average annual rate of change (%) of total and urban population. Southern Europe, Northern Europe, and Europe. 2020-2050.*



Author's elaboration on United Nations, Department of Economic and Social Affairs, Population Division (2018). World Urbanization Prospects: The 2018 Revision, Online Edition.

Within this framework, the specific location of Mediterranean urban regions in-between North-Western affluent cities and developing agglomerations of the world 'South', usually prevents scholars to identify such agglomerations as 'global city regions'. Their characteristics, including hyper-compact forms, dense settlements, poorly organized public services, and limited infrastructural networks, differ from those of the traditional global city regions, and hinder the achievement of a relevant position among the wealthiest European cities. Further discussion is needed on how recent transformations in Mediterranean cities are determined by the diffusion of sparse, low-density settlements thanks to population deconcentration affecting both rural and urban areas. Nevertheless, these regions appear as suitable to understand long-term dynamics of discontinuous and dispersed urban expansion for three main

reasons, as clearly recognized and discussed throughout the book. The first reason refers to urban growth policies depending on Mediterranean deregulation and informality, at least for a long-time window between the 1950s and the 1980s. Second, homogeneous demographic dynamics over space and time – as described by distinct phases of fast and slow population increase – has recently emerged. Lastly, at present, a complex system of interacting agents with local peculiarities, independently from land prices and zoning processes, characterized 'Southern' urban expansion.

5. Future cities and the legacy with socioeconomic crises and environmental challenges

The Lisbon strategy has put the need for sustainable growth at the heart of policy agenda in Europe. The new EU objective of territorial cohesion was added as a third dimension to the old objectives of economic and social cohesion (e.g. Pisani-Ferry 2005). In fact, the main demographic, economic, social, and environmental trends shaping Europe's spatial development represent a challenge for a balanced and sustainable development in the whole Union (Tumpel-Gugerell and Mooslechner 2003). It was recognised that challenges such as regional socio-economic disparities and pressures on the natural heritage require an integrated and multidisciplinary approach in order to correctly monitor their impact as well as a common European-wide policy response (e.g. Drake and Vafeidis 2004). This is particularly true in a context of climate change and increased demand for mitigation and adaptation policies to global transformations, spanning from land-use and demography to socioeconomic structures, hitting distinctively urban and rural areas in Europe.

In this perspective, environmental quality and uncertainty as well as ecological risk in vulnerable regions in Europe, such as the Mediterranean countries, are becoming key words of the knowledge processes, which focus on economic/social dynamics and political actions (e.g. Quaas et al. 2007). Sustainable development has intended, for a long time, how to reconcile growth with environmental quality. In fact, sustainable development involves much more complex aspects of social, economic, and ecological relevance. However, the complexity of the environmental phenomena and their interaction with social and economic processes represents an important challenge for the scientific approach and requires the development of both advanced analytical procedures and adequate policy strategies (Steer 1998).

Problems related to unbalanced resources, economic polarisation, and territorial dichotomy along the Mediterranean basin are clearly the most significant to consider (e.g. Zuindeau, 2007). Their assessment requires an integrated, multidisciplinary approach. In fact, a territory prone to socio-ecological vulnerability represents a critical issue, which is not generally restricted to the resident population, but spreads to other parts of the country (Onate and Peco, 2005). The process often accelerates

territorial unbalances, which may ultimately lead to social conflicts countrywide (Iosifides and Politidis, 2005). Such conflicts may enhance migration movements, representing in the near future a serious obstacle to the achievement of sustainable development in many dry areas of southern Europe.

We assume urbanization as one of the crucial issues of global change. Nowadays, we are increasingly observing a remarkable correlation between socioeconomic changes, urban land use and landscape modifications. Urban expansion pattern is becoming gradually more dispersed and fragmented. Topography and natural amenities seem to contribute to shaping the territorial development in suburban districts, in particular the expansion of urban settlement into high-quality landscapes. These are, in turn, strictly associated with preference for environmental landscape expressed by citizens characterized by uneven socio-economic status, primarily income inequalities. The socio-spatial polarization has been typically observed in Mediterranean cities under structural crises. It results from intense recessions and continuous economic stagnation and may represent a risk in term of sustainability and liveability of modern urban areas. Therefore, it requires a punctual spatial planning activity to restore the image of inclusivity and equality that the "European city" once had.

Official UN forecasts draw future scenarios with differential growth rates for urban and total population. Even more marked will be the difference between the Southern European and Northern European contexts. In the former case, the intensity of the contraction of the total population will increase over time, but the urban population will also experience a slowdown in growth rates, which will nevertheless remain positive until the last period of observation, 2045-2050, when both urban and total population will have a negative growth rate. In the case of Northern Europe, the demographic growth pattern is significantly different. In fact, even though in this case growth rates will be decreasing over time, they will still remain positive in relation to both total and urban population.

Based on these premises, we require more flexible and generalized notions to evaluate urban expansion modes aimed at envisaging well informed socio-territorial policies to contrast unbalanced trajectories of regional development. In recent times, this trend of land development is becoming a common feature of a growing number of cities in advanced economies. Therefore, providing instruments of evaluation of the ongoing situation is a useful and scientific support in the decision-making process to improve the quality and equity of future urbanization in metropolitan areas around the world. Socially cohesive and spatially equitable cities seem to be the necessary antidote to urban crisis in Mediterranean regions. A robust alternative to 'competitive models' of local development seems to be imperative in a context reflective of a sort of 'triple crunch' (austerity policies, climate changes, and increases in oil prices). Going back to the local scale, abandoning interpretative

paradigms oriented toward the logic of 'global networks' paralleled the idea of socioeconomic resilience. While resilience appears as a permanent issue in socioeconomic thought, institutions and policy makers have properly considered the notion of 'resilience' in strategic planning and landscape design only in some cases. Dealing with resilience clearly requires a deep analysis of societies, institutions, and local contexts, understanding the resilient nature beyond the system. This 'holistic' concept further complicated research on local factors of urban resilience and should be imperatively associated with the sustainability challenge. Under this viewpoint, however, resilience theory provides the appropriate knowledge and informs local development built on local-based concepts, rather than focusing only on economic competitiveness factors. These policy perspectives are aimed at envisaging robust economic and social spaces, empowering producers, and consumers to interact locally, to reduce dependency upon distant and larger scale agents, and, in turn, shape the interplay between non-local (large) corporations and the nation state.

6. Concluding remarks

A territorially unbalanced demographic development is an obstacle to social cohesion, economic competitiveness, and socioeconomic development of territories. In this regard, the European Commission wrote in 1999 that "...a polycentric settlement structure across the whole territory of the EU with a graduated city-ranking must be the goal. This is an essential prerequisite for the balanced and sustainable development of local entities and regions and for developing the real location advantage of the EU vis-à-vis other large economic regions in the world" (European Commission 1999: pp. 21). In addition, that kind of models are not compatible with the idea of a sustainable growth model which need to be based upon network of smart and greener cities of medium dimension well interconnected and suited for boost processes of spatial (re)distribution of populations (AISP 2021).

Resilient systems rely upon peculiarities and resources to restart in case of sudden changes. In other words, resilience can be defined as a sort of regional ability to experience positive and socially inclusive economic success, which respects environmental limits and rides global economic purchase. Resilience of metropolitan regions can be further understood as the ability to adapt to economic, technological, and political changes, affecting evolutionary dynamics and trajectories pursued by regional economies. Moving from the structural crisis typical of Mediterranean cities since decades, the present work provides a wide collection of empirical examples and theoretical approaches to shift toward a resilience discourse in urban affairs – focusing on both morphological and functional issues, thus reaching the appropriate vision to interpret metropolitan complexity in an upcoming urban world.

Acknowledgements

Authors would like to thank the organizers of the LIX Scientific Meeting of the Italian Society of Economics, Demography and Statistics and the anonymous reviewers for their excellent work in revising a first version of the paper.

References

AISP. 2021. *Rapporto sulla popolazione. L'Italia e le sfide della demografia.* Bologna: Il Mulino.

BALK D.L., DEICHMANN U., YETMAN G., POZZI F., HAY S.I., NELSON A. 2006. Determining global population distribution: methods, applications and data. *Advances in parasitology*, Vol. 62, pp. 119-156.

BALSA-BARREIRO J., MENDEZ M., MORALES A.J. 2022. Scale, context, and heterogeneity: the complexity of the social space. *Scientific Reports*, Vol.12, 9037

BENASSI F., BONIFAZI C., HEINS F., LIPIZZI F., STROZZA S. 2020. Comparing residential segregation of migrant populations in selected European urban and metropolitan areas, *Spatial Demography*, Vol. 8, pp. 269-290.

BENASSI F., NACCARATO A., IGLESIAS-PASCUAL R., SALVATI L., STROZZA S. 2023. Measuring residential in multi-ethnic and unequal European cities, *International Migration*, Vo. 61, No.2, pp. 341-361.

BENASSI F., SALVATI L. 2020. Urban cycles and long-term population trends in a Southern European City: A demographic outlook, *Applied Spatial Analysis and Policy*, Vol. 13, No.1, pp. 777-803.

CASSIERS T., KESTELOOT C. 2012. Socio-spatial inequalities and social cohesion in European cities. *Urban Studies*, 49(9), pp. 1909-1924.

DRAKE N.A., VAFEIDIS, A., 2004, A review of European Union funded research into the monitoring and mapping of Mediterranean desertification, *Adv. Env. Monit. Mod.*, Vol. 1, pp. 1-51.

EUROPEAN COMMISSION. 1999. *European Spatial Development Perspective: Towards Balanced and Sustainable Development of the Territory of EU*. Luxembourg: Publications Office of EU.

HEIDENREICH M. 2022. Social cohesion in Europe. Between Europe wide-convergence and social and territorial inequalities. In HEIDENREICH M. *Territorial and Social Inequalities in Europe. Challenges of European Integration*, Springer, Cham., pp. 313-339.

IOSIFIDES T., POLITIDIS T., 2005, Socio-economic dynamics, local development and desertification in western Lesvos, Greece. *Local Environment*, Vol. 10, pp. 487-499.

JAMES P. 2014. *Urban sustainability in theory and practice: circles of sustainability*. London: Routledge.

KIRK, D. 1996. Demographic transition theory. *Population Studies*, Vol. 50, No. 3, pp. 361-387.

MARTÍ-HENNBERG J. 2005. Empirical evidence of regional population concentration in Europe, 1870-200. *Population, Space and Place*, Vol. 11, No.4, pp. 269-281.

ONATE J.J., PECO B., 2005, Policy impact on desertification: stakeholders' perceptions in southeast Spain, *Land Use Policy*, Vol. 22, pp. 103-114.

PISANI-FERRY J. 2005, *An agenda for a growing Europe*. Oxford: Oxford University Press.

POLINESI F., RECCHIONI M., TURCO R., RONTOS K., RODRIGO-COMINO J., BENASSI F. 2020. Population trends and urbanization: simulating density effects using a local regression approach, *ISPRS International Journal of Geo-Information*, Vol. 9, No.7, 454.

POTANCOKOVÁ M., STONAWSKI M., GAILEY N. (2021). Migration and demographic disparities in macro-regions of the European Union, a view to 2060. *Demographic Research*, 45, pp. 1317-1354.

QUAAS, M.F., BAUMGARTNER, S., BAKER, C., FRANK, K., MULLER, B. 2007, Uncertainty and sustainability in the management of rangelands, *Ecol. Econ.*, Vol. 62, pp. 251-266.

REES P., VAN DER GAAG N., DE BEER J., HEINS F. (2012). European regional populations: Current trends, future pathways, and policy options, *European Journal of Population*, 28(4), pp. 385-416.

ROBERTS, S. (1991). A critical evaluation of the city life cycle idea. Urban Geography, Vol. 12, No. 5, pp. 431-449.

SALVATI L. 2022. Endogenous population dynamics and metropolitan cycles: long-term evidence from Athens, and eternally Mediterranean city, *European Journal of Population,* Vol. 38, No. 5, pp. 835-886.

SAMWAYS D. 2022. Population and Sustainability: reviewing the relationship between population growth and environmental change, *Journal of Population and Sustainability*, Vol. 6, No. 1, pp. 15-41.

TUMPEL-GUGERELL, G., MOOSLECHNER, P. 2003, *Economic convergence and divergence in Europe. Growth and regional development in an enlarged European Union,* UK: Edward Elgar.

ZUINDEAU, B. (2007), Territorial equity and sustainable development, *Environmental Values*, Vol. 16, pp. 253-268.

Federico BENASSI, Department of Political Sciences, University of Naples Federico II, federico.benassi@unina.it

Luca SALVATI, Department of Methods and Models for Economics, Territory and Finance, Sapienza University of Rome, luca.salvati@uniroma1.it

DISCOVERING INDIVIDUAL PROFILES FROM ADMINISTRATIVE SIGNS OF LIFE USEFUL FOR THE ESTIMATION OF CENSUS RESULTS

Antonella Bernardini, Angela Chieppa, Tiziana Tamburrano

Abstract. The Italian Permanent Population Census (PPC) produces traditional Census results making use of administrative data integrated into statistical registers and survey data. Specific workflows validate administrative records and integrate data related to the same person, producing a standardized data structure that represents the so-called "signs of life" (SoL), referring to a specific reference date or period. SoL classifications and patterns are key for the Permanent Census strategy, especially for the estimation of the usual resident population: each individual in administrative sources is classified as resident according specific conditions related to Sol profiles. Moreover, quality assessment of Census population counts relies on SoL to design an audit survey. SoL can also significantly contribute to estimating thematic aggregates, adding new dimensions to what is collected with the census questionnaire. In this context, continuous evaluation and improvement of SoL classifications are essential. The availability of data from the initial waves of PPC provides a valuable opportunity for experimentation to uncover individual patterns by studying the statistical association between survey responses and the SoL of the same person. In this work, we present the initial results from pattern recognition to evaluate SoL profiles. The data used are derived from the integration of survey data collected in 2021 with administrative SoL for the corresponding year. Multiple Correspondence Analysis and Clustering are employed for an exploratory analysis. Subsequently, a supervised classification tree is used, with the response to the survey as the target variable, and SoL classification is considered among the independent variables. Some patterns and relevant features emerge and point out specific groups of interest as well as issues than call for further analysis and improved SoL classification.

1. Signs of Life for Population Census purposes: the case of the estimation of usual residents

The results of the Italian Permanent Population Census (PPC) are the output of some estimation processes based on survey data, registers, and specific variables derived from administrative sources (Bernardini et al., 2021). These latter sources are designed for administrative purposes, so they need proper processing to meet the quality requirements of official statistics production. Administrative databases also need to undergo processing to derive new features that could serve as variables specifically relevant to the estimation model, in which administrative data are intended to contribute.

The process for producing the census population counts consists of the estimation, for each individual candidate to be resident in Italy, of the dichotomic variable "usual resident (yes/no)" and the categorical one "place of usual residence" (territorial classification of the Italian administrative units). A specific thematic database (or register), called "Integrated Data Base of Usual Residents" (AIDA, from now on) has been implemented in the Italian National Statistical Institute (ISTAT) to exploit, at the individual level, the administrative sources and to derive new valuable information for population count estimation (Bernardini et al., 2019).

The main output of AIDA are the "Signs of Life" (SoL) that could be defined as *structured information derived from administrative sources after proper statistical processing (microlevel linkage, quality evaluation, classification according to expert knowledge, or pattern detection) and designed to support the estimation of usual residents in Italy and their place of residence*.

SoL classifications are key for the Permanent Census strategy. From 2020 on, they constitute the unique source for the estimation of the usual resident population and have also been used as covariates in other Census result estimation models. SoL relevance implies a strong effort to continuously improve and evaluate their quality.

1.1 Initial classification of SoL for Population Census purposes

Each presence of individuals in administrative sources provides data useful to build SoL, according to the definition in the previous paragraph. The sources of the signals are multiple and growing.

The main one is the National Register of the Resident Population (ANPR), that is the national database in which the municipal registry offices (municipal

"Anagrafi") gradually converged during the latest years. ANPR is managed and fed on a local basis by each Italian Municipality, while the Ministry of Interior officers supervise it at the national level. ANPR is the administrative source with the highest quality to get data about people usually living in Italy, because it is specifically designed to store data about resident people, although for administrative purposes, and covers every local territorial unit. Nevertheless, this source still contains coverage errors, due both to individual habits of late or false personal registration and also to living conditions that are less stable than in the past. Data from ANPR are the primary source of the ISTAT Base Population Register (PBR) and determine the first computation of population amounts for specific dates and territorial units before the Census correction is delivered. Only ANPR data that comply with a set of quality checks defined at ISTAT are loaded into the PBR.

Additional sources are currently under study and will soon be introduced, including those related to electricity supply as well as mobile phone big data.

From the integration of all these sources available at ISTAT, direct or indirect signs of life are derived:

- *direct signs of life* – Activities performed by individuals from which a durable period of time (e.g., a year) and a place can be clearly identified (e.g., being a public or private employee, having a regular rental contract);
- *indirect signs of life* – An individual status or a non-professional condition (e.g., children or other relatives as dependents in tax returns).

The integration and loading of the different administrative sources and registers into AIDA are implemented through standardized workflows that run periodically and that produce specific entries in the database, each of these entries representing a structured version of SoL (ISTAT, 2022). Each SoL occurrence corresponds to a specific person and year (e.g., person A, year 2021) and contains various attributes that condense and summarize the integration of the administrative sources. The most important of these attributes are: 1) a sequence that represents the combination of the specific sources presenting information related to that person for the considered year; 2) a sequence that represents the monthly presence of the individual in the administrative sources over the 12 months of the year considered; and 3) the territorial units to which the administrative individual presence is related.

These attributes are the basis for computing SoL classifications that could be useful for Census estimation. Some initial analyses of the association between administrative data and place of usual residence (Chieppa et al., 2018) have determined the classifications currently implemented in AIDA. For direct SoL, the identification of duration patterns in administrative data has been translated into a specific classification that distinguishes between stable/continuous signals,

seasonal signals, discontinuous signals, random signals, and absent or non-useful signals (Bernardini et al., 2019). Additionally, the initial analysis, coupled with expert knowledge, facilitated the formulation of specific computing rules to derive only one 'prevalent place' for each SoL. Moreover, a hierarchy among all indirect signals, established through expert evaluation of different administrative sources, has been utilized to classify each person based on a 'main indirect signal' when multiple signals are present. While these SoL classifications currently implemented in AIDA can label all individuals, there is an ongoing effort to identify new SoL classifications that can enhance the accuracy of predicting usual residence, particularly as additional data sources are integrated.

Table 1 shows the distribution of people with signal in at least one 2021 administrative sources, breakdown according to the presence in ANPR, and the an aggregation of the initial classification of SoL currently available in AIDA.

Table 1 – *Individual profiles based on initial SoL classification and presence in ANPR.*

Presence in ANPR	SoL initial aggregate classification	Counts	%
Not registered in ANPR	Steady signs of work/study	420.287	34,01%
	Signs of university enrollment	26.508	2,14%
	Weak signs of work/study	256.624	20,76%
	Signs others than work/study (permit to stay, rental contracts etc.)	263.722	21,34%
	Episodic presence	268.802	21,75%
Total individual entries with SoL, not registered in ANPR		1.235.943	100,00%
Registered in ANPR	Steady signs of work/study	31.620.418	52,94%
	Retirement/income source signs	16.926.345	28,34%
	Fiscally dependent family member	4.479.095	7,50%
	Weak signs of work/study	1.932.287	3,24%
	Indirect signs of life - several sources	1.218.389	2,04%
	Rental contract	1.042.950	1,75%
	Signs of university studies	991.543	1,66%
	Signs of work/study episodic	422.545	0,71%
	No signs of life	1.096.850	1,84%
Total people registered in ANPR		59.730.422	100,00%

AIDA 2021

The upper rows of the table are related to SoL of people not registered in ANPR but with at least one presence in another administrative source: they're about 1,2

million of people for 2021, in 36,15% of cases have a strong/steady signal that could be considered as candidates to under-coverage of official population register.

The majority (97,46%) of more than 59 million individual records registered in ANPR have also administrative SoL that are coherent with registered place of usual residence in Italy. Nevertheless, 4 million of these individuals have administrative signs out of registered province.

Moreover, there are 2,5% of people in ANPR without any other administrative sign.

These results are the starting point for a deeper analysis on administrative data to improve SoL classifications, as described in following paragraphs.

2. Learning new classifications for SoL from the integration with census survey data

The availability of data from the first waves of PPC is a great opportunity to discover patterns in administrative individual data associated with usual residence. The response to the survey serves as a proxy for each individual's usual residence in a certain territory. Therefore, the survey outcome can be used to evaluate the administrative patterns relevant to predicting the resident population count. These same patterns can be used to determine improved versions of SoL classifications that can constitute covariates in prediction models using administrative data to estimate the population resident in Italy. In this section, we describe the first results of a multidimensional analysis of integrated data aimed at improving the existing initial SoL classification on duration and type of source, as well as basic individual and territorial characteristics.

Multiple Correspondence Analysis (MCA) and Clustering were employed for an exploratory, unsupervised analysis to unveil the complete association structure in the data, identify relevant dimensions, and detect 'natural' clusters. Subsequently, a decision tree classifier was utilized as a tool to provide a more detailed description of specific associations and patterns related to usual residence. The response to the Census survey served as the outcome, guiding the algorithm in its search for final groups.

2.1 Set up the experimental database for learning

Data from different sources have been integrated into a database useful for learning goals through linkage at the individual level of all the different sources and ensuring the coherence of the time or period referenced by the data.

Survey data have been loaded, taking into account the data collection and validation rules. The variables from surveys that could be very useful in analyzing administrative signs are: survey outcome (respondent or not found); place of residence in the previous year; duration of presence; use of the same or other accommodation for systematic travel; some additional information for people found; type of survey (areal or list); survey mode.

From the AIDA database, variables useful for deeper analysis are: the presence or absence of the administrative signal during the last available year; continuity pattern; type or source of signal; and coherence among localizations of signs from different sources.

From PBR, which manages all ANPR validated data but also other individual and territorial attributes, variables considered are: presence or absence on January 1st, on the date of the survey, and on December 31st. Moreover, some territorial classifications of the registered place of residence are also integrated, such as degree of urbanization, regions, and other geographical attributes.

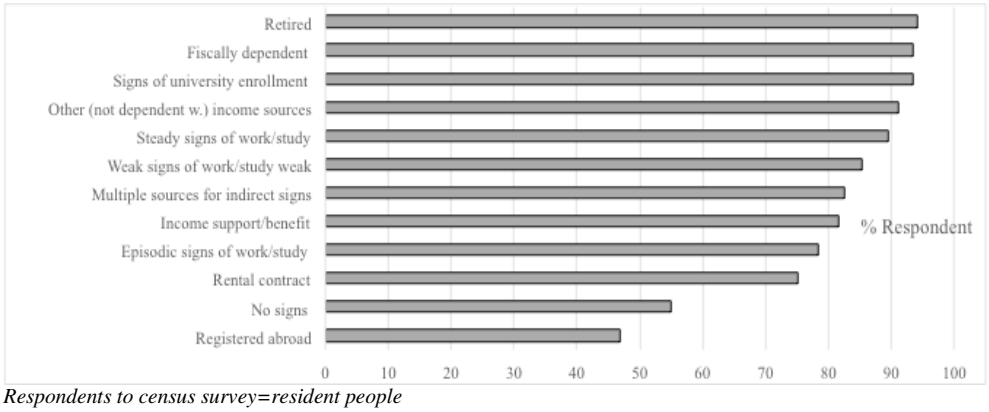
Finally, individual demographic variables are derived from available archives: gender, age, citizenship, and place of birth.

The results described in the following paragraphs are related to the analysis of a specific dataset extracted from the integrated database, where only data related to the year 2021 are considered. This dataset amounts to about 4 million individual records.

3. Multidimensional exploratory analysis on SoL and census survey outcomes

In Figure 1, it is possible to read the response rate at the List Census Survey executed in 2021, with the type of SoL corresponding to sampled people as a breakdown. The survey is executed on a sample of households and individuals extracted from PBR. Only those who confirm that the place of usual residence is the same resulting from the sample list have to respond to the questionnaire, that is to say that respondents are actually residents.

Figure 1 – *Respondents to census surveys and SoL – 2021 L CENSUS SURVEY.*



Respondents to census survey=resident people

For considerations above, the distribution in Figure 1 could be read as a first exploration of the effectiveness of SoL to predict usual residence.

The most evident result is the importance of the pension signal (94,6% of respondents for this group) and of the "indirect" fiscal signals (93,5% of respondents), even stronger than stable work or study signals. For weak signals, i.e., non-continuous, the response rate falls below 90%. People without signals have a response rate of 55%; this group of individuals is clearly critical for prediction based on SoL.

A multidimensional analysis is needed to evaluate the joint effects of SoL classes with other individual or territorial attributes.

MCA is an unsupervised technique for visualizing patterns in large and multidimensional categorical data (Greenacre and Blasius, 2006) by means of identifying principal dimensions that explain and synthesize the variability in the study dataset. Another powerful result when using MCA is the possibility of plotting in the same space defined by the resulting dimensions both categories and cases. When there is a statistical association among them, they are plotted near each other.

In Figure 2, we present the outcomes of a MCA applied to the dataset integrating SoL and survey data. These results aim to assess the relationships among all variables considered in this analysis, including gender, age, citizenship, urbanization degree, regions of residence, response to the survey, and SoL classes. Importantly, age and the degree of urbanization emerge as the primary dimensions explaining the majority of variability in the dataset, along with citizenship. In the graph, age is represented by the horizontal axis, with older ages on the left and younger ones on the right. The territorial dimension, with citizenship, moves along the vertical axis of the scheme: below we have rural areas and medium-sized

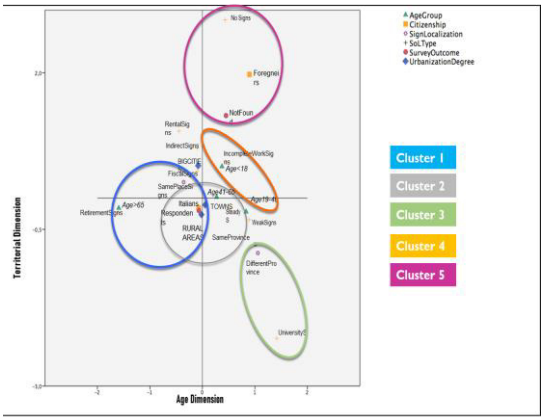
municipalities, while above are more urbanized (cities) areas, with a strong associated presence of foreigners. This association structure is very typical when analyzing Census population phenomena on Italian territory.

A K-means cluster technique (Abidioun et al., 2023) is used to better detect significant patterns. Clusters are plotted in coloured ellipses in the same space resulting from MCA principal dimensions in Figure 2.

Five clusters result from the K-means algorithm:

- Cluster 1, with 54% of cases: Italians, respondents, older people, steady and fiscal and retirement signs; SoL with same localization of PBR.
- Cluster 2, with 27% of cases: Italians, living in medium towns or suburbs, adult ages, stable signs; SoL in different place than PBR but same province. This cluster seems to detect commuters.
- Cluster 3, with 3% of cases: Italians, respondents to survey, living in medium towns, university signs or weak SoL, with different localization than PBR. This cluster rapresents the students living seasonaly where university is located.
- Cluster 4, with 13% of cases: young people, absent or incomplete or fiscal SoL, half respondents half not; critical pattern, difficult to predict place of usual residence; could be Neet (not working or studying) or PBR overcoverage.
- Cluster 5, with 3% of cases: foreigners not founded with survey, registered in cities/urban areas, only rental contract sign or absent/incomplete SoL; critical pattern, difficult to predict if people with this SoL are still usually living in Italy by using only available SoL classes and individual/territorial attributes. Need for more information.

Figure 2 - *Clusters on SoL and survey outcomes.*

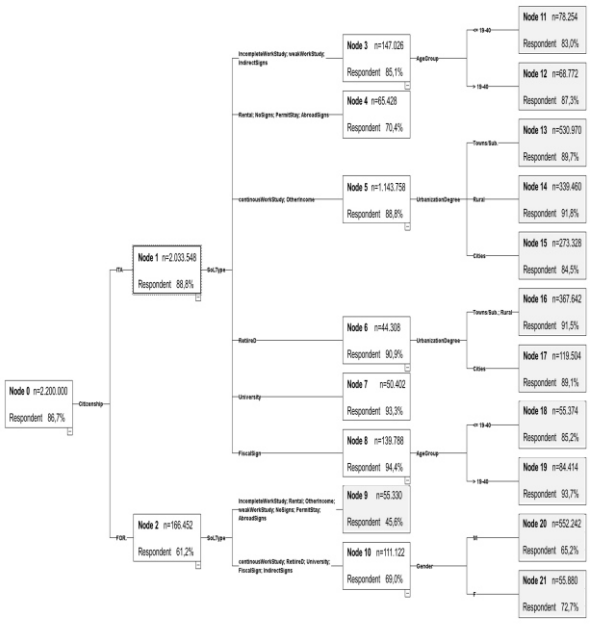


Both MCA and K-means techniques are unsupervised methods; that is, there is no variable to "guide" the pattern detection (Wu, 2012). Clusters derived from MCA and K-means undercover the associations among all considered variables. A supervised classification algorithm is needed to derive patterns that are especially relevant for the prediction of usual residence, considering response to a Census survey as a proxy for being a usual resident. In figure 3, there is a visual representation of the results of a classification tree algorithm: the same variables from MCA and k-means analysis have been used.

The classification tree is a very useful tool to share the results of a pattern recognition and classification model with thematic experts, since the resulting classification could be read through rules that are more easily understood than the parameter output of other predictive models.

The dependent variable chosen is the survey outcome, and the tree classifier adopted is CHAID (Ritschard, 2013), which makes use of the chi-squared association statistic to define, at each level, how to split cases into subgroups, starting from the entire study dataset. On each node, the algorithm splits cases according to rules on different categories of independent variables so that derived subsets are the most associated with the response variable (survey outcome).

Figure 3 - *Patterns according to response to the survey: classification tree result.*



Dataset: census survey sampled individuals CHAID classification tree

The model resulting when using CHAID on the study dataset (see figure 3) does not succeed in identifying rules to accurately predict usual residents in Italy, because almost all subsets detected have a higher percentage of respondents and prediction would be "resident" in almost all cases. Nevertheless, the tree is very useful for a description of individual and territorial profiles associated with the different probabilities of being resident in Italy. Tree results show that first splitting rule coincide with citizenship, that is to say that the patterns of Italians and foreigners respondents are different.

- For Italians, each of the existing SoL classes has a strong association with the probability of being resident: for instance, Italians with a fiscal SoL have a 94,4% of being confirmed, compared to only 70,4% in the case of Italians without administrative signs or only with a rental contract. In some cases (direct signs), the degree of urbanization is needed to better differentiate residents from those who are not found, while in other cases (indirect signs, such as fiscal ones or incomplete or weak signals), the age class is needed.
- Foreigners, on the other hand, have lower probabilities to be found with surveys; this could be related to both undercoverage of the survey or change of usual residence. Moreover, all different available classes of SoL combine in only two relevant groups, forming a critical pattern for prediction: all foreigners without signs, incomplete ones, or only with rental contracts have an almost equal probability of response and not being found; therefore, any prediction one makes (on the probability of response or residence) would make an error with a probability of about 50%. Moreover, the small size of this critical subset (only about 55 thousand individuals out of more than 2 million in the dataset) constitutes an added problem when adjusting a predictive model on these data.

4. Exploring data of critical no-signs group

In previous paragraphs, the group of individuals without signals, both foreigners and Italians, has resulted in a critical or difficult-to-predict pattern by using the variables of the study dataset.

To dissect this group further, additional targeted analyses were undertaken in an attempt to delineate subgroups and identify derived variables that might aid in classifying such cases.

The spatial distribution suggests the presence of distinct subgroups and potential varied living conditions. Noteworthy concentrations of this pattern are evident in several scenarios: municipalities along the country borders (potentially

Italians commuting to neighboring countries); holiday municipalities (lacking signals: "convenience" residents with second homes); southern municipalities where a mix of undeclared workers, genuine unemployed individuals (lacking signals but still residents), and emigrants not yet registered in the census coexist (lacking signals = pending cancellation due to non-residency).

Survey data significantly contribute to supplementing information for this group of people. There are 52.337 individuals without administrative signals who responded to the surveys. 90% of them declared to have resided in the same place or house one year prior, implying that the absence of SoL might not directly lead to removal from the residents register. Furthermore, 18.33% of these individuals indicated a working condition in the Census questionnaire, underscoring potential instances of illicit employment or underreporting in certain labor archives. Survey data from Census questionnaires regarding the marital status of these individuals reveal that 31% are married, suggesting the need to further explore household relationships or indirect signals in AIDA.

5 Some lessons learnt: changes in Population Census design and current experimentations

The use of administrative data for official statistics leads to major changes in the estimation of final official results (UNECE, 2020). Identification of different administrative-specific patterns, associated with each estimation output, is crucial for building the best integration and estimation strategy.

The first results from a multidimensional analysis to evaluate the association of available SoL classifications with usual residence, approximated by survey responses, point out very useful insights. For the majority of people eligible to be residents (at least one administrative sign), existing SoL classes together with individual demographic attributes and the degree of urbanization of the expected place of residence could be used to build accurate predictive models. On the other hand, groups such as foreigners in urbanized areas, people with no signals, and youth from the South with incomplete signs emerge as critical patterns in the sense that they call for an improved SoL classification since the existing classes proved not to be informative enough to predict their usual residence.

To get an accurate estimation also for these critical patterns and to improve existing SoL classes, the current study areas are: 1) improving the pattern recognition analyses by including machine learning techniques and by using all the data from the first cycle (2018–2022) of census waves (Casari and Zheng, 2018); 2) loading new administrative sources with high quality and coverage in AIDA (big data). Moreover, audit surveys are being planned to measure the quality of

population estimates based on administrative data and to gain further insight and data for critical patterns (Solari et al., 2023).

References

ABIDIOUN M. I., EZUGWU A.E., ABUALIGAH L., ABUHAIJA B., HEMING J. 2023. K-Means Clustering Algorithms: A Comprehensive Review, Variants Analysis, and Advances in the Era ff Big Data, *Information Sciences*,Vol. 622, pp. 178-210.

SOLARI F., BERNARDINI A., CIBELLA N. 2023. Statistical Framework for Fully Register Based Population Counts, *Metron*, Vol. 81, pp. 109-129.

BERNARDINI A., CHIEPPA A., CIBELLA N., SOLARI F. 2021. Administrative data for population counts estimations in Italian Population Census. In PERNA C., SALVATI N. and SCHIRRIPA F. (Eds.) *Book of Short Papers*, SIS, pp. 274-278.

BERNARDINI A., CIBELLA N., FASULO A., FALORSI S., GALLO G. 2019. Empirical evidence for population counting: the combined use of administrative sources and survey data. In *ESS Workshop on the use of administrative data and social statistics*, Valencia, Spain.

CASARI A, ZHENG A. 2018. *Feature engineering for machine learning.* Boston: O'Reilly Media, Inc.

CHIEPPA A., GALLO G., TOMEO V., BORRELLI F., DI DOMENICO S. 2018. Knowledge Discovery for Inferring the Usually Resident Population from Administrative Registers, *Mathematical Population Studies.*, Vol. 26, pp. 1-15.

ISTAT. 2022. Nota tecnica sulla produzione dei dati del Censimento Permanente. Roma: Istituto Nazionale di Statistica.

GREENACRE M., BLASIUS J. 2006. *Multiple correspondence analysis and related methods*. New York: Chapman and Hall.

RITSCHARD G. 2013. CHAID and Earlier Supervised Tree Methods. In MCARDLE, J.J. and G. RITSCHARD (Eds.) *Contemporary Issues in Exploratory Data Mining in the Behavioral Sciences*, New York: Routledge, pp. 48-74

UNECE. 2020. New frontiers for censuses beyond 2020. Geneva, United Nations.

WU J. 2012. Cluster Analysis and K-means Clustering: An Introduction. In WU J. (Ed.) *Advances in K-means Clustering*, Berlin, Heidelberg: Springer, pp. 1-16.

Antonella BERNARDINI, Istat, anbernar@istat.it
Angela CHIEPPA, Istat, chieppa@istat.it
Tiziana TAMBURRANO, Istat, tamburra@istat.it

SMALL AREA ESTIMATION OF SEVERE FUNCTIONAL LIMITATION FROM ITALIAN DATA OF THE EUROPEAN HEALTH INTERVIEW SURVEY

Michele D'Alò, Andrea Fasulo, Francesco Isidori, Maria Giovanna Ranalli

Abstract. This paper focuses on the methodology used to estimate the indicator of severe functional limitation (SFL) using data collected from the European Health Interview Survey (EHIS) in Italy. While direct estimates of SFL are reasonably accurate up to the regional level (NUTS2), there is a demand for more detailed estimates at the provincial level (NUTS3), disaggregated by sex and two age groups (15-64 years and 65 years and above). This requires the computation of estimates for 428 unplanned domains. To address this challenge, a small area estimation approach based on an area-level model has been applied, integrating auxiliary information known from administrative registers with EHIS data. To meet the assumptions of the model and ensure in this way a better accuracy of the final estimates, the model has been specified on a log-transformation of direct estimates. The case study presented here is one of the first attempts at obtaining small area estimates for unplanned domains within the EHIS survey and the results obtained are very promising.

1. Introduction

There is a growing demand for increasingly detailed statistical information, particularly regarding estimates of socio-economic indicators at a highly granular level. This demand arises from the need to support urban policies that effectively consider and incorporate specific local characteristics. Decision-makers and policymakers require comprehensive and accurate data to tailor policies that cater to the unique needs and challenges of specific places. Moreover, a considerable number of United Nations Sustainable Development Goals are pursued using survey indicators at a very detailed level.

In order to explore appropriate estimation methodologies for obtaining estimates at a level of granularity beyond that of planned domains by the main social surveys carried out by the Italian National Institute of Statistics (ISTAT), a working group was established, comprising experts in small area estimation from both ISTAT and the Academia. The primary objective of the group was to

delineate the process for computing small area estimates (SAEs) of relevant indicators for unplanned domains, drawing from the main social surveys conducted by ISTAT. In the latest edition of this working group, three subgroups were formed, each with a specific focus on defining the production process of SAE for relevant indicators, collected through three distinct surveys: the European Survey on Income and Living Conditions (EUSILC) for poverty indicators, the Aspect of Daily Life survey (AVQ) for ITC indicators, and the European Health Interview Survey (EHIS) for health indicators. In the previous edition of the working group, there was a dedicated subgroup focusing on SAE of indicators derived from the Labour Force Survey (LFS), particularly at the functional area level. However, since ISTAT has a well-established tradition of applying SAE techniques for LFS indicators, the decision was made to temporarily set aside this specific focus. Nonetheless, the expertise and methodologies developed to produce SAEs for LFS indicators have been valuable for implementing case studies and computing SAEs from other social surveys.

This paper aims to describe the methodology used to estimate the indicator of severe functional limitation (SFL) using data collected through the European Health Interview Survey (EHIS). EHIS gathers information on key aspects of the population's health conditions and the use of healthcare services for citizens aged 15 and above. The adopted sampling design is a two-stage stratified sampling: municipalities are first-stage units, while households are second-stage units. The final sample includes approximately 22,800 households, living in 835 municipalities of different sizes and distributed throughout the national territory. The areas considered in the survey include the five main geographical areas (North-West, North-East, Center, South, Islands) according to the NUTS 1 classification, Italian Regions (NUTS2), and the two autonomous provinces of Bolzano and Trento. Broad areas defined according to the national health program are also domains of interest.

Direct estimates of SFL have an acceptable level of error up to the regional level. Therefore, ISTAT internal request to provide estimates at the provincial level (NUTS3), disaggregated by sex and two age groups (15-64 years; 65 years and above), has required the need to compute estimates for 428 unplanned domains. There are no "a priori" guarantees about the validity of estimates for this level of granularity, as they may have high sampling errors. Therefore, to meet the request to produce an estimate for the indicator at such a level of disaggregation, specific estimation methods for small areas have been adopted.

The paper is structured as follows. In Section 2, a concise overview of the survey sampling design and the methodology employed to calculate the direct estimates of SFL is presented. Section 3 provides a brief description of the small area estimation method applied in the study. In section 4, the main outcomes of the

case study concerning the target parameter are examined in detail. In conclusion, Section 5 presents definitive insights drawn from the main findings and outlines the necessary future work required to further validate the proposed model-based estimates.

2. Sample design, direct estimates and sampling errors

The European Health Interview Survey (EHIS) is conducted in all European Union member states with the aim of computing comparable health indicators at the European level on key aspects of the population's health conditions, the use of healthcare services, and health determinants. The Eurostat methodological manual[1] provides all the recommendations and instructions to best implement the survey. Italy has selected modules on the social participation of people with disabilities (Disability module) and the evaluation of received healthcare services (Patient Experience module) among the additional modules.

The sampling design has the usual structure of most social surveys on households carried out by the ISTAT. This design is based on a two-stage sampling design, with stratification of municipalities based on their population size. The 2019 survey design was integrated with the one used for the Master Sample of the Permanent Census. The selected municipalities are a sub-sample of the 2850 municipalities present in the Master Sample selected for the 2018 Census round. The second-stage units are households, randomly selected (*i*) from the population registers for sample municipalities with fewer than 1000 inhabitants and (*ii*) from the list of households selected for the 2018 Permanent Census for sample municipalities with more than 1000 inhabitants. As stated in the Introduction, the final sample consists of approximately 22,800 households in 835 municipalities of various sizes and distributed throughout the national territory. For further details on the adopted sampling strategy, see ISTAT (2021).

The topics covered in the survey relate to three main areas: health status, health determinants, and access to and use of healthcare services. Most sections of the survey modules refer to the population aged 15 and above, as required by the European Regulation.

The estimates produced by the survey are absolute and relative frequencies, referring to households and individuals. The estimates are obtained using a calibrated estimator (Devaud and Tillé, 2019; Deville and Särndal,1992; Särndal, 2007), with benchmark constraints given by:

[1] https://ec.europa.eu/eurostat/documents/3859598/8762193/KS-02-18-240-EN-N.pdf/5fa53ed4-4367-41c4-b3f5-260ced9ff2f6?t=1521718236000

- distribution in the 21 Italian regions (19 regions plus the provinces of Trento and Bolzano) by sex and seven age groups (0-14, 15-24, 25-44, 45-54, 55-64, 65-74, 75+);
- distribution of the population in the 5 territorial divisions by sex and nine age groups (0-14, 15-24, 25-34, 35-44, 45-54, 55-64, 65-74, 75-84, 85+);
- distribution of the population by citizenship (households of Italian citizens, households of foreign citizens, mixed households).

In particular, the target indicator, SFL, is given by the proportion of people above 15 years of age who have long-term physical, mental, intellectual, or sensory impairments that, when interacting with various barriers, may hinder their full and effective participation in society on an equal basis with others (ISTAT, 2015).

To assess the reliability of estimates, the classification based on the coefficient of variation (CV %) considered by Statistics Canada for the Labour Force Survey is applied. In particular, estimates are classified as follows:

1. estimates publishable without restrictions, CV % $\leq$ 16.5;
2. estimates publishable with caution, 16.5% < CV % $\leq$ 33.3%;
3. estimates not recommended for publication, CV % > 33.3%.

The estimates are based on 2019 EHIS survey data. Table 1 shows the number of estimates falling into the three categories, as well as the number of domains for which direct estimates are not available. The large number (168) of estimates with CV exceeding the threshold for release, added to the unavailability of 21 domains that are out of sample, highlights the need of employing small area estimation methods. These methods allow to enhance the precision of estimates for disaggregated domains by borrowing strength from other areas and exploiting the relationship between the variable of interest and a set of auxiliary variables available at the elementary unit or area level. Due to privacy concerns, integrating information from other sources, such as administrative data with EHIS survey data at the elementary unit level is not feasible. Consequently, only auxiliary information known at the area level can be employed. In this informative context, the applied method is an estimator based on a mixed-effects model defined at the area level, proposed by Fay and Herriot (1979).

Table 1 – *Estimates that can be released, that can be released with warning and estimates that are too unstable to be released, for the indicator SFL – year 2019.*

CV%	Evaluation	Number of estimates
≤ 16.5]	Publishable	87
(16.5; 33.3]	Publishable with caution	152
>33.3	not recommended for publication	168
Not available	Not available	21

3. Small Area Estimation based on a Mixed Area Level Model

Small area estimation based on an area-level mixed model, often referred to as the Fay-Herriot method, is a technique used to estimate the parameters of interest for specific domains (areas) by combining survey data with available auxiliary information at the area level. Let d be the generic small area of interest ($d = 1, 2, \dots, D$), $\hat{\theta}_d$ the direct estimate of the target parameter θ_d related to area *d,* and $\boldsymbol{X}_d$ a set of auxiliary variables known for each area of interest. The area-level mixed model is given by the combination of the following two models:

$$\hat{\theta}_d = \theta_d + e_d$$

$$\theta_d = \boldsymbol{X}_d\beta + u_d$$

where the sampling errors e_d and the area-specific random effects u_d have zero mean. The combination of these two models provides the following mixed-effects model,

$$\hat{\theta}_d = \boldsymbol{X}_d\beta + u_d + e_d, \tag{1}$$

where the random effects u_d are assumed to be independent of the sampling errors e_d, and both are normally distributed. The variance σ_e^2 of the sampling errors is assumed to be known and the other model parameters are estimated by using restricted maximum likelihood method as described e.g. in Rao and Molina (2015, Chapter 5). Denoted with $\hat{\beta}$ and $\hat{u}_d$ the estimate of the fixed effects and the prediction of the area-specific random effects, respectively, the resulting Empirical Best Linear Unbiased Predictor for θ_d can be written as a linear combination of a direct and a synthetic estimator:

$$\hat{\theta}_d^{sae} = \hat{\gamma}_d\,\hat{\theta}_d + (1 - \hat{\gamma}_d)\,\boldsymbol{X}_{\boldsymbol{d}}^{\boldsymbol{T}}\,\hat{\beta} \tag{2}$$

where $\hat{\gamma}_d$ is a shrinkage factor that represents the weight assigned to the direct estimator. In particular, it is given by:

$$\hat{\gamma}_d = \hat{\sigma}_u^2/\hat{\sigma}_e^2 + \hat{\sigma}_u^2, \tag{3}$$

where $\hat{\sigma}_e^2$ is the estimated sampling variance of the direct survey estimate, and $\hat{\sigma}_u^2$ is the estimated variance of the random effects.

4. The application results

The case study's objective is to estimate the SFL indicator at a specific level of disaggregation, defined by the cross-classification of provinces, sex, and two age classes (15-64; 65+), resulting in a total of 428 unplanned domains of interest. The planned domains with a targeted level of accuracy are the regions, having a coefficient of variation (CV) ranging from 5% to 12%. In order to compute estimates for the required level of disaggregation, the mixed area level model is specified using two auxiliary variables available at the area level:

- disability certification, available from INPS, the Italian National Social Security Institute, and
- the hospital attractiveness index, available from the Ministry of Health, that is used to monitor the healthcare service.

Additionally, valuable area level information is gathered from other small area estimation case studies, particularly for poverty estimation, where integration at the unit level between survey data and administrative information was feasible. This supplementary information is accessible through ISTAT's Integrated System of Registers, particularly from the Population Register and the Labour Register integrated with administrative data for income (Baldi et al., 2018). This additional auxiliary information comprises:

- population distribution for 7 age classes;
- population distribution for three education level classes (primary education, secondary education, university degree);
- at risk of poverty rate;
- quintiles of equivalent income at the national, regional, and provincial level;
- population distribution for work income, pension income and capital income grouped in five classes;
- population distribution for four classes according to the average number of working weeks, obtained by dividing the year into quarters.

All these variables' frequencies within the domains of interest have been considered to specify the small area level model.

Figure 1 shows the distribution of direct estimates in the 428 domains of interest. It is evident that the parameter of interest exhibits different intensities in the two age classes, with some overlaps between the upper tail of the estimates concerning the first age class and the lower tail of the estimates concerning the second class. Thus, the distribution of the estimates is characterized by a mixture of two distributions, as depicted in Figure 2. To address this situation, the fixed part of the mixed model is formulated incorporating the interaction of covariates with the two distinct age classes (15-64 and 65+). The final model is chosen through a

stepwise selection of relevant auxiliary information and its interaction with the two age classes. The SAEs of the target variable are calculated using the R package emdi (Kreutzmann et al. 2019), which stands for "Estimating and Mapping Disaggregated Indicators." This package is freely available on the R CRAN platform (https://cran.r-project.org/web/packages/emdi/index.html).

Figure 1 – *Distribution of direct estimates in the domains of interest*.

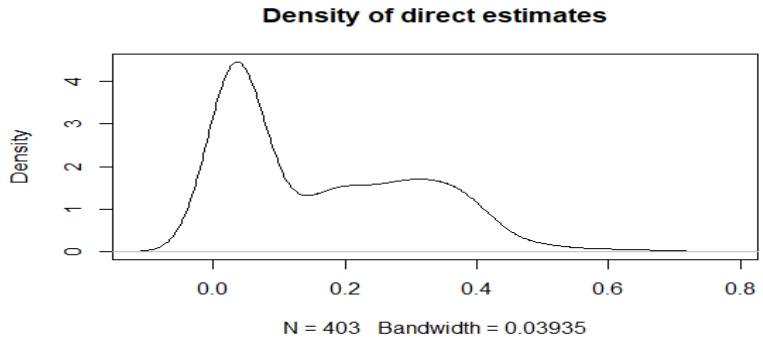
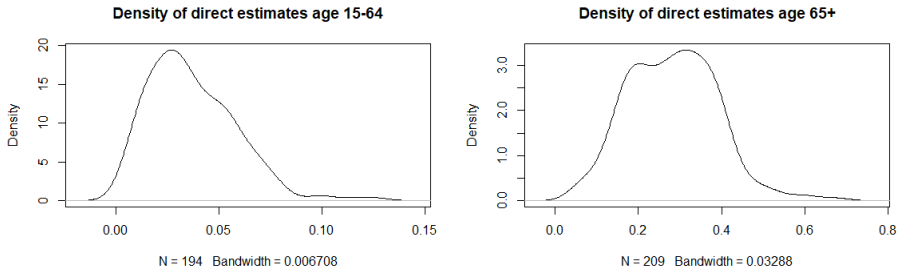


Figure 2 – *Distribution of direct estimates in the domains of interest in the two classes of age.*



The standard Fay-Herriot area-level estimator assumes normality and independence of the error terms. However, in this specific application, this assumption appears to be violated, as shown in Figure 3 which compares the distribution of the realized standardized residuals and of the standardized random effects of the FH model with their normal counterparts. To take this issue into account, the area level model has then been specified on the log-transformation of the direct estimates, and SAEs based on this model have been computed using emdi package. For more details on the log-transformed area-level model, see Kreutzmann et al. (2019). The log-transformation allows for a better suitability of

the fitted area level model to the assumptions of normality of the random error terms, as illustrated in Figure 4.

Figure 3 – *Distribution of the realized standardized residuals and of the standardized random effects of the standard FH model.*

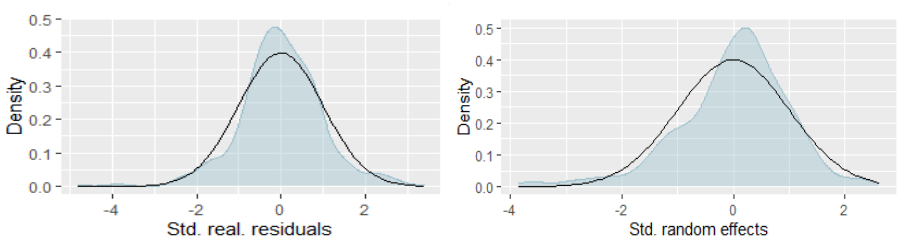
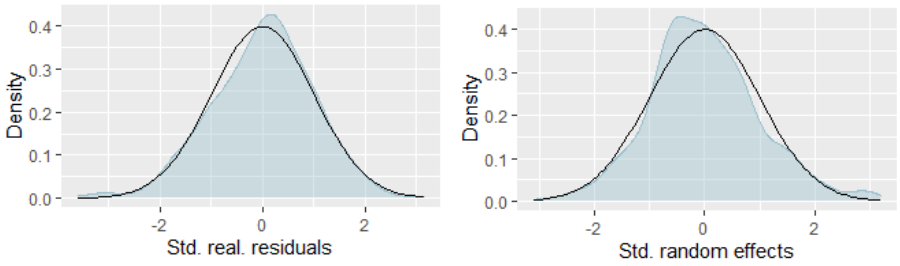


Figure 4 – *Distribution of the realized standardized residuals and of the standardized random effects of the log-transformed FH model.*



As with direct estimates, also their variance estimates can be very unstable, so that smoothing these variance has been considered and used for computing the log-transformed FH SAEs. Assuming that the CV of estimates depends on the area sample size and on the target variable, the smoothing model considered is given by

$$\ln(CV^2(\hat{\theta}_d)) = \beta_0 + \beta_1 \ln(n_d) + \beta_2 \ln(\hat{\theta}_d) \quad (5)$$

Figure 5 clearly illustrates the impact of employing a smoothed set of sampling variance estimates in contrast to the more unstable original estimates: the distribution of the shrinkage factor $\hat{\gamma}_d$ in estimator (2), as defined by expression (3), shows notable improvement with the use of smoothed variances. In fact, by incorporating the smoothed variances, the difference between the shrinkage factors for the two age classes becomes more evident, along with their respective trends with respect to the sample size. This enhancement significantly improves our

understanding of how the shrinkage factor influences the expression of estimator (2) in the determination of the SAEs of interest.

Figure 5 – *Distribution of the estimated shrinkage factor γ_d as a function of the sample size for the log-transformed FH model when the original (left) and the smoothed (right) variance is used to computed SAEs.*

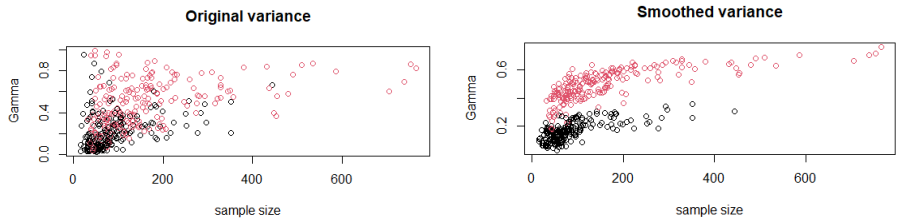


Figure 6 compares the values of the SAEs obtained using the standard FH method and the log-transformed FH method with respect to the direct estimates. The log transformation of data yields estimates that have higher consistency with direct estimates, in contrast to estimates derived from the standard area level model. This highlights the effectiveness of log-transformation in producing more reliable and precise SAEs in this case. This finding is further supported by Figure 7, which clearly shows a significant efficiency gain when applying the FH method to data transformed using a logarithmic function, as opposed to using the original data.

Table 2 displays the distribution of estimates across the three classes of %CV for direct estimates and for the SAEs computed under the standard FH and log-transformed FH models. It can be observed that all FH-LOG estimates have a coefficient of variation below 33%.

Figure 6 – *Direct estimates versus standard FH and LOG transformed FH SAEs.*

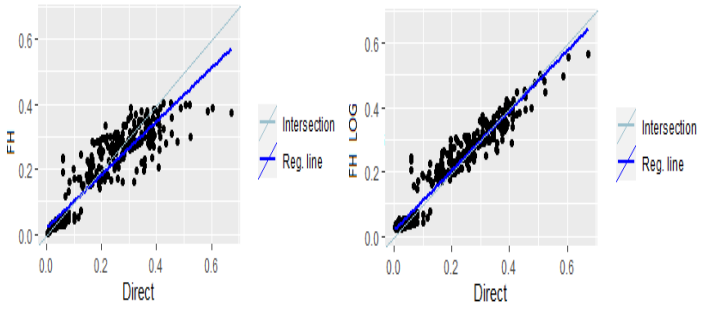


Figure 7 – *Distribution of Coefficient of Variation (CV) of Direct estimates versus standard FH and LOG transformed FH SAEs.*

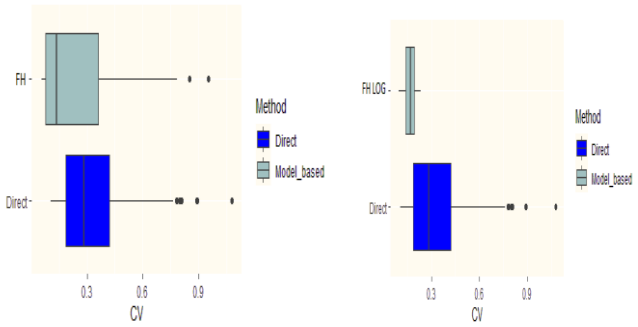
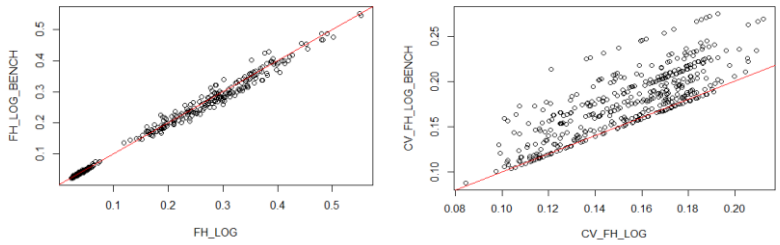


Table 2 – *CV distribution of direct and model based estimates.*

Estimator	CV%			
	< 16.6	*16.6-33.3*	*>33.3*	*Not available*
Direct	87	152	168	21
FH	213	65	150	0
FH LOG	196	232	0	0

Another crucial step is to ensure that all the estimates of the target indicator computed at different levels of disaggregation are consistent. To achieve this, the SAEs were benchmarked to the direct estimates produced for the planned regional domain. This process ensures that SAEs are aligned with the unbiased direct estimates at a regional level, enhancing also in this way the overall accuracy and reliability of the model based SAEs produced for the unplanned domains. The good performance of the FH-LOG estimator in terms of accuracy is further validated in Figure 8. The pictures show a comparison of the distribution of SAEs and their CV, before and after benchmarking. Following the benchmarking procedure, SAEs show small changes compared to the pre-benchmarked estimates, indicating a good fit of the specified model. As expected, the CVs of the estimates after benchmarking slightly increase in comparison to their respective pre-benchmarked counterparts, as the benchmark procedure introduces additional variability due to the adjustments needed to achieve the coherence among estimates.

Figure 8 – *Distribution LOG-transformed FH SAEs (left) and corresponding CV (right) before and after benchmarking.*



5. Conclusions

The application of a small area estimation method based on a mixed area level model with log-transformed data has yielded promising results for the target indicator of Severe Functional Limitation at the required unplanned domains from the European Health Interview Survey. It allows good gains of efficiency of the produced estimates with respect to direct estimates. Nonetheless, several further actions should be implemented to further enhance the small area estimation process. Firstly, as soon as it becomes available, auxiliary information from the ISTAT disability register could be considered. Moreover, it is important to further assess the quality of the SAEs produced, also by means of a process of validation of the estimates carried out by users and thematic experts.

Acknowledgements

We would express our gratitude for the fruitful collaboration of all the members of the sub-working group WP3 for Small Area Estimation (SAE), and in particular to Isabella Corazziari, ISTAT DIRM/DCME/MEB, Lidia Gargiulo and Laura Iannucci from ISTAT SWC, Gaia Bertarelli from the University of Venice, and Francesco Schirripa Spagnolo from the University of Pisa.

References

BALDI C., CECCARELLI C., GIGANTE S., PACINI S., ROSSETTI F. 2018. The Labour Register in Italy: The New Heart of the System of Labour Statistics,

Rivista Italiana di Economia, Demografia e Statistica, VOL. LXXII, No. 2, pp. 95-105.

DEVILLE, J. C., SÄRNDAL, C. E. 1992. Calibration Estimators in Survey Sampling, *Journal of the American statistical Association*, Vol. 87, No .418, pp. 376-382.

DEVAUD, D., TILLÉ, Y. 2019. Deville and Särndal's Calibration: Revisiting a 25-Years-Old Successful Optimization Problem, *Test,* Vol. 28, No. 4, pp. 1033-1065.

FAY, R. E. AND HERRIOT, R. A. 1979, Estimates of Income for Small Places: An Application of James-Stein Procedures to Census Data, *Journal of the American Statistical Association,* Vol. 74, No. 366, pp. 269-277.

KREUTZMANN, A., PANNIER, S., ROJAS-PERILLA, N., SCHMID, T., TEMPL, M. AND TZAVIDIS, N. 2019. The R Package Emdi for Estimating and Mapping Regionally Disaggregated Indicators, *Journal of Statistical Software*, Vol. 91, No. 7, pp. 1-33.

ISTAT. 2015. *Social inclusion of people with functional limitations, impairments or severe chronic diseases, Methodological Note*. Roma: Istituto Nazionale Di Statistica.

ISTAT. 2021. *Condizioni di salute e ricorso ai servizi saniotari in Italia e nell'Unione Europea, Nota metodologica.* Roma: Istituto Nazionale Di Statistica.

RAO J.N.K., MOLINA I. 2015. *Small area estimation*. New Jersey: John Wiley & Sons.

SÄRNDAL, C. E. 2007. Calibration Estimators in Survey Sampling, *Survey Methodology,* Vol. 33, No. 2, pp. 99-119.

Michele D'ALÒ, Istat Dirm/Dcme/Meb, dalo@istat.it
Andrea FASULO, Istat Dirm/Dcme/Meb, fasulo@istat.it
Francesco ISIDORI, Istat Dirm/Dcme/Meb, isidori@istat.it
Maria Giovanna RANALLI, University of Perugia, maria.ranalli@unipg.it

FURTHER DEVELOPMENTS ON THE POWER OF THE DOUBLE FREQUENCY DICKEY FULLER TEST ON UNIT ROOTS

Margherita Gerolimetto, Stefano Magrini

Abstract. In this paper we present some further investigations on the power of the Double Frequency Dickey Fuller test for unit root, recently proposed in literature to capture those situations where the time series might be affected by potential unknown structural breaks, asymmetrically located.
The use of Fourier function to approximate structural breaks has recently received large attention in unit root literature. The idea is that the Fourier approach allows capturing the behavior of a deterministic function form even if it is not periodic, working better than dummy variables, independent of the breaks are instantaneous or smooth and avoiding the problem of selecting the dates and the form of the breaks. The first attempts focused on the adoption of single frequency trig functions. More recently, it has been proposed an approach based on a double frequency in trig functions, which is more likely to capture also breaks that are asymmetrically located. Of this so-called Double Frequency Dickey Fuller test, it has been developed the asymptotic theory and, via simulations, its finite sample properties have been shown with respect to a variety of processes. To the best of our knowledge, however, no results have been presented with respect to the power of the Double Frequency Dickey Fuller test in case of occasional breaks data generating processes.
To address this issue we intend to conduct an extensive Monte Carlo experiment, concentrated on some occasional break data generating processes such as Mean Plus Noise and Markov Switching to evaluate the power of the test to distinguish also among this type of behavior.

1. Introduction

One of the most studied topics in the applied unit root time series literature is whether macroeconomic time series, in particular those considered by Nelson and Plosser (1982) are random walks or stationary processes around a level or a trend. The issue of stochastic versus deterministic trend has important practical policy implications. Until the empirical work of Nelson and Plosser (1982), the general view was that macroeconomic time series were stationary around a deterministic

trend or level (Blanchard, 1981; Barro, 1976). However, after the introduction of Dickey and Fuller's tests for unit root (Dickey and Fuller, 1979, 1981), herafter DF and ADF, Nelson and Plosser (1982) find that with one exception, all historical time series have a unit root. This finding supports the real business cycle hypothesis and goes against the deterministic approach which separates business cycles from trend growth.

The paper by Nelson and Plosser (1982) started a long debate, with subsequent research. Phillips and Perron (1988) depart from the standard DF test assumptions of iid errors and developed a new test (PP test) that is robust to heterogeneity and serial correlation in the errors that has the same limiting distribution as ADF. From a different perspective, Sargan and Bhargava (1983) and Bhargava (1986) suggest tests in the Durbin Watson framework. Following Bhargava (1986), Schmidt and Phillips (1992) proposed a LM (Lagrange Multiplier) test whose power is argued to be larger than DF tests. Kwaitowski et al. (1992) proposed a stationarity test based on Lagrange Multiplier principle to a general error process similar to PP-type test. Leybourne and McCabe (1994) modify the KPSS test to form a stationarity test in the DF-type framework. Another line of research approaches the issue from a Bayesian perspective. In this regard it appears that the results concerning the stationarity of NP data differ with the choice of the prior.

However, as Perron (1989) pointed out, all these tests can be misleading if one does not account for the possibility of structural breaks in the time trend or level. His seminal paper opened an area of research to develop unit root tests that are robust to structural breaks or outliers in the data. This poses a serious problem for applied economists since the number duration and form of structural breaks may not be known. Moreover, detecting the number or the locations of the breaks may in turn cause an unknown pre-testing bias. A complicating factor can be also that a break occurring in a given year sometimes does not display its full impact immediately.

The first studies (Perron, 1989; Zivot and Andrews, 2002; Lee and Strazicich, 2003) use dummies to mimic structural breaks in the series. The drawback of this approach is that it generates too many nuisance parameters. This argument stimulate towards a different set of unit root and stationarity tests. Becker et al. (2006) develop tests which model any structural break of unknown form as a smooth process by means of the Fourier transforms. Several authors, starting from Gallant (1981), show that a Fourier approximation can often capture the behaviour of an unknown function, even if the function itself is not periodic. This testing framework requires only the specification of the proper frequency in the estimating equations thus reducing the number of estimated parameters. This ensures that, compared to dummy-based approaches, the tests have good size and power independently of the time or shape of the break. In this vein, there are recent proposals that generalize the original ideas of Becker et al. (2006), among which Enders and Lee (2012) who

adopt the Fourier transform in a set-up where it is avoided the problem of selecting the dates, number, and form of breaks.

Omay (2015) proposes a test that combines the methodologies of Becker et al. (2006) and Enders and Lee (2012) and considers the use of fractional frequency to improve the fitting. Cai and Omay (2022) propose a double Fourier frequency test that is able to capture breaks that are asymmetrically located.

For all these tests, the literature propose simulation studies to ascertain the size and power properties in finite samples. Most studies (among others, Enders and Lee, 2012; Cai and Omay, 2022) logistic smooth transition autoregressive (LSTAR) processes or exponential smooth transition autoregressive (ESTAR). To the best of our knowledge, none of them considers the case when the time series is generated by occasional break processes such as the Mean Plus Noise (Chen and Tiao, 1990; Engle and Smith, 1999) and Markov Switching (Hamilton, 1989) models that can exhibit a dependence pattern that be difficult to distinguish from a unit root one.

The research question is then to find out whether the Double Frequency Dickey-Fuller based tests have power versus occasional break data generating processes. The structure of the paper is as follows. In the second section, we will present the Double Frequency Dickey Fuller test. In the third section, we focus on occasional break processes. In the fourth section, we will present our Monte Carlo experiment and some conclusions.

2. Double Frequency Dickey Fuller Test

The modification of the DF test to account for a deterministic function d_t moves from the following AR(1) process with a deterministic trend

$$y_t = d_t + \theta y_{t-1} + \varepsilon_t \qquad t = 1, \dots, T \tag{1}$$

where the stationary term ε_t has variance σ^2, d_t is a deterministic function. If d_t is known, model (1) can be directly estimated and, in turn, the unit root hypothesis $H_0: \theta = 1$ can be tested. When d_t is unknown testing for unit root is problematic given the risk of misspecification of d_t. The idea underlying the DF test based on Fourier expansion is that it is often possible to approximate d_t using the Fourier expansions, as in Enders and Lee (2012):

$$d_t = \alpha_0 + \sum_{k=1}^{n} \alpha_k \left(\frac{2\pi kt}{T}\right) + \sum_{k=1}^{n} \beta_k \left(\frac{2\pi kt}{T}\right) \qquad n \leq T/2 \tag{2}$$

where n represents the number of cumulative frequencies included in the approximations and k represents a particular frequency. It is interesting to observe

that in the absence of a nonlinear trend, all values $\alpha_k = \beta_k = 0$, so that the usual Dickey Fuller specification appears. Usually the number of frequencies n should be kept small to avoid overfitting; in particular, in the original idea of Enders and Lee (2012), the Fourier approximation is adopted for a single frequency (n=1) as follows

$$d_t = \sum_{i=0}^{1} c_i t^i + \alpha \sin\left(\frac{2\pi kt}{T}\right) + \beta \cos\left(\frac{2\pi kt}{T}\right) \quad (3)$$

that includes, via the first term in the sum where $i = 0,1$, both the intercept and the trend plus intercept versions and it also approximates, via the sinusoidal waves the smooth breaks. In expression (3), k is the frequency to be determined over a pre-given interval. However, as pointed by Omay (2015) the breaks caused by sudden geo political events and financial crisis are stochastically distributed and asymmetrically located. Following this logic, Cai and Omay (2022) relax the assumption that the frequency is identical and propose a more general set up where:

$$d_t^{Dfr} = \sum_{i=0}^{1} c_i t^i + \alpha \sin\left(\frac{2\pi k_s t}{T}\right) + \beta \cos\left(\frac{2\pi k_c t}{T}\right) \quad (4)$$

and within the framework of a DF unit root test, the model with optimal frequencies k_s and k_c is (Double Frequency Dickey Fuller, DFDF herafter):

$$y_t = \sum_{i=0}^{1} c_i t^i + \alpha \sin\left(\frac{2\pi k_s t}{T}\right) + \beta \cos\left(\frac{2\pi k_c t}{T}\right) + \theta y_{t-1} + \varepsilon_t, \quad (5)$$

and the test statistic for the unit root hypothesis $H_0: \theta = 1$ is:

$$\tau^{Dfr} = \frac{T(\hat{\theta}-1)}{\sqrt{T^2 \delta_{\hat{\theta}}^2}} \quad (6)$$

where $\hat{\theta}$ and $\delta_{\hat{\theta}}^2$ are OLS estimators of θ and standard errors. The asymptotic distribution of the test statistic τ^{Dfr} only depends on the frequencies k_s and k_c and the critical values are tabulated (Cai and Omay, 2022, table 1).

If a nonlinear trend is not actually present in the data, a standard unit root test, such as DF or ADF, is more powerful and there is no need of Fourier terms. So, before adopting the DFDF test with a predetermined frequency pair $(k_s,\ k_c)$, it recommended to test $H_0: linearity$ versus $H_1: non\ linearity$ via an adjusted F test:

$$F^{Dfr}(k_s, k_c) = \frac{\frac{SSR_0 - SSR_1(k_s,k_c)}{2}}{\frac{SSR_1(k_s,k_c)}{T-q}} \quad (7)$$

SSR_0 and $SSR_1(k_s, k_c)$ represent sum of squared residuals without and with Fourier components, q is the number of regressors. If H_0 is rejected, a functional form with Fourier components is suggested.

Selecting the double frequency is done with a grid search to find the optimal pair (k_s^*, k_c^*), through the minimization of the SSR. This leads to the modified F test:

$$F^{Dfr}(k_s^*, k_c^*) = max_{(k_s,k_c)} F^{Dfr}(k_s, k_c)$$

where $(k_s^*, k_c^*) = argmax F^{Dfr}(k_s, k_c)$. Minimizing SSR is equivalent to maximizing the F^{Dfr} test statistic under the condition of maximum frequency k_{max} and a searching precision of Δk (critical values tabulated).

3. Occasional break processes

For all the above mentioned tests based on the Fourier approximation, the literature propose simulation studies to ascertain the size and power properties in finite samples. To the best of our knowledge, none of them considers occasional break processes in mean, such as Mean Plus Noise and Markov Switching whose patterns can be sometimes non easily distinguishable from strong dependent ones (Granger and Hyung, 2004).

The idea of occasional break processes is that the number of breaks that can occur in a specific period of time is somehow bounded. More formally, we assume, that the probability of breaks, p, converges to zero slowly as the sample size increases, i.e. $p \to 0$ as $T \to \infty$, yet $lim_{T\to\infty} Tp$ is a non-zero finite constant. This implies that letting p decrease with the sample size, realization tends to have just finite breaks.

The Mean Plus Noise model (Chen and Tiao, 1990; Engle and Smith, 1999) is a binomial model, characterized by sudden changes only

$$\begin{aligned} y_t &= m_t + \varepsilon_t, \\ m_t &= m_{t-1} + q_t \eta_t \end{aligned} \tag{8}$$

where ε_t is a noise variable, the occasional level shifts m_t are controlled by two variables q_t (date of breaks) and η_t (size of jump). η_t is an i.i.d. $N(0, \sigma_\eta^2)$ although the normality assumption can be dropped. $q_t$is assumed to be an i.i.d. sequence of Bernoulli random variables such that $P(q_t = 1) = p$.

The structural changes might also occur gradually, in this case a Markov switching model (Hamilton, 1989) is more appropriate to describe the behaviour of q_t. More in details, it is given s_t a latent random variable that can assume only values

0 or 1 and is assumed to be a Markov chain, with transition probability $p_{ij} = P(s_t = j|s_{t-1} = i)$. Then it is possible to use a switching model for q_t such that q_t=0 when s_t=0 and q_t=1 when s_t=1. In this specification a regime with s_t=1 represents a period of structural change, regardless of the value of s_{t-1}.

4. Monte Carlo experiment

The experiment investigates the performance of the DFDF and the adjusted F test. The DFDF test is applied using the optimal (k_s^*, k_c^*) frequencies identified in the implementation of the F test.

The sample size is T=50,150, 300, the number of simulations is 2000 and we consider the following occasional break data generating processes (DGPs):

1) Mean Plus Noise, where p = 0.005, 0.01, 0.05, 0.1, σ^2 =1, σ_η^2= 0.1
2) Markov Switching, where p, q==(0.95, 0.95); (0.95, 0.99); (0.99, 0.95); (0.99, 0.99), σ^2 =1, σ_η^2= 0.1. The initial state s_1 is generated by a Bernoulli random variable with p=0.5

The percentage of rejection of the null hypothesis of the DFDF and F tests, is an estimate of the power of both tests, that have been implemented in the version with linear trend as well as in the version with intercept only, i.e. constant level. The considered nominal sizes are 5% and 1%.

The results are presented in the set of tables below (Tables 1-4). As we can see, the DFDF test confirms its excellent power properties in both occasional break DGPs even at the smallest sample size. Instead, the performance of the F test is low, in particular for Mean Plus Noise when p is low. This is not surprising, given that the smaller is the probability of jumps, the less the DPG shares nonlinear features. It also must be noticed that the power improves when p grows and in general with the sample size. This same pattern, although at a somewhat less evident extent, characterizes also the Markov Switching DGP.

Overall, our simulations confirm the very good power performance documented in the literature of the DFDF test, but cast some doubts on the power properties of the test F for occasional break DGPs. This issue is very crucial and should be considered with great care, given that the F test is preliminary to the DFDF test, hence a failure of the F test in rejecting the null hypothesis of linearity would imply a wrong use of the standard unit root test in place of the DFDF with consequent further wrong inference.

Table 1 – *Mean Plus Noise, percentage of rejection of null hypothesis (nominal size=5%).*

p	Linear Trend		Constant level	
	Test F	Test DFDF	Test F	Test DFDF
T=50				
0.005	0.002	1	0.004	1
0.01	0	1	0.004	1
0.05	0.002	1	0.008	1
0.1	0.004	1	0.028	1
T=150				
0.005	0	1	0.038	1
0.01	0.012	1	0.044	1
0.05	0.06	1	0.26	1
0.1	0.176	1	0.47	1
T=300				
0.005	0.02	1	0.12	1
0.01	0.072	1	0.216	1
0.05	0.4	1	0.738	1
0.1	0.614	1	0.874	1

Table 2 – *Markov Switching, percentage of rejection of null hypothesis (nominal size=5%).*

p,q	Linear Trend		Constant level	
	Test F	Test DFDF	Test F	Test DFDF
T=50				
0.95,0.95	0.042	1	0.09	0.984
0.95,0.99	0.014	1	0.04	0.994
0.99,0.95	0.056	1	0.15	0.976
0.99,0.99	0.034	1	0.094	0.994
T=150				
0.95,0.95	0.67	1	0.744	0.994
0.95,0.99	0.652	1	0.75	1
0.99,0.95	0.686	1	0.714	0.996
0.99,0.99	0.686	1	0.746	0.99
T=300				
0.95,0.95	0.92	1	0.904	0.998
0.95,0.99	0.952	1	0.928	0.998
0.99,0.95	0.918	1	0.92	0.994
0.99,0.99	0.94	1	0.926	0.998

Table 3 – *Mean Plus Noise, percentage of rejection of null hypothesis (nominal size=1%).*

p	Linear Trend		Constant level	
	Test F	Test DFDF	Test F	Test DFDF
T=50				
0.005	0	1	0	1
0.01	0	0.998	0	1
0.05	0.002	0.992	0.002	1
0.1	0.	0.998	0.008	1
T=150				
0.005	0	1	0.014	1
0.01	0.002	1	0.018	1
0.05	0.03	1	0.13	1
0.1	0.086	1	0.286	1
T=300				
0.005	0.012	1	0.072	1
0.01	0.044	1	0.144	1
0.05	0.28	1	0.618	1
0.1	0.492	1	0.78	1

Table 4 – *Markov Switching, percentage of rejection of null hypothesis (nominal size=1%).*

p,q	Linear Trend		Constant level	
	Test F	Test DFDF	Test F	Test DFDF
T=50				
0.95,0.95	0.008	0.984	0.09	0.984
0.95,0.99	0.006	1	0.04	0.994
0.99,0.95	0.012	1	0.15	0.976
0.99,0.99	0.01	1	0.094	0.994
T=150				
0.95,0.95	0.498	1	0.744	0.994
0.95,0.99	0.482	1	0.75	1
0.99,0.95	0.518	1	0.714	0.996
0.99,0.99	0.526	1	0.746	0.99
T=300				
0.95,0.95	0.862	1	0.904	0.998
0.95,0.99	0.878	1	0.928	0.998
0.99,0.95	0.846	1	0.92	0.994
0.99,0.99	0.88	1	0.926	0.998

References

BECKER R., ENDERS W., LEE J. 2006. A Stationarity Test in the Presence of an Unknown Number of Smooth Breaks, *Journal of Time Series Analysis*, Vol. 27, pp. 381-409.

BHARGAVA A. 1986. On the Theory of Testing for Unit Roots in Observed Time Series, *Review of Economic Studies*, Vol. 53, pp. 369-384.

BARRO R. 1976. Rational Expectations and the Role of Monetary Policy, *Journal of Monetary Economics,* Vol. 2, pp. 1-32.

BLANCHARD O. 1981. What is left of the multiplier accelerator? *American Economic Review,* Vol. 71, pp. 150-154.

CAI Y., OMAY T. 2022. Using Double Frequency in Fourier Dickey-Fuller Unit Root Test, *Computational Economics*, Vol. 59, pp. 445-470.

CHEN C., TIAO G.C. 1990. Random Level-Shift Time Series Models, ARIMA Approximations, and Level-Shift Detection, *Journal of Business & Economic Statistics*, Vol. 8, pp. 83-97.

DICKEY D., FULLER W. 1979. Distribution of the Estimators for the Autoregressive Time Series with a Unit Root, *Journal of the American Statistical Association*, Vol. 74, pp. 427-431.

DICKEY D., FULLER W. 1981. Likelihood Ratio Statistics for Autoregressive Distribution of the Estimators for the Autoregressive Time Series with a Unit Root, *Journal of the American Statistical Association*, Vol. 74, pp. 427-431.

ENDERS W., LEE J. 2012. A Unit Root Test Using a Fourier Series to Approximate Smooth Breaks, *Oxford bulletin of Economics and Statistics*, Vol. 74, pp. 574-599.

ENGLE RF., SMITH AD. 1999. Stochastic Permanent Breaks, *Review of Economics and statistics*, Vol. 81, pp. 553-574.

GALLANT A.R. 1981. On The Bias in Flexible Functional Forms and an Essentially Unbiased Form: the Fourier Flexible Form, *Journal of Econometrics*, Vol. 15, pp. 211-245.

GRANGER C.W.J., HYUNG N. 2004. Occasional Structural Breaks and Long Memory with an Application to the S&P500 Absolute Returns, *Journal of Empirical Finance*, Vol. 11, pp. 399-421.

HAMILTON J.D. 1989. A New Approach to the Economic Analysis of Nonstationary Time Series and the Business Cycle, *Econometrica*, Vol. 57, pp. 357-384.

KWAITOWSKI D., PHILLIPS P., SCHMIDT P., SHIN Y. 1992. Testing the Null Hypothesis of Stationarity Against the Null Hypothesis of a Unit Root, *Journal of Econometrics*, Vol. 54, pp. 159-78.

LEE J., STRAZICICH M.C. 2003. Minimum Lagrange Multiplier Unit Root Test with Two Structural Breaks, *Review of economics and statistics*, Vol. 85, pp. 1082-1089.

LEYBOURNE S.J., McCABE B.P.M. 1994. A Consistent Test for A Unit Root, *Journal of Business & Economic Statistics*, Vol. 12, pp. 157-66.

NELSON C., PLOSSER C. 1982. Trends and Random Walks in Macroeconomic Time Series, *Journal of Monetary Economics,* Vol. 10, pp. 139-162.

OMAY T. 2015. Fractional Frequency Flexible Fourier Form to Approximate Smooth Breaks in Unit Root Testing, *Economics letters*, Vol. 134, pp. 123-126

PERRON P. 1989. The Great Crash, the Oil Price Shock and the Unit Root Hypothesis, *Econometrica*, Vol. 57, pp. 1361-401.

PHILLIPS P., PERRON P. 1988. Testing For a Unit Root in Time Series Regression, *Biometrika*, Vol. 75, pp. 335-346.

SARGAN J., BHARGAVA A. 1983. Testing Residuals from Least Squares Regression for being Generated by the Gaussian Random Walk, *Econometrica*, Vol. 51, pp. 153-174.

SCHMIDT P., PHILLIPS P. 1992. LM Tests for a Unit Root in the Presence of Deterministic Trends, *Oxford Bulletin of Economics and Statistics*, Vol. 54, pp. 257-287.

ZIVOT E., ANDREWS D. 1992. Further Evidence on the Great Crash, the Oil Price Shock and the Unit Root Hypothesis, *Journal of Business and Economic Statistics*, Vol. 10, pp. 251-270.

MARGHERITA GEROLIMETTO, Università Ca' Foscari Venezia, Dipartimento di Economia, margherita.gerolimetto@unive.it

STEFANO MAGRINI, Università Ca' Foscari Venezia, Dipartimento di Economia, stefano.magrini@unive.it

THE ROLE OF STATISTICS IN SHAPING THE TERRITORY TO ADDRESS DEMOGRAPHIC DECLINE AND SUPPORT DEVELOPMENT. THE CASE STUDY OF SICILY

Bianchino Antonella, Camisasca Michele, Dolce Alberto, Lasco Federico

Un territorio non è un semplice agglomerato di case, uffici, strade, industrie, semafori, automezzi ecc. È soprattutto un concetto "identitario", uno spazio allo stesso tempo fisico e simbolico, in grado di generare senso di appartenenza nelle persone che ci vivono e attrazione per le altre
Antonio Romano

Abstract: This paper highlights the significance of generating territorial statistics as a strategic tool for cohesion-oriented policies. Cohesion policy aims to reduce territorial disparities and requires relevant data. Integrating the territorial dimension in policy models is crucial to address demographic and environmental goals. An operational experiment in Sicily focuses on generating such statistics for better programming. The post-World War II era saw extensive research on local public goods and services, crucial for achieving territorial cohesion. Historical reflections and approaches from the 1970s emphasize optimal levels of government. Recent policies focus on organizational structures for cohesive capacity. Integrating these analyses and utilizing statistical information enhances policy planning for territorial cohesion. Capacity building for Cohesion Policy has evolved since 1988, recognized as crucial for development and sustainability. Some contributions have analyzed limited absorption capacity of financial resources for cohesion investments, proposing organizational schemes for optimal division of labor among actors. The Sicilian 2021-2027 ERDF Regional Programme was inspired by similar models. Statistical data serves as a robust basis for multi-level implementation processes. Sicily identified 29 functional areas based on official indicators, fostering inter-municipal cooperation, population, and economic growth. The "Knowledge and Identity" project empowers strategic decision-making and long-term planning to counter depopulation and support territorial development. ISTAT's role as an official statistical producer fulfills social responsibility and supports effective local development strategies. The bottom-up approach empowers local communities for effective growth prospects.

1. Territorial statistics in the era climatic and demographic crisis

Rex Stout, known to the most as the "literary" creator of detective Nero Wolfe, said that "there are two kinds of statistics, the kind you look up and the kind you make up".

The story we want to tell in this essay concerns the importance of statistics "to be maked up" so that they can constitute a strategic tool for the construction of cohesion-oriented policies in multilevel governance contexts.

As reported by the portal of the Department for Cohesion Policies of the Italian Presidency of the Council of Ministers: "cohesion policy has the aim of increasing opportunities for economic and social development to help reduce the gaps and disparities between territories, acting in particular in less developed areas and for the most fragile communities and people. It is based both on the Treaty on the Functioning of the European Union (art. 174) and on the Italian Constitution (art. 3 paragraph 2 and art. 119 paragraph 5), which require special interventions to promote harmonious development and to remove economic and social imbalances.

It is a policy with medium-term objectives which involves various levels of government (central and local) and attributes a formal and fundamental role to economic and social partnership, financing plans, programs and individual projects owned by both central, regional or local authorities"[1]. The Italian Constitution specifies that "to promote economic development, cohesion and social solidarity, to remove economic and social imbalances, to encourage the effective exercise of individual rights, or to provide for purposes other than the normal exercise of their functions, the State allocates additional resources and carries out special interventions in favor of specific Municipalities, Provinces, Metropolitan Cities and Regions"[2].

Planning a medium-term policy aimed at reducing the gaps between different territorial aggregations (the Italian Constitution provides for four levels of government with a territorial dimension, without counting the island dimension introduced in paragraph 6 of art. 119), requires knowing how much necessary to implement interventions capable of acting on the variables that the Constitution itself identifies. An adequate planning exercise therefore requires that the entire kit of necessary cognitive tools, endowed with the adequate degree of significance, be available at each relevant territorial aggregation level. It is evident that the aggregation of the available information and their comparability for each possible territorial aggregation are indispensable tools for qualifying needs and identifying resources necessary to reduce levels of inhomogeneity whose very conceptual dimension is both an object of preliminary recognition and a policy objective, to be set and ongoing monitored.

The need to adopt this "drill-in, drill-out" approach to territorial statistics becomes indispensable where short-medium term development and cohesion policies aim to act in contexts characterized by demographic decline and costs deriving from changes in the climate model. The exogenous nature of the environmental and demographic variables, in the short and medium-term policy models, can be reviewed if, in the

[1] https://politichecoesione.governo.it/it/la-politica-di-coesione/ .
[2] Constitution of the Italian Republic (art. 119 paragraph 5).

models themselves, the optimal territorial dimension of the policies is conceived as an endogenous transmission variable: population movements, the organization of work and residences, the permeability of water and sewage infrastructures and so on, act on the nature and scale of the climatic and demographic impact, if the territorial dimension of the intervention is adequately identified.

Addressing demographic and environmental targets needs a new way of thinking about policy models, in which the optimal territorial dimension (levels, layers and boundaries) is a dynamic and endogenous dependent variable. The memorandum of understanding[3] signed in July 2022 by the Presidents of ISTAT and the Sicilian Region launches an operational experimentation of this endogenous role of the territorial dimension of the policy models and the territorial component of the 2021-2027 ERDF Regional Programme. The methods and results of the experiment "to make-up" territorial statistics in Sicily for a better programming of the next decade are the subject of the following pages.

2. Territorial dimension of policy: from static approach to organizational governance

Since the years following the Second World War, the local nature of the creation of public goods and the management of related services has been the target of a close attention from public policy analysts, economists and statisticians. The analysis tools built to support the Territorial Cohesion Policy cannot ignore the impact of the rationale underlying the economic analysis of local public goods and services. The goal of homogenizing the levels of development and opportunities for access to growth between territories and communities passes through an understanding of the economic, social and political mechanisms underlying the creation and management of local public goods and services systems.

The contribution of public policy analysts starts from the founding reflection of Werner Hirsh, who in 1959 in the Review of Economics and Statistics published was the first to address the issue of how to accompany the growth processes that emerged in metropolitan areas with suitable public policies. The approaches developed in the 1970s (for all, Oates 1972) are oriented towards the optimal dimension of the level of government which must guarantee the production and management of public goods for territories and communities which are starting to present different levels of development. More recently, with the expansion, especially at the supranational level,

[3] MoU on the subject of support for social, economic and environmental statistical analyzes for the programming of the unitary cohesion policy 2021-2027 of the Sicilian Region, of the related planning, implementation, monitoring and evaluation tools and for the strengthening of the statistical function in associated form, signed on 11 July 2022, given the appreciation of the Regional Government https://www2.regione.sicilia.it/deliberegiunta/file/giunta/allegati/N.334_28.06.2022.pdf).

of Government Capacity Building-oriented policies that intend to support growth and development processes, the issue of local policies has shifted its focus on the organizational structure (and the related strengthening measures) of policies that act unevenly on development processes in order to guide their cohesive capacity. The two lines of analysis do not have adequate levels of integration and comparison.

The attention to the organizational dimension of the governance structures, defined ex-ante, on which the analyzes of capacity building are concentrated, and the importance attributed to the cost structure at different levels of government in the models of optimal level of government, both lead to not adequately consider the induced effects of spillovers and economies between communities and territories with variable levels of unevenness. The result is analysis with little argument to say about the organizational processes that integrate the levels of government, proposing effective and efficient forms of division of labour. But, at the same time, they use ineffectively the growing presence of statistical information with increasing capillarity at a territorial level, to articulate adequate proposals for implementing the policies targeted to territorial cohesion.

3. The Optimal Level of Government: beyond the Black Box

The issue of the territorial dimension of policies was born to answer the question which is the best level of government to manage public goods that have specific characteristics of a local dimension. Theoretical studies based the arguments on the assumption that the average per capita cost of local public goods decreases as the population served increases. This effect depends on the presence of scope economies (a single input can be used for the provision of several goods) and of increasing returns to scale in production technologies.

The literature has also highlighted the presence of economies of density, linked to the advantage of sharing the cost of providing a given service or public good among a greater number of users: as in the typical case of water, electricity or transport networks. At the same time, the literature has highlighted the presence of grounds for the fact that there are costs increasing with the demographic dimension in the production of public goods: the synthesis of citizens' preferences grows in complexity as their number increases. Accountability also becomes more onerous in larger communities.

From the two pioneering contributions of Oates (1972) and Mirrlees (1972), to that of Dixit (1973), up to the most recent work by Ladd in 1992, followed by Dollery and Fleming in 2005 and by Gómez-Reino and Martinez-Vazquez of 2013 modeling has maintained a constant logical structure. Demand and supply factors interact to determine a non-monotonous trend in the relationship between the size of the territory and the community and the efficiency in the production of public goods. The typical U-

shaped cost curve of the demographic and territorial dimension emerges as the synthetic theoretical basis of the arguments on the optimal size of the level of government most suited to the provision of public goods.

The models are built to answer the question of the optimal level of governance. The analysis is developed to provide an single answer, depending on the cost structure of public good/service and the characteristics of the communities/territories considered, as clearly emerges from the empirical analysis and from the theoretical reflection of Sabrina Iommi (Iommi and others, 2013) for the accurate research of IRPET on Italy and Tuscany.

The original theoretical apparatus provides policy making with a *black box* in which the factors determining the effectiveness and efficiency of the selection processes of public goods are homogenized, as well as the organizational dynamics of the administrative machines that implement their creation and management, the determinant role of public finance rule, the absorption capacity of the production systems that actually create public goods and related services, the socio-economic and environmental evolution of the local contexts (communities and territories) in which public goods/services are created and provided.

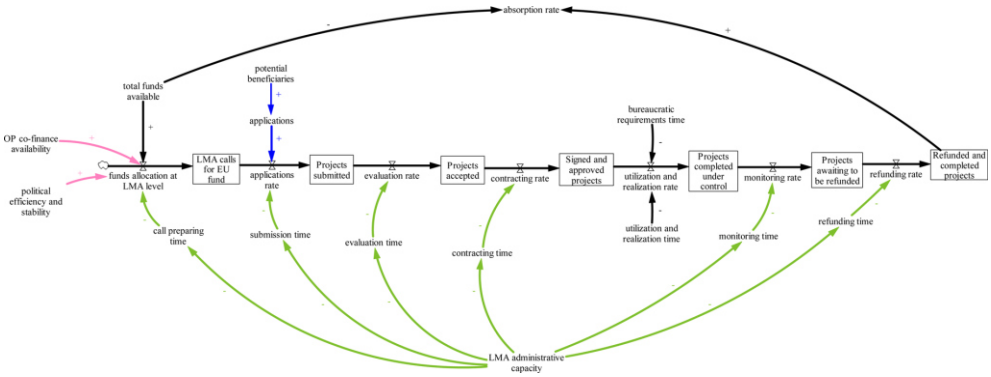
Demographic and environmental dynamics, directly or indirectly, through their redistributive effects, affect the relative relevance of the variables considered by the literature and the direction of their impacts: an analytical framework suitable for policy making must consider the complex structure of the different effects. In other words, the optimal division of functions between the different levels of government (not the unique optimal level) must be the necessary result of any analysis that addresses the issue of the implementation of public investments between territories and communities. Such an analysis suitable for Cohesion Policy cannot ignore a strongly articulated dimension of the statistical knowledge of territories and communities.

4. Optimal organization for policy implementation capacity: balancing power between government levels

The literature on Capacity Building for Cohesion Policy has been evolving steadily since its introduction in EU planning in 1988. Strengthening the operational capacity of actors involved in implementing Cohesion Policy has been recognized as crucial for addressing development gaps, economic growth, citizens' quality of life, and policy sustainability. The literature on capacity building, developed within the debate on development policies supported by international agencies (OECD and UN, in the first place), focuses on the dysfunctions of administrative machines in countries subject to external financing actions. This aspect has been externally oriented in the European debate on cohesion policies, emphasizing the progressive adjustment to common performance standards, organizational models in line with Union's priorities, and

operational alignment with the "Acquis Communautaire," especially regarding directives governing Community Funds for Cohesion Policy. The organizational responsibilities have often been excluded from policy discussions, neglecting the organizational requirements from Community legislation in specific sectors, such as Water Services and Energy.

Figure 1 – *Systemic map of Cohesion Policy implementation 'pipeline'.*



Cunico G. et al. 2021, fig.3 p. 21.

Recently, some contributions have emerged aimed at analyzing the causes of the persistence of a limited level of absorption capacity of financial resources intended for investments for Cohesion, by territorial areas and specific economic sectors (for all, Cunico et al. 2023). The models used have developed organizational schemes, which, albeit still only implicitly, provide organizational analysis of investments implementation which are a prelude to identifying an optimal division of labor among the various actors of multilevel governance. It's reported the Cohesion Policy implementation map drawn up by Cunico and others in 2021, for example only, which graphically summarizes the dynamic model underlying the analysis developed by the authors in different contributions. Beyond the possible more detailed articulations of the process, it is evident that the structure of the model includes political-institutional, administrative and private actors operating at different levels of government in different phases.

Just limiting to considering the project submission phases and the acceptance and approval phase (i.e. the selection of the operations phase), for the implementation of the programs, it is evident how the model is able to analyze the interaction between actors at different levels of government. Similarly, the considerations relating to the definition of the priorities of the Programs and of the specific financial allocations, even if not adequately modeled, lend themselves to an expansion that develops in a unitary model the interactions involving citizens, interest groups, the actors of the

private sectors and of the private social sector, but above all the political decision-makers who interact with public opinion and the administrative machines at the various levels of government. An overlapping layer structure of the proposed system dynamics model constitutes a line of research of great perspective to point to results useful for exploring the organization optimal structures for the division of labor between political, administrative and interest representation institutions operating at different levels of governance.

5. The shape of the territory and the function of statistics in the sicilian experience

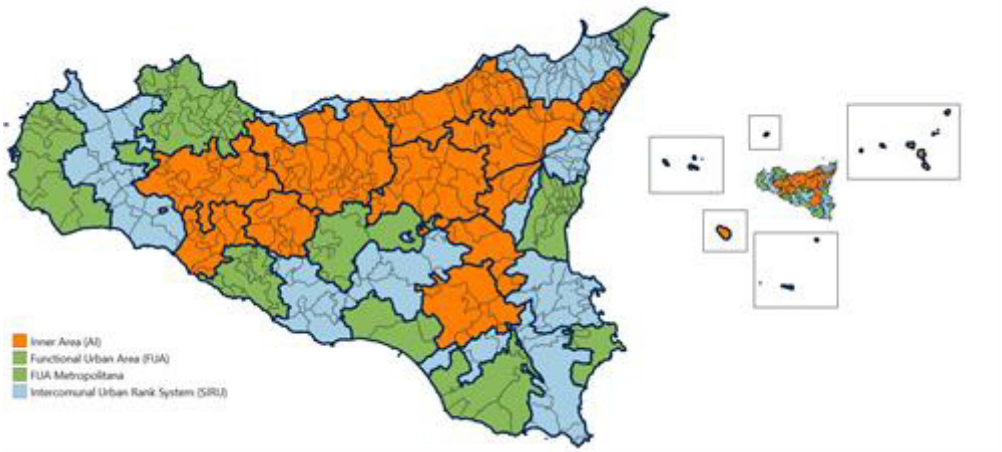
The Sicilian 2021-2027 ERDF Regional Programme's programming phase is inspired by the analytical scheme mentioned earlier. The use of statistical information with high granularity at a territorial and community scale forms the robust methodological foundation for governing multi-level implementation processes, adhering to regulatory requirements and tools for associationism of Local Authorities and Union of Municipalities. During the 2014-2020 programming period, the EU and Italy focused on addressing challenges in urban areas, emphasizing connections between urban and rural regions through the National Strategy for Inner Areas (SNAI). In the 2021-2027 cycle, Sicily divided its territory into three homogeneous zones: Functional Urban Areas (FUA), Inner Areas (AI), and Urban Inter-Municipal Systems (SIRU). These aggregations, based on regulatory indications and the Pact for Italy, ensure transparency and reliability with official statistical data, enabling effective management of programs and resources. Access to territorial statistical data is crucial for designing, implementing, and evaluating public policies, ensuring they are measurable and rely on precise information over time and space.

The identification, measurement, and analysis of these characteristics require stringent statistical criteria and methods based on official statistics due to their regulatory and guaranteeing nature. Additionally, the chosen aggregation solution disregards previous administrative aggregations constructed for other purposes. Its aim is to foster inter-municipal cooperation and the sharing of common functions and services, creating conditions to promote demographic and economic growth and improve the quality of life for residents.

The National Institute of Statistics is fully committed to expanding the offering of territorial statistics on the main environmental and socio-economic phenomena in the country, providing punctual and accurate information that reflects the dynamics and transformations of Italian society. Analytical and continually updated knowledge of the social, economic, demographic, and environmental conditions of the territory is fundamental for good governance and the development and growth policies of the territory itself. Only through correct and in-depth knowledge of the local reality can

informed and conscious decisions be made, directing policies towards concrete and lasting outcomes. In the 2021-2027 cycle, in Sicily, within the framework of territorial policies, the aggregation of territories was based on characteristics aimed at ensuring or at least promoting the achievement of specific objectives. These objectives include internal homogeneity for Inner Areas and functionality for Urban Areas and SIRUs (Single Islands Rural Areas).

Figure 2 – The Sicilian Territorial Aggregations for the 2021-2027 programming cycle



The construction of these new territories has inevitably raised the fundamental issue of lacking knowledge and identity for these 29 Areas, whose consolidation has become the driving force behind strategic decisions for their revitalization. Through the "Knowledge and Identity" project, overseen by ISTAT and the Sicilian Region and formalized through a memorandum of understanding signed in 2022 the statistical function has been identified as the propelling engine for these new areas to acquire in-depth and necessary knowledge of their territories. This knowledge is aimed at defining an optimal long-term strategy and promoting countermeasures against demographic decline and territorial development. The ultimate goal was to move beyond mere data "about" territories and progress towards data "for" and "with" the territories themselves.

At the core of the project, from the construction of the Areas to the planning and implementation of policies and resources, lies statistical information, which plays a fundamental role in supporting decisions in the social and economic realms. It aids in defining priorities and shaping the trajectories of territorial and public administration development. Statistical information is crucial for capturing transformations in the production and environmental systems, while also providing a comprehensive understanding of citizens' life paths and the services dedicated to them. For the new

territorial aggregations, a sound understanding of statistical information is crucial. It aids in developing strategic and operational guidelines for the Regional Programme and ensuring an optimal process of programming, implementation, and monitoring. This facilitates coherent identification of needs and related objective formulation and offers valuable support in selecting the most effective implementation tools, including monitoring and evaluation systems.

The project aimed to empower competent regional offices and territorial coalitions by facilitating their acquisition of significant knowledge and fostering a strong identity perception. This enabled a more informed and effective approach in defining strategies, programs, plans, and agreements to optimize resource utilization. A comprehensive analysis was conducted on the characteristics and needs of the regional territory and specific areas within the territorial coalitions, comprising various Local Authorities. This process involved utilizing a set of relevant indicators to determine priorities, action plans at the local level, and to monitor and evaluate the impacts of the implemented policies. As a result of this indicator framework, an editorial series of "Territorial Area Dossiers" was created to enable new territories to gain knowledge and identity for conscious and effective resource planning, programs/agreements, and the development of various Territorial Strategies (ERDF Sicily 2021-2027).

Another vital objective of the Sicilian experience was to strengthen the independent capabilities of competent regional offices and territorial structures engaged in conducting surveys, data collection, processing, dissemination, and storage of statistical data related to various aspects of territorial cohesion policy. This objective led to the development and sharing of models for producing and utilizing integrated databases. These databases were derived from the integration of statistical and administrative sources and provided valuable insights. The collaboration with institutions such as the National Institute of Statistics (ISTAT), the Sicilian Region, Local Authorities, and other public and private entities facilitated the availability of these data.

6. Some concluding remarks

The identification of homogeneous territories, based on appropriate criteria such as aggregations of municipalities with specific characteristics, facilitates and supports their involvement in both Italian and international political agendas. While it may seem straightforward to direct public policies and resources towards either densely populated or more disadvantaged areas, which represent the two extremes of the territorial continuum ranging from purely urban to strongly rural, it is more complex to identify other territories that do not belong to either of these categories.

Within the ISTAT-Regione Siciliana experience, ISTAT fully fulfills its implicit mandate or social responsibility as the official producer of statistical information. It responds to the increasing need for information in a complex and articulated society

and provides informative support to public decision-makers and, thus, the entire community in a democratic system (Notarstefano, 2020). On the other hand, the Regione Siciliana fulfills its duty to support the co-design of Territorial Strategies as required by European regulations, overcoming local shortcomings in statistical expertise and local planning, resulting in a reduction of drafting times from several years in the 14-20 cycle to a few months in the current cycle.

Through this experience, accordingly with the bottom-up construction of strategies and the local vision, the shared identification of specific needs and opportunities for each community has become possible, as well as their development and growth potential, aiming at building effective local development strategies in the short and long term. The main results of this experience include a significant increase in the connection between cohesion policies and local development objectives, greater local governance and knowledge capacity of cohesion policies, strengthening of functions shared by Territorial Authorities, structural reinforcement of Sicilian Unions (with a higher average number of municipalities and population), identification of a set of municipal indicators useful for the tasks of planning, programming, and management of Local Authorities, consistent with Cohesion Policies.

It also has provided analytical support for the construction of the Territorial Strategy for each Area, identifying needs and criticalities to address through public spending, offers transparent and objective information on their territory and needs to all residents, allows for correct and shared evaluation of the impacts of public action, and disseminates methods and objective criteria for information sharing and the construction of territorial strategic visions.

Acknowledgements

A special thanks is extended to all the people involved in the project "Knowledge and Identity," particularly: the editors of the "Territorial Area Dossiers" series, Alberto Dolce, Rosario Milazzo, Agata Madia Carucci, and Giuseppe Lecardane, under the supervision of Federico Lasco (Regione Siciliana) and Antonella Bianchino (ISTAT). We also want to express gratitude to the ISTAT Working Group, led by Agata Maria Madia Carucci and Giuseppe Lecardane, and comprising Cira Acampora, Beniamino Barile, Salvatore Coppola, Daniela Fusco, Maria Teresa Iuliano, Valeriana Leporanico, Maria Antonietta Liguori, Maria Rosaria Mercuri, Roberto Antonello Palumbo, Alessandra Rodolfi, and Salvatore Vassallo. Additionally, our thanks go to the Regione Siciliana Working Group, "Territorial Policies of Regione Siciliana for the 2021-2027 cycle," which was established in connection with the programming activities of the European Regional Development Fund 2021-2027. This group, coordinated by Domenico Spampinato (NVVIP) and Vincenzo Falletta (DRP) and composed of members from NVVIP and the Department of Programming Area 8, Urban and Territorial Development Planning and Management (DRP), includes Pietro Barbera, Marco Consoli, Alberto Dolce, Maria Teresa Giuliano, Elisabetta Mariotti, Rosario Milazzo, and Ornella Pucci.

References

AIVAZIDOU, E., CUNICO, C., MOLLONA, E. 2020. Beyond the EU structural funds' absorption rate: How do regions really perform ?, Economies, Vol. 8, No. 3.

ANCI-IFEL, "L'Italia delle città medie" in "i Comuni – Quaderni di analisi" from Centro Documentazione e Studi Comuni Italiani, 2013.

ARMAO G., DOLCE A., LASCO F., MILAZZO R., SPAMPINATO D. 2022. The estimation of the costs of insularity through a regressive econometric model applied to Sicily, The Italian Journal of Economic, Demographic and Statistical Studies, Vol. 76, No. 4, pp. 4-12.

BRUNAZZO, M. 2016. The history and evolution of cohesion policy. In S. Piattoni, & L. Polverari (Eds.), Handbook on cohesion policy in the EU (pp. 17–35). Edward Elgar Publishing. https://doi.org/10.4337/9781784715670.00014.

CONSIGLIO ITALIANO PER LE SCIENZE SOCIALI, "Società e territori da ricomporre Libro bianco sul governo delle città italiane", 2011

CORTE DEI CONTI. 2020. Relazione annuale 2020-2021-2022: I rapporti finanziari con l'Unione Europea e l'utilizzazione dei Fondi comunitari, Presidenza del Consiglio dei Ministri

CUNICO A., AIVAZIDOU E., MOLLONA E. 2021. Beyond financial proxies in Cohesion Policy inputs' monitoring: A system dynamics approach, Evaluation and Program Planning, No. 89, 101964.

CUNICO A., AIVAZIDOU E., MOLLONA E. 2023. Investigating Supply and Demand in European Cohesion Policy: Micro-Foundations of Macro-Behaviours, Journal of the Knowledge Economy, https://doi.org/10.1007/s13132-023-01430-6

DIXIT A. 1973. The Optimum Factory Town, Bell Journal of Economics and management Science, Vol. 2, No. 4, pp. 637-651.

DOLLERY B., FLEMING E. 2005. A Conceptual Note on Scale Economies, Size Economies and Scope Economies, Australian Local Governments, Working Paper Series in Economics, No. 6, University of New England School of Economics.

FORMEZ, Dossier regionale -Regione Autonoma Sicilia, programmazione 2021-2027", 2021

GÓMEZ-REINO J.L., MARTINEZ-VAZQUEZ J. 2013. An international perspective on the determinants of local governments fragmentation, in LAGO-PEÑAS S. and MARTINEZ-VAZQUEZ J. (Eds.), The Challenge of Local Government Size. Theoretical Perspectives, International Experience and Policy Reform, by Edwar Elgar.

IOMMI S., LATTARULO P., MARINARI D. 2013. Dimensioni dei governi locali, offerta di servizi pubblici e benessere dei cittadini, Firenze, Studi e approfondimenti IRPET.

ISMERI EUROPA 2023. Inquadramento strategico del tema della capacità amministrativa in Italia nell'ambito d'azione dell'OT11, Scuola Nazionale dell'Amministrazione, Roma - 6 giugno 2023.

LADD E., 1992. Population Growth, Density and the Costs of Providing Public Services, Urban Studies, Vol. 2 No. 29, pp. 273-295.

MANESTRA S., MESSINA G., PETA A. 2018. L'Unione (non) fa la forza ? Alcune evidenze preliminari sull'associazionismo comunale in Italia, Questioni di Economia e Finanza (Occasional Papers), n. 452.

MIRRLEES J.A. 1972. The Optimum Town, Swedish Journal of Economics, Vol. 1, No 74, pp. 114-135.

NOTARSTEFANO G., Statistica e politiche pubbliche, in Politiche pubbliche. Analisi e valutazione, di Antonio La Spina· 2020

OATES W. E. 1972. Fiscal Federalism, New York, Harcourt Brace Jovanovich.

OCSE, Applying the Degree of Urbanization A Methodological Manual To Define Cities, Towns And Rural Areas For International Comparisons, , 2021 edition.

REGOLAMENTO DISPOSIZIONI COMUNI (RDC) UE 1060 del 2021 e Regolamento UE FESR1058 del 2021

Antonella BIANCHINO, Istituto Nazionale di Statistica, bianchin@istat.it
Michele CAMISASCA, Istituto Nazionale di Statistica, michele.camisasca@istat.it
Alberto DOLCE, Istituto Nazionale di Statistica, dolce@istat.it
Federico A. LASCO, Agenzia di Coesione Territoriale
federico.lasco@agenziacoesione.gov.it

REBUILDING A PSEUDO POPULATION REGISTER FOR ESTIMATING PHYSICAL VULNERABILITY AT THE LOCAL LEVEL: A CASE STUDY OF SPATIAL MICRO-SIMULATION IN SONDRIO[1]

Alberto Vitalini, Simona Ballabio, Flavio Verrecchia

Abstract. A wide range of user groups, ranging from policy makers to media commentators, is increasingly seeking more detailed spatial information on health-related topics. This information is needed to gain a better understanding of their communities, more effectively allocate resources, and plan activities and interventions in a more efficient manner. However, due to the sensitivity of the topic of health, it can be challenging to obtain detailed or significant local data. To meet this need, small area estimation (SAE) methodologies are popular as a means of providing spatially detailed insights. Among the various SAE methodologies available, static spatial microsimulation has enabled the simulation of previously unknown variables, such as physical vulnerability, smoking, alcohol consumption, and obesity at the municipal level. This paper presents the initial results of application of static spatial simulation, in order to create synthetic population dataset and estimate "physical vulnerability" of elderly in the municipalities in the province of Sondrio. Physical vulnerability is measured by the prevalence of people who report suffering from chronic or long-term illnesses and who have limitations in their daily activities. A combinatorial optimization (CO) algorithm called simulated annealing, developed at the University of Leeds, is used to simulate the distributions of the "physical vulnerability". This algorithm combines public microdata from the Multiscopo Survey-Aspects of Daily Life-2021 and data from the 2021 Permanent Population Census, which are disseminated in Istat's public databases.

1. Background

With the aging population, the number of elderly individuals affected by chronic illnesses and disabilities is increasing, posing challenges for healthcare systems and policymakers in providing effective care and assistance. It is projected that the aging population will significantly increase in the coming years, putting pressure on

[1] The work is the joint responsibility of the authors. Paragraph 1 is attributed to Flavio Verrecchia, paragraph 2 to Alberto Vitalini, paragraphs 3 and 4 are attributed to Simona Ballabio.

healthcare systems and leading to rising healthcare costs. For example, in Lombardy, it is estimated that the number of individuals aged 65 and older will reach 2.9 million by 2070, surpassing the figure by over half a million compared to 2023 (Istat, 2023).

Identifying older people at risk of physical decline is important as it allows healthcare professionals and policymakers to prioritise care and support for those in greatest need. Frail elderly individuals are more likely to experience negative health outcomes such as falls, hospitalisations, and disabilities (Clegg et al., 2013). In this context, having data on physical vulnerability can help better allocate resources and develop more targeted interventions. For instance, if a particular area exhibits a higher prevalence of physical vulnerability among the elderly, policymakers can allocate more resources to address the issue in that area.

The need for health information at both individual and community levels is crucial, but unfortunately, there is a chronic lack of available data. Official statistics from ISTAT, the Italian National Institute of Statistics, are limited due to data protection regulations that require statistical units to be non-identifiable for data release. This means that information on the entire population at the municipal level is extremely limited, especially in the health domain. Even the sample surveys conducted by ISTAT as part of official statistics are not conclusive for obtaining municipal level information. These surveys are designed to provide reliable information at the national, regional, and geographic levels but may not be sufficient for local objectives. For example, if we wanted to estimate the number of physically frail elderly individuals at the municipal level, we could only calculate a value for the municipalities where there are survey respondents, and the sample size would likely be too small to provide accurate information. To overcome this problem, policymakers could invest in data collection and analysis, but resources and capacity to gather and analyse accurate and reliable data from many local structures -such as small municipalities- are limited.

2. Methods and Data

Considering the limitations of available data sources, it is necessary to explore and evaluate alternative solutions to obtain the required information. Among these solutions, methods of "small area estimation" (SAE) are popular in providing detailed information on specific areas. Small area estimation methodologies are statistical techniques used to estimate parameters in areas where the sample size of a survey is too small for reliable estimation and/or where data have not been collected (Asian Development Bank, 2020).

Traditionally there are two types of small area estimation – direct and indirect estimation. Direct small area estimation is based on survey design and includes three

estimators called the Horvitz-Thompson estimator, GREG estimator and modified direct estimator (Asian Development Bank, 2020). On the other hand, indirect approaches of small area estimation can be divided into two classes – statistical (Rao, 2003) and geographic approaches (Rahman *et al.*, 2010). Within the geographic approach, spatial microsimulation models (SMMs) have been widely applied on health outcomes and behaviours in populations (Smith *et al.*, 2021).

This paper focuses on the potential of the spatial microsimulation approach, demonstrating its application in estimating the number of physically vulnerable elderly individuals in the municipalities of the province of Sondrio.

There are two types of spatial microsimulation: static and dynamic. In practice, while a static microsimulation model provides an estimated population by synthesizing data at a specific point in time, a dynamic microsimulation model incorporates the ability to model changes and transitions over time, allowing the population to age and evolve within the simulation (Tanton, 2014). In this paper spatial static microsimulation is used.

The spatial microsimulation takes in two types of data:

1. Statistics on small areas, such as aggregated census tables for each municipality in a region, provided by ISTAT . These tables were constructed based on the data from the 2021 Permanent Census of the Population (Istat, 2021). Specifically, two tables were used: the first one containing the distribution of residents in two age groups (65 to 74 years old and 75 years and older) in each municipality of the province of Sondrio, by sex and citizenship (Italian or foreign/stateless), and the second one containing the distribution of residents aged 65 and older in each municipality of the province of Sondrio, by educational attainment in four categories: no education or elementary school, middle school diploma, high school diploma, and university degree or postgraduate degree.
2. Microdata from a sample survey. In this paper, publicly available microdata from the "Multipurpose Survey on Households: Aspects of Daily Life" (2021 edition), specifically related to the geographic distribution of Northwestern Italy, were used. It should be noted that these data do not include the ISTAT code for the municipality or province of residence of the interviewee, only the region code.

Spatial microsimulation utilizes the microdata from Multipurpose Survey on Households as a container of "donor records" for constructing a micro-population in each individual municipality, with gender, age, educational attainment, and citizenship characteristics that allow for the most accurate reproduction of the distribution of residents in the cells of the census tables.

Within the family of static spatial microsimulation techniques, three types of alternative algorithms dominate the literature) and allow for the creation of simulated

spatial microdata: iterative proportional fitting (IPF), combinatorial optimization (CO), and generalized regression reweighting (GREGWT) (O'Donoghue *et al*., 2014).

In this paper, the combinatorial optimization (CO) algorithm called simulated annealing, implemented in the Flexible Modelling Framework (FMF) software developed at the University of Leeds (Harland *et al*., 2012; Harland, 2013), is used for spatial microsimulation.

2.1. The Six Steps of Spatial Microsimulation

The complete process of microsimulation can be divided into six conceptual steps:

1. Definition of small areas.
2. Definition of variables to be estimated.
3. Definition of "constraining" variables.
4. Construction of the microsimulation model.
5. Calculation of the percentage of people within a municipality that fall into the category of the variable of interest.
6. Evaluation of the results.

In the first step (point 1), the municipalities of the province of Sondrio are considered as small areas, and the population aged 65 and older is the target population for the study.

In the second step (point 2), the concept of "physical vulnerability" is defined. Physical vulnerability is a complex and multifactorial concept, which makes its accurate measurement difficult. For example, physical vulnerability can be influenced by factors such as age, chronic conditions, cognitive decline, and social support, which can be challenging to measure comprehensively (Rockwood *et al.*, 2005).

In this study, based on the publicly available sample information from the "Multipurpose Survey on Households: Aspects of Daily Life," an attempt was made to create a measure of "physical vulnerability" that is easy to calculate, easy to understand, and at the same time, meets the needs of potential users (in our case, local administrators) and is comparable at the territorial level. The proposed measure allows for identifying individuals at higher risk of physical vulnerability based on their self-reported health status and limitations in activities due to health problems and was derived by combining the responses of two survey questions formulated as follows: "How is your overall health?" and "To what extent do you have limitations in your usual activities lasting at least 6 months due to health problems?" The responses to these questions are converted into a binary variable, where a value of

"1" indicates that the person is physically vulnerable, and a value of "0" indicates that the person is not physically vulnerable. Specifically, a person is considered physically vulnerable if they report their health as "poor" or "very poor" and, at the same time, report limitations in their usual activities due to health problems. In all other cases, they are considered not vulnerable.

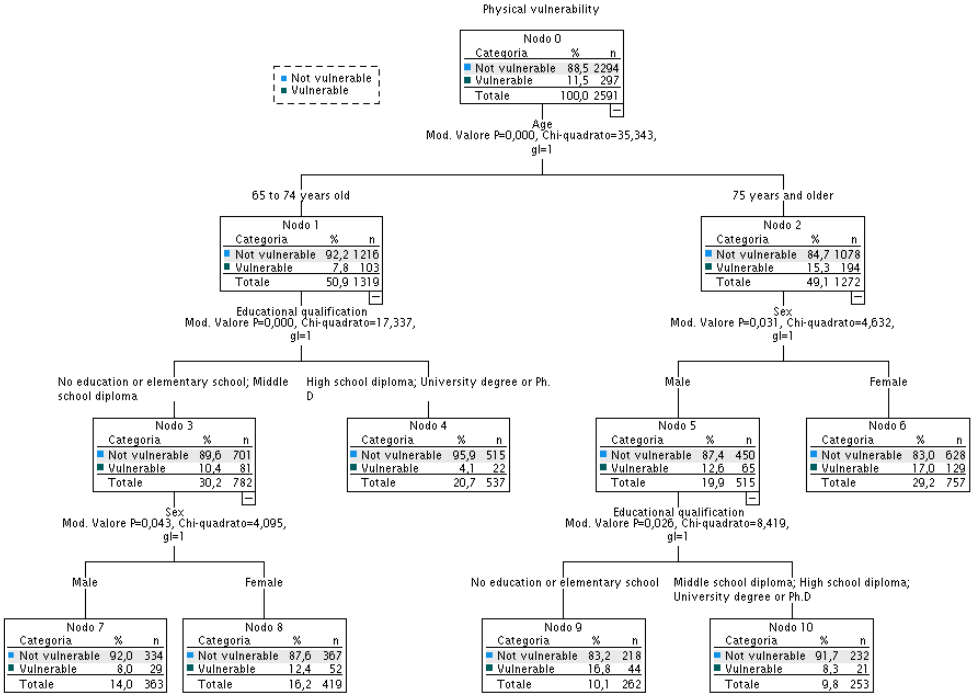
In the third step (point 3), the "constraining" variables considered are age, gender, educational attainment, and citizenship.

In the context of spatial microsimulation, from a technical standpoint, the categories of the "constraining" variables must be the same in both the tables derived from census data and the sample microdata. From a substantive standpoint, the selection of "constraining" variables should be guided by their capacity to accurately explain variability in physical vulnerability. In other words, when choosing these variables, it is important to pick ones that are strongly related with physical vulnerability. This ensures that the spatial microsimulation model will be effective in capturing and representing the real-world variations in vulnerability.

The choice of age, gender, and educational attainment variables fully satisfies this requirement as they provide valuable insights into the physical vulnerability of the elderly and are well supported by the literature (Galluzzo *et al*., 2018; European Institute for Gender Equality, 2021; Petrelli *et al.*, 2019). Age is an established predictive factor, as physical decline and chronic health conditions increase with advancing age. Gender is relevant as women tend to live longer but are more susceptible to chronic health conditions and disabilities. Higher levels of education are associated with better health outcomes and reduced risk of disability and chronic health conditions.

Preliminary analyses, based on classification trees, of the microdata from the Multipurpose Survey reveal a significant predictive capacity of these variables for the physical vulnerability status of individuals (Figure 1) consistent with the literature.

Figure 1 - *Decision tree with the target variable "Physical Vulnerability" and predictor variables "Age," "Educational Attainment," and "Gender".*



Source: Multipurpose Survey on Households: Aspects of Daily Life, ed. 2021 - Northwestern Italy.

Simulated annealing (point 4) is stochastic computational technique for finding near globally-minimum-cost solutions to large optimisation problems: in our case, to select a configuration of microdata in each municipality that closely reproduces the census tables constructed from the Istat census data used in the process. For a mathematical description of the algorithm, please refer to Appendix A and D in Harland et al. (2012).

The functioning of the algorithm can be, however, clearly explained in a more intuitive way. The algorithm starts by randomly selecting a certain number of individuals, which we can refer to as "donors," from the sample survey data and placing them in a specific municipality until reaching the correct number of residents for that municipality (provided by the Census). The procedure is repeated for all municipalities. At this point, the tables obtained for each municipality will not be identical to the Istat source tables, except for the total number of people.

Simultaneously, an error measure is created that compares the values in the cells of the Istat census tables with those obtained from the simulated data. The error measure used in this study is the Total Absolute Error (TAE), which is the sum of the absolute differences between the simulated and actual values in each cell. The TAE[2] calculates the number of people in the population that have been misclassified.

The simulated annealing algorithm works by exchanging individuals between the simulated population and the sample of individuals and checking if this exchange improves the error measure. If so, the algorithm keeps the exchange; otherwise, it cancels it and looks for another person in the sample to perform the exchange. The process continues until successive modifications no longer improve the error measure or until the number of cycles set by the analyst is exhausted.

When comparing a single variable, such as gender, the process is relatively simple. However, when there are multiple variables to compare, replacing one individual may improve the distribution fit of one variable to the census data but worsen it for another. Additionally, there may be no suitable individual in the sample for replacement.

Spatial microsimulation allows estimating variables in the simulated micropopulation that are not present in the Istat census tables but are present in the sample of individuals (point 5). By taking an individual from the sample and including them in the simulated population, the algorithm also "carries over" the associated non-constraining variables, referred to as additive variables. For example, whether the individual is physically vulnerable or not. At the end of the simulation, it is sufficient to count how many records exist in each municipality with the additional variable "physically vulnerable" equal to 1. This operation allows calculating the number of vulnerable people and the rates of physically vulnerable elderly individuals for the two age classes: 65-74 and 75 years and above. The absolute estimates are not definitive since, implicitly using the data of the resident population from the census, we do not consider the elderly individuals living in care facilities such as nursing homes and therefore overestimate the number of physically vulnerable residents. To control for this source of distortion, we use census information on the "number of disabled adults and elderly residents in care facilities - nursing homes" (Istat, 2021), Resident population in cohabitation by type of cohabitation and sex - Lombardy). We subtract the number of disabled adults and elderly residents in care facilities - nursing homes from the total number of elderly individuals and apply the previously calculated rates to obtain the definitive estimates of the absolute number of physically vulnerable elderly individuals living in households. For this purpose, we calculate the vulnerability rates within each municipality.

[2] $TAE = \sum_i \sum_j |O_{ij} - E_{ij}|$. Where O_{ij} and E_{ij} are the observed and expected counts for the i,j-th cell, respectively.

To conclude the section on the method, it is necessary to evaluate the simulation results (point 6).

Firstly, the estimation of the variable under study will be more accurate the stronger the correlations between the constraining variables and the additional variables in the microdata of the Multiscopo sample survey (and if these correlations represent the variations in each simulated area considered). It should be emphasized that this requirement applies not only to spatial microsimulation techniques but also to estimation techniques for small areas based on regression models. To be valid, these techniques require a high predictive capacity of the independent variables in the chosen regression model underlying the simulation. In our study, this requirement is satisfied by the results of the analyses on the sample data, as stated in point three.

Secondly, the spatial microsimulation method ends with point estimates and does not provide uncertainty intervals around the point estimates. Currently, despite some promising attempts to solve this problem (Whitworth et al., 2017; Moretti and Whitworth, 2021), the accuracy of the estimates is generally assessed through a combination of internal validation based on the Total Absolute Error (TAE) and external validation with respect to some related results whose distribution is known (Smith et al., 2011, Edwards et al., 2012).

In this paper, internal validation of the model was performed using TAE, which, as mentioned earlier, is a measure of statistical fit that compares the observed values in the initial tables provided by the Census with the tables calculated from the estimated microdata. In our case, the final TAE is equal to 86: in other words, considering the distribution of over 43,000 units, the produced tables deviate from the census tables by only 86 units, i.e., 0.2%. This result confirms the excellent fit of the model to the initial data.

3. Results and implications for policymakers

The main result of applying spatial microsimulation is to quantify the studied problem, which in this case is the physical vulnerability of the elderly at the municipal level for the two age groups 65-74 and 75 years and above (Table 1).

Table 1 – *Extract from the table of estimates for elderly residents in nursing homes and vulnerable elderly individuals living at home by municipality code in province of Sondrio.*

Municipality Code	Residents aged 65-74, living at home, "physically vulnerable" "	Residents aged 75 and above, living at home, "physically vulnerable""	Residents in nursing homes, care homes for disabled adults and the elderly	Total residents aged 65 and above
14001	6	9	0	99
14002	38	61	0	879
14003	5	13	0	145
…	…	…	…	…
14076	1	3	0	49
14077	7	18	0	205
14078	24	53	52	748
Tot. Prov.	1657	3310	1152	43788

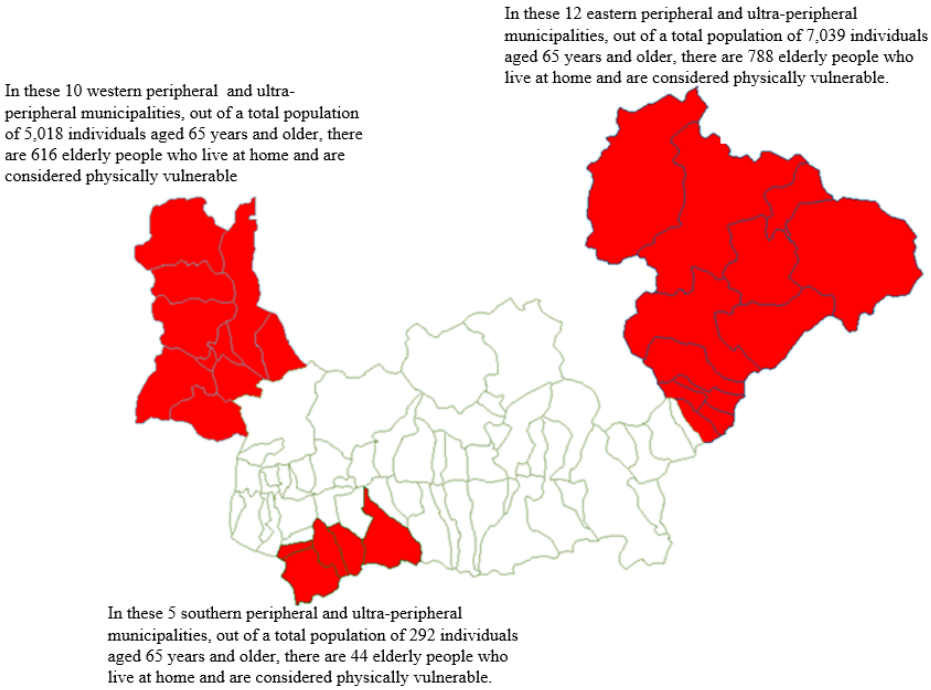
Note: tables is only for illustrative purposes and therefore does not present all values.

Synthetic data from spatial microsimulation is valuable for social policy planning, specifically in estimating the number of physically vulnerable elderly individuals in regions facing challenges related to healthcare, connectivity, transportation, population density, and social services.

Using the typology developed for internal areas (Istat, 2022), peripheral and ultra-peripheral areas in Sondrio were analysed to determine the number of vulnerable elderly people living in these areas with limited amenities. In particular peripheral and ultra-peripheral areas are characterised by their distance of over 40 minutes from a "hub," which refers to municipalities providing secondary education options (comprising at least one high school, be it scientific or classical, as well as at least one technical or vocational institute), a DEA-level hospital, and a railway station meeting at least the silver standard.

The choropleth map (Figure 2) visually illustrates the peripheral and ultra-peripheral areas in red, along with the estimated count of physically vulnerable elderly individuals residing within them.

Figure 2 – *Choropleth map with peripheral and ultra-peripheral municipalities highlighted in red and the estimated total number of vulnerable elderly individuals, living at home.*



4. Conclusion and limitations of the study

This study has demonstrated that spatial microsimulation can be used to estimate the number of vulnerable elderly individuals in the province of Sondrio. Using simulated data generated through this technique, local administrators can identify areas with the highest number of vulnerable elderly individuals and concentrate resources and services to support these individuals, such as planning the construction of elderly care facilities, enhancing transportation services for people with physical limitations, or providing home support for those who have difficulty moving.

However, the technique has some limitations, particularly in its ability to provide uncertainty intervals around the central estimates. Nevertheless, this does not mean that spatial microsimulation approaches cannot have potential advantages once the issue of estimation accuracy is resolved. Therefore, in our opinion, the potential

opportunities offered by spatial microsimulation approaches should not be ignored in further development.

References

ASIAN DEVELOPMENT BANK. 2020. *Introduction to small area estimation techniques. A Practical Guide for National Statistics Offices*. Manila: ADB.

CLEGG, A., YOUNG, J., ILIFFE, S., RIKKERT, M. O., ROCKWOOD, K. 2013. Frailty in elderly people*, Lancet* (London, England), Vol. 381, No. 9868, pp. 752–762.

EDWARDS, K.L., TANTON, R. 2012. Validation of Spatial Microsimulation Models. In TANTON, R., EDWARDS, K. (Eds.) *Spatial Microsimulation: A Reference Guide for Users. Understanding Population Trends and Processes*, Vol. 6, Dordrecht, Springer.

EUROPEAN INSTITUTE FOR GENDER EQUALITY. 2021. *Gender Equality Index 2021: Health.* Luxembourg: Office of the European Union.

GALLUZZO L., O'CAOIMH R., RODRÍGUEZ-LASO Á., BELTZER N., RANHOFF AH., VAN DER HEYDEN J. 2018. Incidence of frailty: a systematic review of scientific literature from a public health perspective, *Annali Istituto Superiore Sanità*, Vol. 54, No. 3, pp. 239-245.

HARLAND, K., HEPPENSTALL, A.J., SMITH, D., BIRKIN, M. 2012. Creating Realistic Synthetic Populations at Varying Spatial Scales: A Comparative Critique of Population Synthesis Techniques, *Journal of Artificial Societies and Social Simulation,* Vol. 15, No, 1.

HARLAND, K. 2013. Microsimulation model user guide ver. 1.0 (Flexible Modelling Framework). *Woking paper 6/13*, School of Geography, University of Leeds, United Kingdom.

ISTAT. 2021. Censimenti permanenti. Roma: Istituto Nazionale di Statistica.

ISTAT. 2022. La geografia delle aree interne nel 2020: vasti territori tra potenzialità e debolezze, *Statistiche Focus*. Roma: Istituto Nazionale di Statistica.

ISTAT. 2023. Previsioni della popolazione - anni 2021-2070. Roma: Istituto Nazionale di Statistica.

MORETTI, A., WHITWORTH, A. 2021. Estimating the Uncertainty of a Small Area Estimator Based on a Microsimulation Approach, *Sociological Methods & Research*, pp. 1-31.

O'DONOGHUE, C., MORRISSEY, K., LENNON, J. 2014. Spatial Microsimulation Modelling: A Review of Applications and Methodological Choices, *International Journal of Microsimulation*, Vol. 7, No. 1, pp. 26–75.

PETRELLI, A., DI NAPOLI, A., SEBASTIANI, G., ROSSI, A., GIORGI ROSSI, P., DEMURU, E., COSTA, G., ZENGARINI, N., ALICANDRO, G., MARCHETTI, S., MARMOT, M., FROVA, L. 2019. Atlante Italiano delle Disuguaglianze di Mortalità per Livello di Istruzione, *Epidemiologia e prevenzione*, Vol. 43, No. 1S1, pp. 1-120.

RAHMAN, A., HARDING, R., TANTON, S. L. 2010. Methodological Issues in Spatial Microsimulation Modelling for Small Area Estimation, *International Journal of Microsimulation,* Vol. 3, No. 2, pp. 3-22.

RAO J.N.K. 2003 *Small area estimation.* Hoboken, New Jersey: John Wiley & Sons.

ROCKWOOD, K., SONG, X., MACKNIGHT, C., BERGMAN, H., HOGAN, D. B., MCDOWELL, I., MITNITSKI, A. 2005. A global clinical measure of fitness and frailty in elderly people, *CMAJ: Canadian Medical Association journal*, Vol. 173, No. 5, pp. 489-495.

SMITH, D. M., HEPPENSTALL A., CAMPBELL, M. 2021. Estimating Health over Space and Time: A Review of Spatial Microsimulation Applied to Public Health, *J-Multidisciplinary Scientific Journal*, Vol. 4, No. 2, pp. 182-192.

SMITH, D.M., PEARCE, J.R., HARLAND, K. 2011. Can a Deterministic Spatial Microsimulation Model Provide Reliable Small-Area Estimates of Health Behaviours? An Example of Smoking Prevalence in New Zealand, *Health Place*, Vol. 17, No. 2, pp. 618–624.

TANTON, R. 2014. A Review of Spatial Microsimulation Methods, *International Journal of Microsimulation*, Vol. 7, No. 1, pp. 4-25.

WHITWORTH, A., CARTER, E., BALLAS, D., MOON, G. 2017. Estimating Uncertainty in Spatial Microsimulation Approaches to Small Area Estimation: A New Approach to Solving an Old Problem, *Computers, Environment and Urban Systems*, Vol. 63, pp. 50-57.

Alberto VITALINI, Istat, Sede della Lombardia, vitalini@istat.it
Simona BALLABIO, Istat, Sede della Lombardia, ballabio@istat.it
Flavio VERRECCHIA, Istat, Sede della Lombardia, verrecchia@istat.it

INTEGRATION BETWEEN DATA FROM REGISTER AND SAMPLE SURVEYS: ENTERPRISES CLASSIFIED BY USE OF ICT AND ECONOMIC INDICATORS

Alessandra Nurra, Giovanni Seri, Valeria Tomeo

Abstract. Business investments in Information and Communication Technologies (ICT) that impact production processes represent an important lever for enhancing business productivity. The data presented in this work offer new indicators and classifications by integrating the phenomenon of digitization with elements of economic performance. The Italian National Institute of Statistics (Istat) annually measures the digital transformation of enterprises with at least 10 persons employed through a sample survey that is harmonized at the European level and focuses on the use of ICT and e-commerce. The survey primarily adopts a qualitative approach, and to explore the economic characteristics of companies on the basis of their level of ICT adoption it is essential to integrate productivity and profitability indicators. Particular attention was given to the possibility of integrating the ICT data with the statistical information system, known as FRAME SBS (Structural Business Statistics), which is used for estimating structural economic variables related to business accounts on the entire population of enterprises, as defined by the SBS Regulation. The calibration estimator methodology has been chosen as the most suitable approach for integrating the data from the FRAME SBS with the ICT sample survey, aiming at analysing indicators based on data stemming from both sources. Specifically, the analysis includes the presentation of structural, productivity and competitiveness measures examined in relation to the indicators of the digital intensity of enterprises with at least 10 persons employed, such as the use of personal computers and the internet, web presence, and online sales.

1. ICT and competitiveness

Investments in Information and Communication Technologies (ICT) that impact production processes play a crucial role in driving business productivity growth. The new Regulation on European business statistics emphasizes the significance of measuring the digital economy and the use of ICT, recognizing their influence on competitiveness, growth, and the need to promote European strategies and policies related to the completion of the digital single market. Both the European

Commission and Parliament highlight the importance of monitoring the digital progress of Member States through the Digital Economy and Society Index (DESI). Additionally, there is a focus on measuring EU trajectories against targets defined by the Digital Decade Policy Programme 2030.

The Italian National Institute of Statistics (Istat) annually measures the digital transformation of enterprises with at least 10 persons employed through a sample survey that is harmonized at the European level. This survey focuses on the utilization of ICT and aims to investigate the extent of adoption of various emerging technologies, considered by policymakers as enabling behaviours and virtuous processes that can enhance the competitiveness of businesses. The survey primarily adopts a qualitative approach, and it is essential to integrate indicators of productivity and profitability. This integration allows for an examination of the economic characteristics of companies based on their level of ICT adoption. Furthermore, it enables the creation of an integrated database suitable for cross-sectional and panel analyses, particularly for larger enterprises included in the survey (all enterprises with at least 250 persons employed are included in the ICT survey).

In literature, there are numerous studies related to the impact that the adoption of information and communication technologies (ICT) can have on innovation processes, production mechanisms (such as robotics or additive manufacturing), organizational functions, as well as firm and overall system performance. However, this positive relationship between ICT usage and firm productivity has not always been consistently demonstrated, particularly in advanced countries. This has led to the development of two distinguishable schools of thought known as "techno-optimists" and "techno-pessimists" (Andrews *et al.*, 2016; Cette *et al.*, 2016). As a result, some studies argue, for various reasons, that Italy belongs to the countries where the use of ICT struggles to become a lever for improving the entire economic system. This could be attributed to factors such as the prevalence of small enterprises that hinder the diffusion of new technologies (Accetturo *et al.*, 2013), inefficiencies in management selection (Pellegrino and Zingales, 2017), limited investment in human capital (Bugamelli and Pagano, 2004), or the low effectiveness of innovation support policies (Bronzini and Piselli, 2016).

In recent years, Istat has also conducted several analyses on the competitiveness of Italian companies in relation to their levels of digitization and capital endowments (both human and physical) at both national and regional levels. These analyses have identified, on one hand, the presence of virtuous digitization behaviors, and on the other hand, the need for adequate physical and, more importantly, human capital inputs to effectively transform ICT into growth opportunities. Special attention was dedicated to exploring the potential integration of data between the statistical information system, known as FRAME SBS (Structural Business Statistics), which estimates structural economic variables based on business accounts, and sample

surveys such as the one focused on the use of ICT (Istat, 2020). The FRAME SBS variables are derived from administrative data or statistical estimation methods. They constitute a comprehensive archive that encompasses the entire population of enterprises as defined by the SBS Regulation.

2. Methodology

The aim of this work is to integrate the information gathered by a sampling (ICT) and exhaustive (FRAME SBS) sources by combining qualitative indicators and economic variables through a method qualifying the results in terms of comparability and consistency according to Istat statistical standards.

Different approaches have been considered: macro methods such as balancing and iterative proportional fitting, and micro methods including statistical matching with or without weights, consistent repeated weighting, and calibration estimators (Seri *et al.*, 2016). The calibration estimates methodology was applied as the most suitable approach for integrating data from the FRAME SBS with the ICT survey. It utilizes the interaction between an exhaustive register and sample data to produce economic indicators. This is the same methodology used in the production process of the ICT sample survey. However, the initially proposed set of indicators does not directly or indirectly replicate published estimates. Instead, it utilizes the information derived from the combination of the two data sources in a manner that ensures substantial or complete consistency with both sources. Considering the significance of the Digital Intensity Index (EDII) as an indicator used by Eurostat, based on the use of 12 digital activities and now being employed for the Digital Decade, we have also incorporated this indicator into our analysis.

Several differences in applying method are listed as follows:

- considering t as reference year of the ICT survey, the statistical archives used for estimations consist of FRAME SBS refers to the previous year (t-1), while the ICT sample survey of year t utilizes ASIA referred year (t-2). The given population is identified based on updated information from the reference year (t-1);
- a small percentage of enterprises interviewed by the survey are no longer part of the considered population, mostly because they have fewer than 10 persons employed. These units have been excluded from the sample, and the data set used for indicator estimations is based on matching observations between the ICT survey units and the units in the FRAME SBS. Anyway, more than 90% of the ICT respondents is eligible for the analysis;
- the estimation domains are redefined by matching the survey's required domains (with small details) to align with the study's objectives (tables of

indicators with specific characteristics) and the quality of the results obtained. Specifically, the territorial level has been reduced, excluding regional information;

- in terms of the calibration model used for weights in the ICT survey (totals for variables such as the number of enterprises and number of employees by NACE and geographical level), the utilization of FRAME SBS was crucial to consider known totals related to Value Added, Turnover, and Gross Operating Margin.

The methodological framework utilized is structurally the same as that of the ICT survey. Therefore, to assess the accuracy and precision of the produced estimates, the same criteria accompanying the currently published estimates can be adopted. The ReGenesees software (Zardetto, 2015), which implements the methods commonly used in Istat for economic surveys, was employed for the analysis. Furthermore, since the adopted strategy generates microdata files with a weighting system that represents the entire population (similar to the survey) it was possible to reproduce the estimates of the ICT indicators (which is not the objective of this work to replicate). These estimates are entirely consistent with the published ones, providing reassurance about the consistency of the results (the consistency with FRAME SBS is guaranteed by design in the new estimation domains). However, it is important to reiterate that the objective of this work was to define a series of indicator tables that combine information from FRAME SBS and the ICT survey. Therefore, it is advisable to exclusively use the dataset of elementary units (and their corresponding weights representing the entire population) that enabled the achievement of this objective.

2.1. ICT indicators

The ICT indicators have been constructed in such a way as to divide the population of enterprises into exhaustive classes according to different types of ICT adoption. As previously mentioned, the proposed set of indicators does not replicate, either directly or indirectly, published estimates, but it leverages the informative interaction between the two sources in substantial or complete coherence with both. Additionally, for enterprises with at least 10 persons employed, the structural characteristics are presented (export propensity, group membership), as well as productivity and efficiency (labor productivity) and competitiveness (labor cost per employee) indicators in relation to indicators measuring the extent of ICT utilization.

The ICT indicators are as follows:

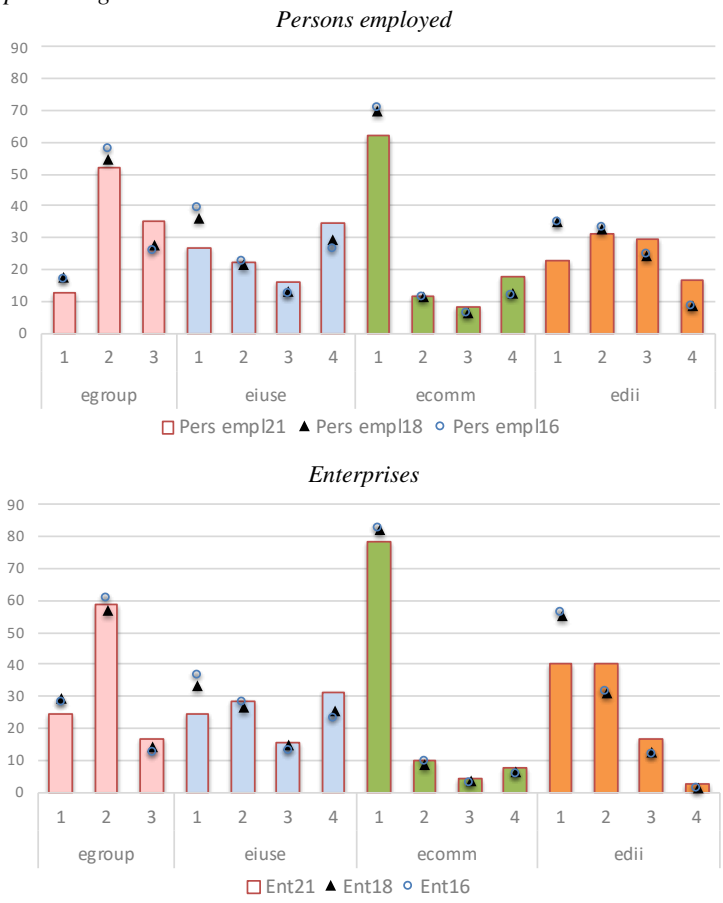
Table 1 – *Indicators that measure the extent of ICT adoption.*

ICT Indicators	Values	Description
EGROUP		Enterprises classified by level of Internet usage (PC online, web, e-sales)
	1	Enterprises without/with PC and Internet but not having a website and online sales
	2	Enterprises connected to the Internet and having a website but not making online sales
	3	Enterprises connected to the Internet and making online sales
EIUSE		Enterprises classified by penetration rate of Internet usage among persons employed
	1	Enterprises without PC and Internet + those using PCs and less than 25% persons employed are connected to the Internet
	2	Enterprises with % of persons employed using Internet between 25% and less than 50% of the total
	3	Enterprises with % of persons employed using Internet between 50% and less than 75% of the total
	4	Enterprises with at least 75% of persons employed using the Internet
ECOMM		Enterprises classified by selling online or not
	1	Enterprises not having website or having it but without shopping cart functionality and not selling online
	2	Enterprises having website with shopping cart functionality but not selling online in previous year
	3	Enterprises selling online in previous year without shopping cart functionality (indirect e-commerce proxy)
	4	Enterprises selling online in previous year with shopping cart functionality
EDII		Digital intensity index used by Eurostat based on use of 12 digital activities
	1	Very low intensity – if enterprises use 0-3 digital activities
	2	Low intensity – if enterprises use 4-6 digital activities
	3	High intensity – if enterprises use 7-9 digital activities
	4	Very high intensity – if enterprises use 10-12 digital activities

3. Key findings

Data reveals the increasing evolution of digital usage (edii2 and edii3, eiuse4) and the decreasing of enterprise that have low use of Internet (egroup1, eiuse1, ecomm1, edii1). Confirming ICT annual results, data reveal that only few enterprises are engaged in e-commerce (ecomm2, ecomm3, ecomm4) and in very high level of ICT activities adoption (edii4). (Figure 1)

Figure 1 – *Persons employed and enterprises by ICT indicators. Year 2016, 2018 and 2021, percentage values.*



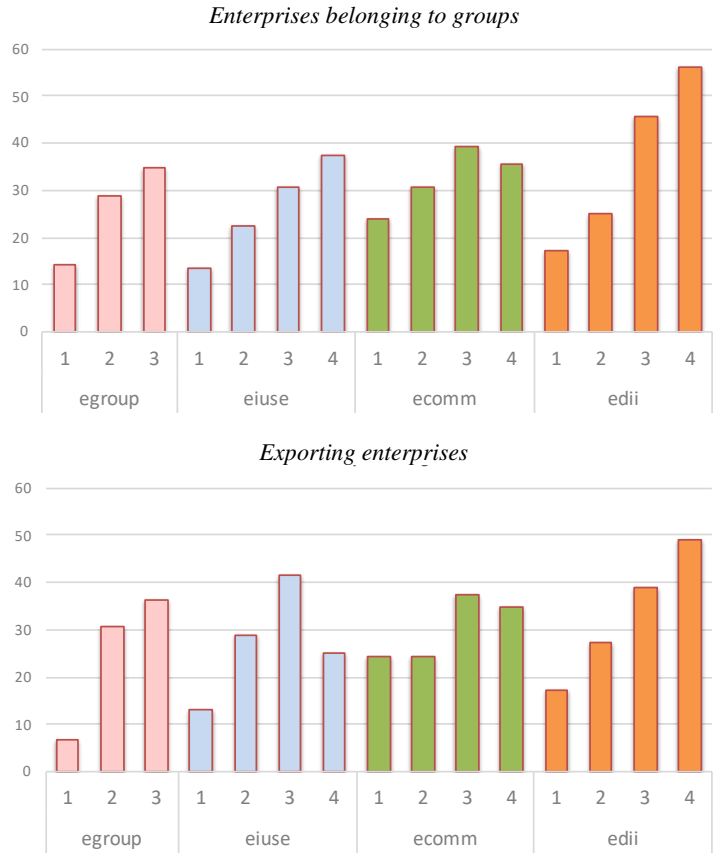
Authors' elaborations on Istat data

The share of companies with at least 10 persons employed belonging to groups increases as it grows digitization indicators or if they use third-party online sales channels beyond their own (ecomm3).

The same occurs for exporters, except for the last level of the eiuse indicator, where there is a prevalence of companies with high technological content and low propensity to export, such as software production and IT consulting.

A similar situation is observed for the level 4 of ecomm indicator, in which companies are concentrated in the accommodation services sector. (Figure 2)

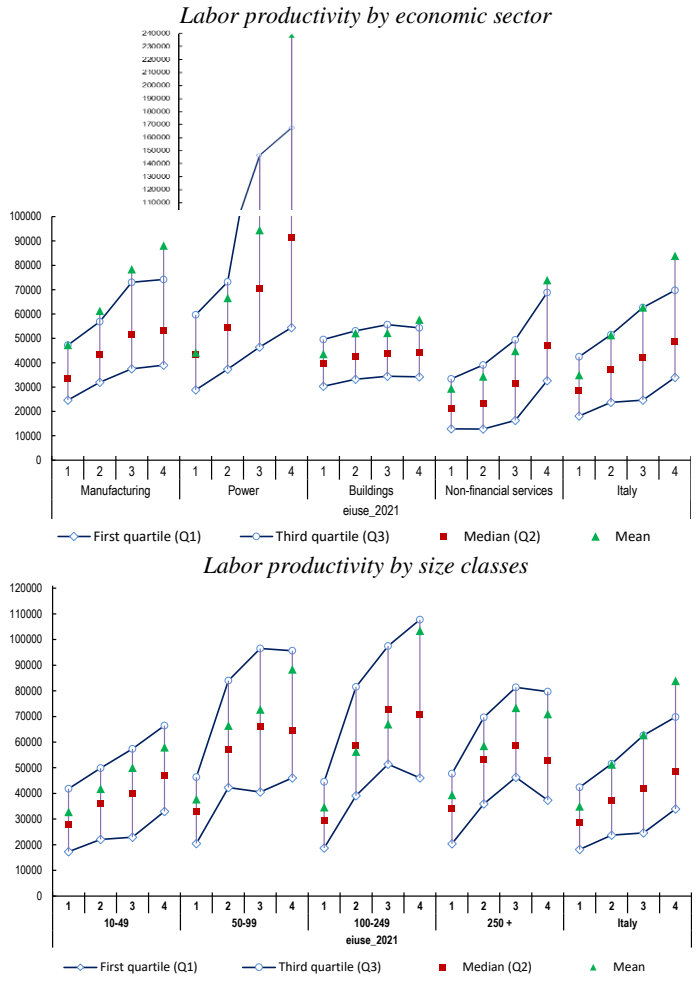
Figure 2 – *ICT indicators by structural characteristics (belonging to groups and exporting enterprises). Year 2021.*



Authors' elaborations on Istat data

There is a positive correlation between the adoption of technologies (eiuse) and the labor productivity measured in terms of value added per persons employed, especially in manufacturing and energy companies and, in particular, for those already in a position of high productivity compared to the others (third quartile). This positive correlation is evident also in terms of classes of persons employed and is very similar in all the years analysed. (Figure 3)

Figure 3 – *Labor productivity and levels of Internet usage (eiuse) by economic sector and by size classes of persons employed. Year 2021.*



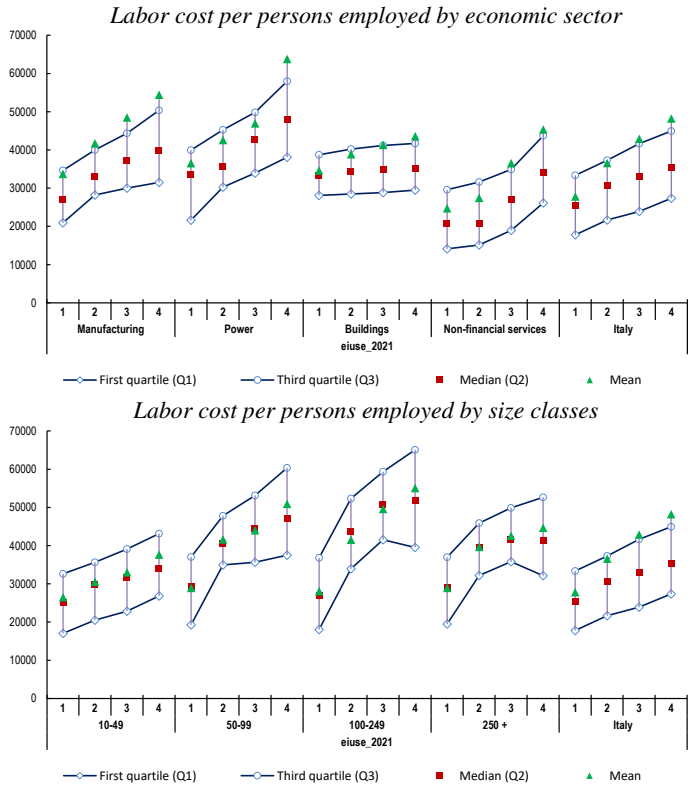
Authors' elaborations on Istat data

The interquartile gap seems to increase mostly as the technology indicator rises, highlighting a greater dispersion around the average of the indicator and therefore a greater differentiation between those who get the most benefits and those who fall behind.

In general, the percentage growth of the economic indicators considered at the level of persons employed class or macro-sector shows larger increases, especially during the initial transition from low or no ICT usage to the subsequent level. However, the relationship between ICT and economic indicators becomes less evident in subsequent jumps and, at times, even negative.

The positive correlation is evident also between the adoption of technologies (eiuse) and the labor cost per persons employed in terms of economic sector and size classes of persons employed. (Figure 4)

Figure 4 – *Labor cost per persons employed and level of Internet usage (eiuse) by economic sector and by size classes of persons employed. Year 2021.*



Authors' elaborations on Istat data

When analysing the data regarding the distribution of companies across quartiles of labor productivity, a general trend emerges: there is an increase in the proportion of companies belonging to higher levels of the chosen ICT indicator as the quartile of economic performance grows (from the first to the second and third quartile). The share of larger-sized companies is greater than that of smaller-sized companies in the higher quartiles of productivity, particularly at higher levels of ICT adoption (Figure 5). Smaller-sized companies show productivity improvements at lower levels of ICT adoption compared to what is required for larger-sized companies.

Figure 5 – *EDII indicator levels by quartiles of labor productivity and size classes of persons employed. Years 2018 and 2021.*

p.e. classes	quartile classes	edii_2018 1	2	3	4	edii_2021 1	2	3	4
10-49	<=Q1	74,1	19,5	5,8	0,6	46,6	40,4	11,0	2,1
10-49	Q1 - Q2	55,6	31,3	12,3	0,9	48,4	40,1	10,8	0,7
10-49	Q2 - Q3	54,9	31,8	12,4	1,0	42,4	40,0	15,3	2,3
10-49	>Q3	45,4	39,2	14,0	1,5	32,7	42,6	22,6	2,2
50-99	<=Q1	64,0	27,6	6,6	1,8	46,0	33,4	16,6	4,0
50-99	Q1 - Q2	43,0	37,3	17,5	2,2	26,3	40,1	29,6	4,1
50-99	Q2 - Q3	37,4	39,7	19,3	3,6	11,9	37,7	39,9	10,4
50-99	>Q3	27,2	41,3	26,7	4,8	7,0	35,8	49,7	7,5
100-249	<=Q1	61,7	26,4	11,0	0,8	25,5	36,5	26,9	11,0
100-249	Q1 - Q2	30,9	43,4	23,6	2,0	15,1	30,8	40,0	14,1
100-249	Q2 - Q3	18,6	48,0	29,3	4,1	5,6	23,3	48,6	22,5
100-249	>Q3	16,8	39,6	34,9	8,7	2,2	16,2	49,6	32,0
250 +	<=Q1	43,9	38,6	15,5	2,1	48,1	35,1	13,2	3,6
250 +	Q1 - Q2	18,1	37,3	36,5	8,1	28,4	46,9	20,3	4,5
250 +	Q2 - Q3	12,1	33,7	43,7	10,5	19,3	42,2	36,1	2,4
250 +	>Q3	4,9	30,2	47,8	17,1	12,8	42,8	37,0	7,3
Total	**<=Q1**	**73,0**	**20,4**	**5,9**	**0,7**	**46,7**	**39,7**	**11,2**	**2,3**
Total	**Q1 - Q2**	**55,0**	**31,6**	**12,5**	**0,9**	**47,3**	**39,7**	**12,0**	**1,0**
Total	**Q2 - Q3**	**52,0**	**33,1**	**13,5**	**1,4**	**38,9**	**40,7**	**17,7**	**2,7**
Total	**>Q3**	**40,3**	**39,6**	**17,6**	**2,5**	**27,5**	**41,4**	**26,9**	**4,1**

Authors' elaborations on Istat data

4. Results and future work

The integration of microdata between the ICT survey and the SBS Frame has allowed for expanding the range of available indicators by leveraging the combined information from the two sources.

During the calibration process, the regional level of estimation detail was lost, but it was possible to maintain the other estimation domains by fixing the known totals to other economic variables.

The methodology presented here for the ICT survey has also been applied to other surveys, such as the Community Innovation Survey (CIS), and has provided equally interesting results.

References

ACCETTURO A., BASSANETTI A., BUGAMELLI M., FAIELLA I., FINALDI RUSSO P., FRANCO D., GIACOMELLI S., OMICCIOLI M. 2013. Il sistema industriale italiano tra globalizzazione e crisi. *Questioni di Economia e Finanza, Occasional Papers 193*, Banca d'Italia.

ANDREWS D., CRISCUOLO C., GAL P. 2016. The Best versus the Rest: The Global Productivity Slowdown, Divergence across Firms and the Role of Public Policy. *OECD Productivity Working Papers 2016/05*, OECD Publishing, Paris.

BRONZINI R., PISELLI P. 2016. The Impact of R&D Subsidies on Firm Innovation, *Research policy*, Vol. 45, No 2, pp. 442-457.

BUGAMELLI M., PAGANO P. 2004. Barriers to Investment in ICT, *Applied Economics*, Vol. 36, No 20, pp. 2275-2286.

CETTE G., FERNALD J.G., MOJON B. 2016. The Pre-Great Recession Slowdown in Productivity, *Working Paper 2016/08*, Federal Reserve Bank of San Francisco.

ISTAT. 2020. Integration between data from register and sample surveys: enterprises classified by use of ICT and economic indicators, Experimental Statistics. Roma: Istituto Nazionale di Statistica.

PELLEGRINO B., ZINGALES L. 2017. Diagnosing the Italian Disease. *Working Paper 23964*, National Bureau of Economic Reasearch.

SERI G., ICHIM D., LUCHETTI F., COSTA S., NURRA A., MASTROSTEFANO V., SALAMONE S., PASCUCCI C., ORSINI D. 2016. Integrazione del Frame con altre Indagini e Fonti Amministrative ai fini della Produzione di Indicatori Complessi, *Istat Working Papers 17*, Istat.

ZARDETTO D. 2015. ReGenesees: An Advanced R System for Calibration, Estimation and Sampling Error Assessment in Complex Sample Surveys, *Journal of Official Statistics*, Vol. 31, No 2, pp.177-203.

Alessandra NURRA, Istat, nurra@istat.it
Giovanni SERI, Istat, seri@istat.it
Valeria TOMEO, Istat, tomeo@istat.it

BUILDING AN INTEGRATED DATABASE FOR THE TRADE SECTOR FOR THE PERIOD 2010- 2022

Maria Rita Ippoliti, Luigi Martone, Fabiana Sartor[1]

Abstract. This paper aims at describing the process of combining and merging data from the retail trade survey with data stemming from the survey on business confidence in retail trade, for the period from 2010 to 2022. While our previous paper focused on macro data, this paper explores the comparability of micro data. Although the samples of the two surveys involved in the study have a different sectoral coverage, combining them interestingly allows checking the consistency between the evaluations expressed by the economic operators in the qualitative survey and the indicators of the quantitative survey. The merged dataset is a valuable resource to study the enterprises' expectations in the retail trade sector as it includes precious information such as retail trade index, economic activity, channel of distribution and size of the enterprises and enterprises' judgements and expectations on sales for the main economic and business variables.
The paper is organised as follows: a general overview of data sources and a description of the key steps for combining the databases of the two mentioned surveys. The following part describes more in detail all actions implemented to build the integrated dataset and analyses the units involved in both samples through descriptive statistics, graphs and statistical tests. Comparing the two samples highlighted that the number of units involved in both surveys has increased over the years and that there is a good correspondence between qualitative data and quantitative data.

[1] Though the article is the result of a joint work, the single paragraphs are attributed as follows: paragraph 1 and 2 to Fabiana Sartor; paragraph 3 to Maria Rita Ippoliti; paragraph 5 and 6 to Luigi Martone. The published articles are exclusively expressing the authors' opinions; Istat shares no responsibility for the published contents.

1. Introduction

In recent years integration of data from different sources in economic analysis has gained a key role as it allows the study of the trend of the Italian production, exploiting the interactions of available sources.

Qualitative data generally provide information on macroeconomics, catching the overall picture. When compared to traditional quantitative data, qualitative variables have the advantage of a better timeliness (Lui et al., 2011a e 2011b).

This study aims at evaluating the consistency among judgments and expectations of the economic agents of trade sector and data on turnover returned by the same agents in quantitative surveys. That allows exploring more in detail the behavior of the operators in trade sector, evaluating any distortions in the answers provided to the qualitative survey.

Although efforts to integrate data from different sources have already been carried out (see Margani and Orsini, 2020), the present study represents the first work concerning the trade sector, integrating micro data from the Retail Trade Survey and the Business Confidence in Retail Trade concerning years from 2010 to 2022 (see Ippoliti M., Martone L., Sartor F. 2021 for macro data integration).

The integrated database, which incorporates information on national sales turnover, economic activity classification, employment size of enterprises, channel of distribution, enterprises' judgments and expectations on main economic and business variables, appears remarkably useful to study overall enterprises' expectations in trade sector.

2. Trade surveys

2.1. Business Confidence Survey in Retail Trade (FIDCOM)

The European Commission coordinates a harmonised project for Member States to provide data on Business Confidence Survey in Retail Trade (FIDCOM). These surveys allow to compare information on the economic evolution of retail trade at European level (NACE Division G, except for Division 46 - Wholesale trade, except of motor vehicles and motorcycles and for Group 47.9 - Retail trade not in stores, stalls or markets including retail sales via mail order or via Internet).[2] The survey asks enterprises to express their opinions (judgements and expectations over the following 3 months) about the main economic variables (orders placed with

2 Divisions of NACE Section G involved in the survey are Division 45 (Wholesale and retail trade and repair of motor vehicles and motorcycles) and Division 47 (Retail trade, except of motor vehicles and motorcycles).

suppliers, employment, selling prices), giving therefore an updated overview on the evolution of the sector and of the perceived economic uncertainty. Respondents are requested to state their consideration on their total sales in the last three months, on their current volume of stock and on prices charged by their suppliers. Additional questions are asked to know their expectations for the following three months on volume of orders, employment, prices they charge and on total sales. The nominal sample of the Business Confidence Survey in Retail Trade comprises about 1.000 commercial enterprises. Three stratification criteria are used: enterprise employment size class (1-2 employees, 3-5; 6-999; at least 1.000 employees), geographical area (North-West, North-East, Centre, South and the Islands) and main activity (45.1 sales of motor vehicles; 45.2-45.4 maintenance of motor vehicles and sales of accessories; 47.1, 47.2 retail sales of food, drinks and tobacco; 47.3 retail sales of automotive fuel; 47.4-47.7 retail sales of other goods). Enterprises with less than 1.000 employees are randomly sampled, while all units with more than 1.000 employees are included in the sample. The data processing method sets out the estimate of the frequency percentages of each reply option relating to each item of the questionnaire. For this purpose, the processing of the micro data is based on a double weighting system: a) the frequencies of each reply option are firstly weighted using the number of employees declared by the enterprise at the time of the interview (internal weight); b) subsequently fixed weights reflecting the distribution of the added value of the reference sector (external weight) are used. Since March 2015, the aggregation procedure uses an external weighting structure derived from the added value at factor cost referred to 2012. Each variable is measured calculating balances as percentage differences between favourable and unfavourable responses. Weighted balances are seasonally adjusted if needed. The Index of Business Confidence in Retail Trade is calculated as the arithmetic mean of seasonally adjusted balances based upon opinions and expectations on sales and upon judgments on volume of stocks (the above-mentioned values have inverse signs).

2.2. Retail Trade Survey (VEN)

The Retail Trade Index (VEN) produces a short-term indicator measuring the changes in the value and volume of sales. The reference population of the survey is all enterprises having retailing as their main economic activity, except retail trade of motor vehicles and motorcycles and automotive fuel. Therefore, the survey covers the retail trade sector only partially (NACE Rev. 2, G 47 - Retail trade, except of motor vehicles and motorcycles not including automotive fuel)[3]. Estimates of Retail

[3] According to NACE Rev. 2, Retail trade (Division G47) is first classified by type of sale outlet (retail trade in stores: groups 47.1 to 47.7; retail trade not in stores: groups 47.8 and 47.9). For retail sale in

Trade Survey provide a timely indicator of economic performance and strength of consumer spending. Monthly indices on retail trade are released at national level, consistently with the European Union Regulations concerning short-term statistics (see European Regulations n. 1165/98 and n. 1158/2005)[4]. The sample of the survey involves over 8.000 enterprises, which are resident in Italy.

The sample is stratified considering the following variables: main activity according to NACE Rev. 2 and employment size class (1-5, 6-49 and at least 50 employees). The sample includes all large retailers (at least 50 employees) and a representative sample of smaller enterprises. Every year, enterprises belonging to the sample strata of employment size class 1-5 and 6-49 are partially replaced either if they stayed in the sample for more than 3 years, or if they changed their main economic activity or if they closed.

The sampling design of the survey rotates some units out and rotates new units in each year (belonging to employment size classes 1-5 and 6-49 only) to share burden and refresh the sample. Typically, every year 2.500 to 3.000 enterprises are replaced, therefore 60% of the sample does not rotate. This aspect gains relevance when comparing retail trade indicators with qualitative data, as the Retail Trade Survey struggles to keep track of quick evolution of stores' closures and openings. According to their distribution channel, enterprises in the retail trade sample can be classified into large-scale distribution, small-scale distribution, internet sales and non-store sales. Within the weighting structure of Monthly Retail Trade Index (base=2015), large scale-distribution accounts for 46.4% of total turnover, while small-scale distribution reaches 48.0% of total turnover. Retail trade indices are calculated as weighted means of the sub-indices of each stratum. Concerning the calculation method of the indicator, the synthetic index numbers are constructed as weighted averages of indices related to the domains identified by the intersection of the stratification variables (main activity and employment size). The Laspeyres index is used to calculate aggregate indices up to the retail trade total. The weights are based on turnover data from SBS of the year 2015. Value of sales indices measures the retail trade turnover over time at current prices and, therefore, incorporates the effects in changes of quantity sold and prices. In order to determine estimates on the volume of sales, value of sales indices are processed to allow removing price effects on turnover, using the Harmonised index of consumer prices (HICP). Monthly data are first revised in the following month after publication (which occurs 38 days past the reference period). On annual basis, provisional indices are subject to a second revision to calculate the finale estimates.

stores, there exists a further distinction between specialised retail sale (groups 47.2 to 47.7) and non-specialised retail sale (group 47.1).

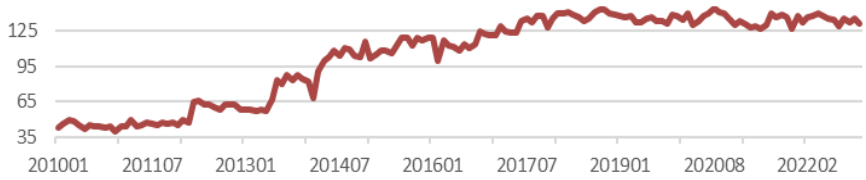
[4] See https://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CONSLEG:1998R1165:20120621:EN:PDF

3. The integrated dataset: building and analyses

As a first step, the dataset of Business Confidence in Retail Trade (FID.COM) was extracted, than it was merged with the microdata from Retail Trade Survey (VEN). ASIA code (an identifying code of each enterprises according to the statistical archive of active businesses) and reference period (YEARMONTH) were both used as key variables for merging the two datasets.

Figure 1 shows the movement in the total amount of businesses involved in both surveys (retail trade and confidence in retail trade) by year. From the graph, it is clear that the number of these enterprises grew during the considered period, going from 43 at the beginning of 2010 to around 130 at the end of 2022, with a significant increase starting from the first months of 2014, when the Statistical Portal of Enterprises was launched by Istat for data collection purposes. The integrated VEN.FIDCOM dataset contains all micro data from the Retail Trade Survey and the Business Confidence in Retail Trade for each enterprise and for each month of the year, even if the enterprise is involved in only one of the two surveys. Furthermore, it is worthwhile noting that within the integrated database, in addition to the above—mentioned variables, for each enterprise we calculated the year-on-year growth in turnover. This latter variable is equal to null when the value of turnover is missing in one the considered periods.

Figure 1 – *Enterprises responding both to the Retail Trade Survey and to the Business Confidence in Retail Trade Survey (years 2010-2022).*

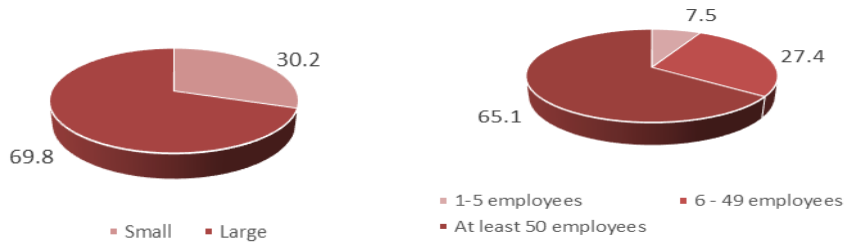


Source: Elaboration on ISTAT data

Figure 2 shows that the number of enterprises responding to both the Retail Trade Survey and the Business Confidence in Retail Trade Survey mostly belong to large-scale distribution (69.8% against 30.2% of enterprises belonging to small distribution).

In order to have more reliable information on the size of the enterprises involved in the study, we chose to consider the number of employees. As a result, the graph on the right of the Figure 2 shows that the amount of large enterprises in the overlapping database is higher (65.1%) when compared to small (7.5%) and medium-sized enterprises (27.4%).

Figure 2 – *Percentage of responding enterprises by channel of distribution and percentage of responding enterprises by employment size class in the overlapping database.*



Source: Elaboration on ISTAT data

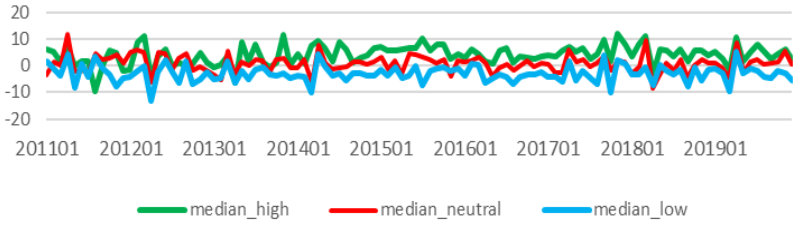
Interestingly not only descriptive statistics, but also graphical analysis are effective to examine the movement in the year-on-year growth rate of turnover in comparison with the responses to Business Confidence Survey on the evolution of the economic trend. Enterprises involved in both surveys were classified according to their evaluation of the retail trade movement (confidence). For each month three groups of enterprises were created: a group answering "High", a group answering "Neutral" and a group answering "Low"[5].

Then, for each enterprise and for each month, we calculated the year-on-year growth rate of turnover using micro data from Retail Trade Survey. Subsequently, we made a graphical analysis of the median of the year-on-year growth rates for each response group ("High", "Neutral" and "Low").

Median was preferred over mean because the value of the mean can be distorted by the outliers. In fact, since the year-on-year growth rates are calculated for each enterprise, either really high or really low values are the likely outcomes in certain months.

[5] *The question asked is "How have your SALES developed over the past 3 months?" Responses are "Sales have: 1=increased; 2=remained unchanged; 3=decreased."*

Figure 3 – *Graph of the median: years 2011-2019.*



Source: Elaboration on ISTAT data

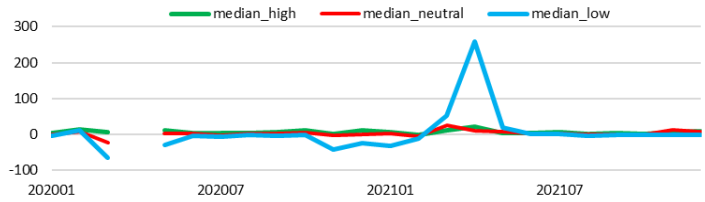
The median of the "High" response in the judgment on the sales trend generally shows higher values when compared to the median of the "Low" response. Furthermore, the graph reveals that the median of the year-on-year growth rate of the "Neutral" response is almost always placed between high and low. In particular, from 2016 the graph shows a bigger distance among the series because of a higher amount of enterprises involved in both surveys. The graphical analysis, therefore, shows a good alignment between qualitative and quantitative data.

3.1. Focus on 2020-2022

To evaluate the effect of pandemic on data, series were split: 1st period from 2010 to 2019 and 2nd period from 2020 to 2022 (see figure 4). Business Confidence data were not collected in April 2020 because of the pandemic emergency, therefore the graphs detects a discontinuity in the series. Consequently, as the year-on-year growth rates were analysed, the whole year 2021 appears to be affected by the pandemic too.

In particular, in April 2021 the enterprises giving the "Low" response still have a high year-on-year growth rate as they were closed in April 2020.

Figure 4 – *Graph of the median : years 2020-2021.*

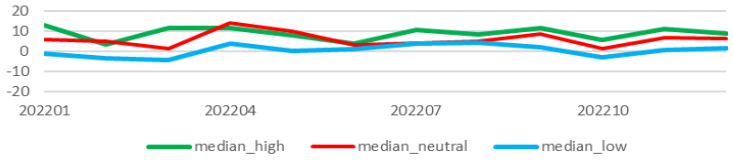


Source: Elaboration on ISTAT data

From 2022 the pandemic effect on data seems to fade (see figure 5). In the second semester of the year, the value of the median of year-on-year growth rates related to

the "Neutral" response is almost always placed between the “High” and “Low”, just like in the pre-pandemic years.

Figure 5 – *Graph of the median: years 2022.*

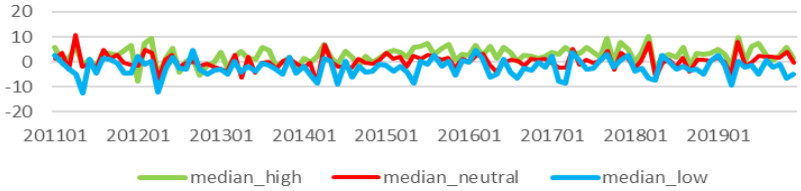


Source: Elaboration on ISTAT data

4. Forecast analysis and Median test

Experts often wonder whether qualitative surveys can be used to forecast quantitative data. For this purpose, we calculated the percentages of enterprises declaring an increase/invariance/decrease in sales in connection to the business trend expectations of the Business Confidence, in order to verify if there was an alignment between the enterprises’ forecasts and actual values declared in the following months[6].

Figure 6 – *Graph of median forecast: years 2011-2019.*



Source: Elaboration on ISTAT data

Graphs at time *t* (see figure 6) seem to confirm a good alignment between the two surveys. Therefore this alignment between the two surveys can also be used for forecasting purposes since the median associated with the "High" response of the judgment on sales forecasting, in most cases has higher values than those recorded for the “Low” response.

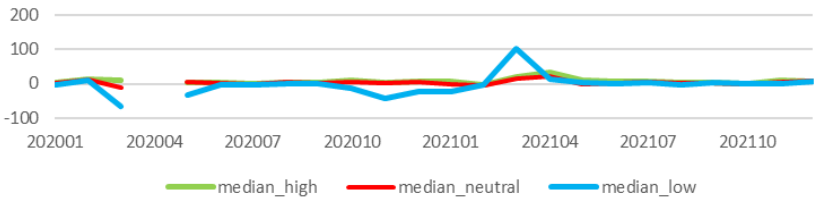
[6] *The question asked is: “How do you expect your SALES to change over the next 3 months?” Responses are “Sales will: 1=increase; 2= remain unchanged; 3=decrease.”*

4.1. Focus on 2020-2022

To allow forecast analyses, time series were split: 1st period from 2010 to 2019 and 2nd period from 2020 to 2022. Figure 7 shows a discontinuity in the time series due to the lack of Business Confidence data in April 2020.

Business expectations in 2021 were affected by the pandemic emergency as in April 2021 the enterprises that chose the "Low" response have a high year-on-year growth rate in turnover as they were closed in April 2020.

Figure 7 – *Graph of median forecast: years 2020-2021.*

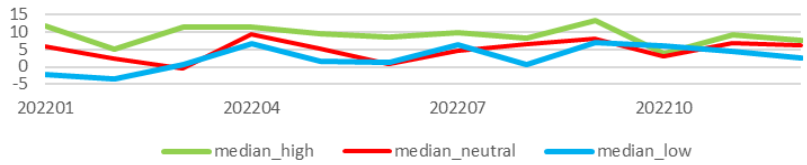


Source: Elaboration on ISTAT data

During pandemic, enterprises struggled to define their business expectations.

Starting from 2022 (see figure 8), especially from June, the value of the median of the year-on-year growth rate for the "Neutral" response mostly falls between "High" and "Low", as it happened in the pre-pandemic years.

Figure 8 – *Forecast graph of the median : years 2020-2021.*



Source: Elaboration on ISTAT data

4.2. Median test

In addition to descriptive statistics and graphical analysis, statistical tests were carry out on the medians of the year-on-year growth rates at time *t* for the "High", “Neutral" and "Low” responses for the qualitative variable “Sales trend”, starting from the dataset of enterprises involved in both surveys.

Enterprises involved in both surveys were sorted according to the way they answered to the item regarding sales performance (confidence). Then, three groups of enterprises were created for each year: the group that answered "High", the group that answered "Neutral" and the one that answered "Low".

Hence, for each enterprise and for each year, we calculated the year-on-year growth rate of turnover, using data stemming from Retail Trade Survey. We used the non-parametric Wilcoxon test, which verifies the hypothesis that two samples are drawn from populations with coincident medians.

Table 1 – *Median and median test (year 2011-2022).*

	MEDIAN			TEST ON MEDIAN(*P-VALUE*)		
YEAR	High	Neutral	Low	High – Low	High – Neutral	Neutral –
2011	3.4	1.0	-2.4	0.000	0.204	0.000
2012	0,9	**1,7**	-3,4	0.010	0.477	0.000
2013	0.5	-0,2	-3.3	0.000	0.400	0.000
2014	5.0	0.2	-3.2	0.000	0.001	0.000
2015	6.3	1.7	-1.8	0.000	0.000	0.000
2016	3.8	0.8	-2.6	0.000	0.000	0.000
2017	5.7	0.9	-2.5	0.000	0.000	0.000
2018	4.5	-0.6	-2.7	0.000	0.000	0.041
2019	4.3	1.3	-2.2	0.000	0.000	0.000
2020	6,8	3,1	-5,6	0.000	0.000	0.000
2021	5.3	4.7	-0.3	0.001	**0.962**	0.004
2022	8.8	6.1	0.6	0.000	0.000	0.000
2011-2022	5.4	1.5	-2.5	0.000	0.000	0.000

Source: Elaboration on ISTAT data

Test results confirm a good alignment between the two surveys. Apart from year 2012, the alignment of the median of the year-on-year growth rate for the different groups was verified, i.e. the median of the year-on-year growth rates of the “High” group is always higher than that of the “Neutral” group, which is higher than that of the “Low” group. In the very first years and especially when comparing the “High-Neutral” groups, tests accept the null hypothesis of equality of the medians.

This is mainly due to the smaller amount of enterprises involved in both surveys in the first years considered, when compared to the last period. Year 2021, on the other hand, remains a critical year because, as shown also by the graphic analysis, it displays a particular trend due to the pandemic.

5. Conclusions

Analyzing the two survey samples highlights that the number of units involved in both surveys has increased over the years and it shows a good alignment between qualitative and quantitative data.

The graphical analysis of the median of the year-on-year growth rates of turnover in connection to the responses to the Business Confidence Survey confirms a good alignment between qualitative and quantitative data.

The forecast analysis shows a good alignment of the two surveys, since business expectations on turnover recorded by the Business Confidence Survey appear to have a similar pattern of the turnover declared by the same enterprises in the quantitative survey. Findings of the graphical analysis appear to be well substantiated by the results of statistical tests on the median of the year-on-year growth rates at time *t* for the "High", "Neutral" and "Low" responses of enterprises involved in both surveys. Apart from a discrepancy recorded in the first years because of a small amount of enterprises involved in both surveys, the results of the tests confirm a good alignment between the two surveys.

References

LUI S., MITCHELL J., WEALE M. 2011a. Qualitative Business Surveys: Signal or Noise? *Journal of the Royal Statistical Society*, Vol. 174, No. 2, pp. 327-348.

LUI S., MITCHELL J., WEALE M. 2011b. The Utility of Expectational Data: Firm-Level Evidence Using Matched Qualitative-Quantitative UK Surveys, *International Journal of Forecasting*, Vol. 27, No. 4, pp. 1128-1146.

MARGANI, P., ORSINI D. 2020. L'indagine sulla Fiducia delle Imprese Manifatturiere e l'indagine sul Fatturato e Ordinativi dell'industria a Confronto: Un Dataset Integrato sulle Imprese Italiane del Comparto Manifatturiero negli Anni 2005-2019. *Working papers 5/2020*, Roma: Istituto Nazionale di Statistica.

IPPOLITI M., MARTONE L., SARTOR F. 2021. Trade Surveys: Qualitative and Quantitative Indicators, *Rivista Italiana di Economia, Demografia e Statistica*, Vol.75, No 4.

Maria Rita IPPOLITI, Istat, marippoliti@istat.it
Luigi MARTONE, Istat, martone@istat.it
Fabiana SARTOR, Istat, sartor@istat.it

STATISTICAL REGISTER OF PLACES: OPPORTUNITIES FOR SUSTAINABLE AND CLIMATE CHANGE RELATED INDICATORS

Damiano Abbatini, Tiziana Clary, Raffaella Chiocchini, Davide Fardelli, Angela Ferruzza, Luisa Franconi, Fabio Lipizzi, Stefania Lucchetti, Stefano Mugnoli, Enrico Orsini, Andrea Pagano, Alberto Sabbi, Gianluigi Salvucci, Assunta Sera, Pina Ticca

Abstract. Statistical information related to Climate Change and Sustainability ask for an integrated approach related to economic, social, environmental and institutional goals, from global to local and from local to global to leave no one behind. The statistical measures are useful to build a common language crucial for monitoring. The Register of places is a complex system with several components. The challenge is the production of spatial information able to respond to the heightened need of detail statistical data. The construction process is complex and faces several issues first of all the very high number of objects involved and the integration of components stemming from different sources independent from each other. The final integrated product allows the possibility of geo-referencing information for flexible outputs. This information has the potential to increase statistical measures and analyses related to Climate Change and sustainability.

1. From global to local and from local to global: sustainability and climate change

The revolution of the integrated frameworks (*2020 UNECE CC Core indicators, 2022 UN FDES Indicators, 2023 Measuring Hazardous Events and Disasters (MHED) Core indicators, UN-IAEG SDGs Indicators)* related to Climate Change and Sustainability proposes that the economic, social, environmental and institutional goals have to be developed considering an integrated approach from global to local and from local to global to leave no one behind. The statistical measures are useful to build a common language crucial for monitoring. In this paper these concepts will be considered: the importance of the Statistical Register of Place (RSBL) and of the GIS for the statistical measures and analyses on Climate Change and Sustainability will be made explicit through examples.

In the context of the UN – Inter Agency Expert Group on SDGs (UN-IAEG-SDGs), Working Group on Geospatial Information produced the Geo White Paper on disaggregation by geographic location and Statistical Commission adopted SDGs Geospatial Roadmap. According to these documents, producing and using geographically disaggregated SDG data is essential. The disaggregation of statistical

measures by geographic location provides a mechanism to achieve a greater analytical potential of the data, turning them into a high quality, accessible and timely tool for the generation of information that allows for more accurate and real-time decision making. Disaggregation by geographic location, alone or in combination with other dimensions (sex, age, income, migration, disability status), allows uncovering the existing hidden societal disparities, bringing to the fore of analysis vulnerable, precarious and marginalized segments of the population and territory. The effectiveness of the statistical measures depends not only on the statistical design of the data, but also on an adequate geographical disaggregation that can demonstrate geographical variations of social, economic, environmental phenomena. This involves the creation of a spatial data infrastructure enabling standardized location references for mapping spatial location to statistical data units. The statistical data should be referenced to the finest geographical scale possible, down to a geographic coordinate. The disaggregation that can be achieved in the calculation of the climate change and sustainability indicators or in the generation of statistical data depends on the territorial/administrative geographic units or geographies defined for statistical purposes. The geographical units organized in a standard hierarchical classification allow the statistics to be disaggregated through the spatial reference codes assigned to the primary observations. The assignment of a unique identifier to each location area allows linking with other statistical and geospatial data associated with the same geographic space. The geocoding of statistical data considerably expands the analytical possibilities, including integrating them into indicators and other data, but also analyzing the data from a geographical point of view. This association of geographical reference to statistical data allows statistics to be produced for a wide range of applications and geographical contexts.

The provision of these common geographies and their life cycles allow the generation of statistical data in a consistent manner, through cartographic grids or units with administrative or statistical boundaries. Likewise, these allow statistical data to be aggregated/disaggregated at different levels for the purpose of their integration. Common geographies help to build a common language, to integrate among domains and to analyze interlinkages that can make explicit trade-off or synergies in the phenomena and in the actions. Concretely the territorial disaggregation could be the basis to integrate economic, social, environmental and institutional domains to produce statistical information on Climate change and Sustainability (Climate change Reports on Urban Areas, Urban environmental Reports, SDGs Report) and the RSBL could be the key element in the construction of statistical measures and specific analyses to build the common language from the global to the local. In any case is important to clarify that the use of administrative

data and of Statistical Registers is essential but is a big challenge for methodological and institutional reasons related also to confidentiality issues.

2. From administrative to statistical data, a big challenge for sustainability and climate change: Statistical Register of Places (RSBL)

Istat is changing its production processes aiming to an Integrated System of Statistical Registers: at the very heart of it lies the Statistical Register of Places (RSBL). The geographical statistical information of RSBL integrated with the statistical information of other Registers (socio-demographic or economics) has an increasing potential to consider statistical measures related to climate change and sustainability, for instance, considering the following thematics:

- ex-ante analysis of areas presenting high risks of flood, or earthquake, or fires;
- ex-post analyses of areas hit by natural disasters;
- production of tool-set of statistical indicators ready for hazardous events;
- green cover in urban areas using high resolution remote sensed images, via the production of vegetation indices, and extraction of statistical information linked to the total vegetation cover in the major Italian urban centres that are very useful to consider sustainability and climate change indicators;
- air pollution analysed considering very detailed territorial area and linked with exposed population;
- land pollution areas linked with exposed population;
- analyses related to land consumption, protected areas, energy consumption.

To understand better the increasing potential of these analyses RSBL is illustrated. RSBL is a complex system with several components; for each register component, variables are being built detailing several characteristics of the entity under study and information on their quality. The challenge is the production of spatial information able to respond to the heightened need of detailed statistical data integrating the different components. To improve integrated statistical analyses, the detailed geography of the statistical units of all the social and economic statistical registers is an essential condition. RSBL has the goal to release detailed geography and to use them to integrate social and economic data also in statistical surveys.

The components of RSBL are:

- Territorial Information System of Administrative and Statistical Units (Situas) related to municipalities and administrative and statistical territorial units;
- Enumeration areas composed by many different archives of geographic data for 800000 georeferenced enumeration areas and 1,1 million micro-zones (infrastructures, green areas, ….);

- Addresses and geographic coordinates: many administrative archives of data for 30 million CUI (Unique Identification Code) of addresses, their geographic coordinates and related Quality indicators;
- Buildings and dwellings: based on administrative archives from Real Estate Registry, from Cadastral agency, from geographic agencies and from open sources, referred to 29 million buildings of which 14.4 million are residential.

The integration process has seen different methods applied to different entities in order to reach the highest possible quality. First results are the production of a preliminary 1km population grid, of dwelling data and of enumeration areas.

3. RSBL SITUAS: a forthcoming dynamic portal to enquire structure and changes of territories

At the heart of the Register of Places lies the Territorial Information System of Administrative and Statistical Units (Situas). The System represents the continuously updated version of the Register as far as administrative and statistical units are concerned. The System allows enquiring the list of active territorial units at any specific date since the foundation of the State. To date, Situas allows to gain information on municipalities, provinces, metropolitan cities, regions and their aggregations as well as NUTS, Labour market areas, Industrial districts, Functional urban areas (EU Tercet Regulation). Next to these, information on various units dimensions are present as well as their classifications (e.g. for municipalities, coastal areas, degree of urbanization, classifications for policies, etc.).

All these data and typologies are going to be released via a dedicated portal on Istat web site. Situas portal will also deliver to users advanced functionalities such as the history of municipalities since 1861, the search for codes or names of administrative units, information on variations occurred in specific period of time and the new service of the reconstruction of codes since 1991. The portal will be equipped also of a Glossary of terms and a Download area with various types of side information: shape file of the units, maps, information at censuses, specific tables, etc. The portal, currently in its final developing stage, will also allow machine-to-machine dialogue with other systems in order to satisfy the needs of agencies, organisations and ministries.

4. RSBL Enumeration areas

The Census Cartography (BT) represents an updated photograph of the territorial boundaries adopted (enumeration areas, inhabited localities and productive areas) which includes new urban areas. Therefore, the BTs are, the representation of geographic objects describing both the settlement plot of the country and its evolution from the medium to long term, although with some approximation. The

definition of the Census Maps is strictly a matter for local authority[1]. The main goal of the updating activity is to define a plot of the National territory connected to the changes that have occurred in terms of urban expansions, new building aggregates and population data. From 2018, the Italian population census survey has marked the definitive transition from the traditional enumeration to a "register-based" system built on the so-called Permanent Population and Housing Census (UNECE, 2021). The change in the census survey strategy has also modified the use of the Census Maps at Municipality level[2].The new edition of the 2021 Census Maps is not used as a basis for census data collected in a specific year, but rather for the dissemination of the 2021 sub-municipal data. It is noteworthy that the new Enumeration Areas coverage, called "microzones", inherits the rules and geometric objects of the 2011 Census Maps; but to better spread sub-municipal data, it is necessary to improve the quality of the drawing and increase the internal homogeneity of the polygons. Analyzing the absolute values, the number of EAs passes from 402,677 (2011) to about 800,000 (2021). 2021 EAs have many new elements compared to the past, even if they have not changed their main characteristics; it remains valid the municipal validation process that Istat planned during the past census surveys. The EA delineation process consists of the following main steps:

- Automatic integration of the 2011 EAs polygons with other thematic cartography (open and commercial), following precise overlapping priorities.
- Verification and geometric cleaning of EAs. This operation adds further features to the EAs. In addition, the layers obtained were examined by a photo-interpretation operation to verify the most recent urban expansions.

In phase 1, many GIS geo-processing tools were implemented to accelerate the production process and improve its quality. In the second step of the workflow, some tools were developed to reduce editing errors and inconsistencies[3].

The overlapping methods are based on the properties of topological spaces and on the operations (inclusion and intersection) between geometric objects associated with them[4].The general rule for the new 2021 EAs provides that new polygons are drawn just within a pre-existing EAs 2011. The new polygons are drawing according to their importance on the territory, such as for airport areas, hospitals, schools, town halls, etc. This new layer constitutes, therefore, the overcoming of the traditional

[1] According to the provisions of art. 39 of the Presidential Decree 223/89, articles 9 and 10 of the Registry Law (Law 24 December 1954, n. 1228), Chapters VII and VIII of its Implementation (Presidential Decree 30 May 1989, n. 223), local authorities are obliged to update the Census Maps, taking into account all the changes of their territories (Istat, 1992).

[2] See the General Census Plan www.istat.it/it/files//2018/09/PGC-POPOLAZIONE-ABITAZIONI-2022.pdf.

[3] Cfr. Laaribi A., Peters L., 2019

[4] For a discussion of the topological operations applied to GIS see Egenhofer and Franosa, 1991

census Enumeration Areas, used almost exclusively for census survey; it gains, some specific characteristics, which make it suitable for other purposes, also related to Climate Change and Sustainability analyses. Each new EA 2021 is identified following a criterion of by its land cover/use homogeneity. characterizes the area in terms of use[5] and land[6] cover (Directive EU 2007/2).

Figure 1 – *Enumeration Areas 2021: examples of land classification.*

Source: Istat elaboration on Istat and AGEA data

Undoubtedly, the main items represented on Istat cartography are inhabited localities (Istat, 1992). They can be divided into three categories:

- Urban centre: groups of houses, distant from each other no further than 70 meters and connected by roads. They must have public services;
- Inhabited nucleus: small settlements of grouped houses distant from each other no further than 30 meters without public services; they must include at least 15 households and 15 buildings;
- Production plant: a locality in a non-urban area, with at least 10 firms or 200 employees; it must be large at least 5 hectares.

The remaining territory represents the extra-urban areas. 2021 EAs are designed on the base of their specific location and topological relations with contiguous polygons. In accordance with international standards[7] the geographical characteristics of the EAs 2021 are closed polygons, cover all the municipal territory, and are consistent with the administrative hierarchy: Region, Province, Municipality, inhabited localities.

[5] https://www.eea.europa.eu/help/glossary/eea-glossary/land-use
[6] https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Glossary:Land_cover
[7] Cfr. UNITED NATIONS, 2009, and Egenhofer and Franzosa, 1991.

5. RSBL-Addresses

The addresses component of RSBL should include all the addresses in the national territory. Every address has been admitted and identified in RSBL with a unified address code (CUI). The attribution of a code simplifies the integration with other registers and avoids linkage errors. Every CUI has a geographic coordinate and/or enumeration area. The geographic information is always accompanied with quality indicators both of coordinate and of geocoding. An important step is the validation of the association between address, enumeration area and municipality. The geocoding process has a key role, because it allocates every address in an enumeration area and in one municipality. High quality process will implicate not distorted output at the municipal sub-level for statistical units. The enumeration area for addresses with coordinate has been calculated according to several level of coordinates (punctual/interpolated/approximated) available in RSBL via GIS tools.

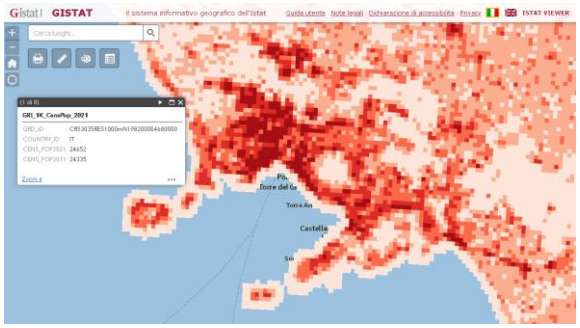
Figure 2 – *Addresses and Geographic Coordinates.*

The use of coordinates allows and strengthens the consistency of the association between each street number within the current enumeration area, within the effective municipal boundaries and it will allow maintaining the correct association over time albeit territorial variations.

The geocoded addresses from 2011 Population Census and the National Archive of Addresses of Urban Streets (ANNCSU) have been used to strengthen the geocoding process and to fill uncovered areas. 30 million of CUI, georeferenced at 80% with geographic coordinates, are now available. The Integration of RSBL and RBI (Register of Individuals) resulted into 97,6% of resident population geocoded at enumeration area. Such integration allowed provisional Population grids statistics see Figure 3; the final one is going to be elaborated in the next months.

Population grids are an alternative to population statistics for administrative areas and are a powerful tool to describe society and to study the interrelationships between human activities and the environment. They are particularly useful for analyzing phenomena, and their causes, which are independent of administrative boundaries, such as flooding, urban sprawl, air pollution.

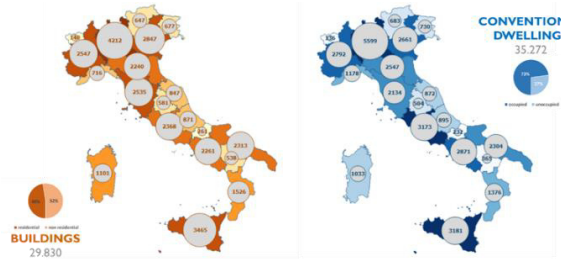
Figure 3 – *Population Grids (2021 provisional).*



6. Building and Dwelling Register

The Building and Dwelling Register component of RSBL (Building in the following) considers different sources: the Cadastral Administrative Archive, Regional Cartography, National Geoportal, Open Street Maps. In the Building Register, individual units, located in residential or non-residential buildings, are always associated with at least one of these elements: individuals and legal entities holding rights (ownership, rental, etc.), cadastral categories that identify the intended use, cadastral data, geographic coordinates of the buildings, and finally, addresses.

Figure 4 – *Building and Dwellings Register (data expressed in thowsand of units).*

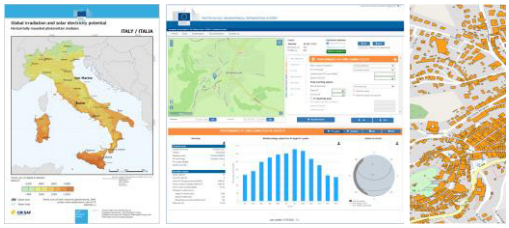


By means of these elements, it is possible to accurately position population and other statistical units on the territory with a high level of quality (Figure 4).

One of the main strengths of this product is the availability of geographic coordinates to which the examined statistical units are assigned. This new technical capability offers numerous possibilities for statistical production: spatial analysis to identify geographic patterns, clusters, or geographical distributions; geocoding data from different sources using geographic location as a key. For example, it is possible to more accurately measure the proportion of buildings and population located within a protected area; finally, geographic coordinates enable the creation of informative

visualizations such as thematic maps or density maps. Moreover, the use of geographic coordinates can enable the creation of new contextual variables that enrich spatial analysis and further explore the relationships between statistical units and the surrounding environment. A brief example is the use of the distribution of buildings across the Italian territory to calculate the photovoltaic potential of specific areas. This approach is interesting and can be a valid method to assess the opportunity and feasibility of installing photovoltaic systems.

Figure 5 – *Integrating new variables and indicators using GIS and coordinates.*



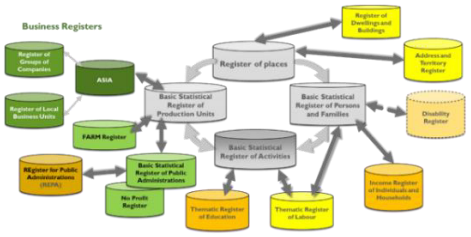
credits: EU Photovoltaic Information System (PVGIS), RSBL Building Register

To calculate the photovoltaic potential, it is possible to use models or simulation tools that take into account parameters such as average solar radiation, solar panel efficiency, shading, and other site-specific factors (first two images on the left in figure 5 by PVGIS). An approximate estimation of the amount of solar energy that could be generated from a given set of buildings in a specific area (the third image on the right of figure 5, by RSBL Building Register) could be obtained. Analysing the photovoltaic potential based on the distribution of buildings is only part of the equation.

7. Integrated System of Statistical Registers (ISSR)

The Integrated System of Statistical Registers (ISSR) is the basis for surveys and statistical production. It integrates information relating to: (i) individuals, families and cohabitation; (ii) economic units; (iii) places; and (iv) activities as showed in figure 6. In this system, every unit is linked to places through RSBL. The base registers are connected by codes and are maintained updated over time using mainly administrative sources. Therefore, RSBL assume a dual role: (i) geo-referencing and/or geocoding the statistical units (demographic/economic) and (ii) spatial data production (e.g. surfaces, altitudes, distances, contiguities, statistics on buildings, population grid, etc.).

Figure 6 – *Integrated System of Statistical Registers (ISSR)*



The register has been built only once and it is kept updated. RSBL is a multidimensional register integrating components with heterogeneous nature. RSBL, with the other Registers, provides a bridge between statistical units, such as individuals and families in RBI and economic in the Labour Register (RTL) and business Register (ASIA). Methodological approaches have been used to consider the interlinkages and integration by code taking in account confidentiality issues: every kind of information with geographic coordinates could be integrated. The final integrated product will allow the possibility of geo-referencing information for flexible outputs. Climate change and sustainability statistics can be then improved considering anthropic pressure.

8. WebGIS tools to share geo-referenced statistics

The Integration of Statistical and Geospatial Information is essential when dealing with Sustainability and Climate Change statistics.

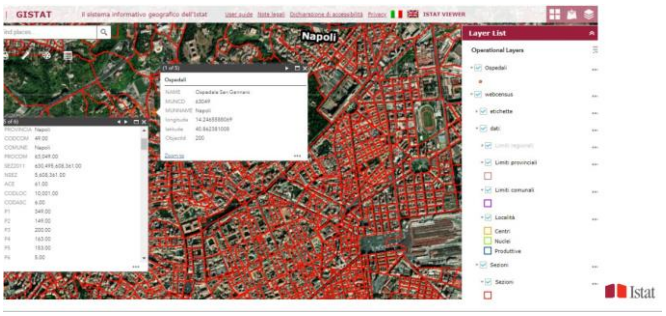
The WebGIS tools and the GeoPortals are essential components to share and to make interoperable geospatial statistics on the Web. Together with the adoption of technical standards and according to national and international frameworks, they contribute to the development of a Spatial Data Infrastructure (SDI). The metadata are also fundamental to make the spatial data discoverable and re-usable; so a GeoMetadata Catalogue should be implemented to be the access point for a Statistical Geoportal.

Using those tools, geospatial data and geo-referenced statistics can be published through the WebMapServices technology; those can be consumed by Web Applications to visualize, to geographically navigate, to query and to analyse geospatial data. In that way, the users can interact with geospatial statistics. Based on the overlay GIS principle, users can combine several layers, coming from Internet WebMapServices or from local computers, to enhance the traditional statistical analysis with spatial operators (distance, adjacency, inclusion, etc.).

The power and the importance of the Integration of Statistical and Geospatial Information is related to the data analysis (such as affected areas, proximity,

isolation, etc.) but also to the creation of new statistical information, not possible without the combination of the two data sources (See fig. 7).

Figure 7 - *Census mapping cartography and indicators with hospitals overlaid.*



9. Sustainability, Climate Change, territorial and geographic data: statistics to not leave behind are necessary and possible

Geostatistical and territorial analyses are then integration factors from global to local and from local to global and are essential to build a common language: considering geostatistical analyses the integration among economic, social, environmental, institutional domains could improve looking forward to Climate change and Sustainability. RSBL provides geographic and administrative data, including identifying codes, geographic boundaries, territorial characteristics and other information useful for analysis and the production of official statistics on Climate Change and Sustainability. RSBL is currently used as a basis for statistical activities and it is a fundamental tool for processing, analyzing, and representing official statistics on the Italian territory. Coordinates allow, infact, to consider the geographical characteristics of an area, such as the presence of rivers, mountains, lakes, or national parks; this information can influence various aspects, such as environmental quality, tourism activities, or the availability of natural resources.

References

EGENHOFER M.J., FRANZOSA R.D. 1991. Point-Set Topological Spatial Relations. *International Journal for Geographic Information Systems*, Vol. 5. No. 2, pp. 161-174.

ISTAT. 1992. Anagrafe della Popolazione: Metodi e Norme. Roma: Istituto Nazionale di Statistica.

LAARIBI A., PETERS L. 2019. *Gis and the 2020 census: modernizing official statistics.* California: Esri Press.

UNECE. 2021. *Guidelines for Assessing the Quality of Administrative Sources for Use in Censuses.* Geneva: United Nations.
UNITED NATIONS. 2009. Handbook on Geospatial Infrastructure in Support of Census Activities. Studies In Methods. *United Nations Publications*, Series F, No.103.
UNITED NATIONS. 2016-2023. Inter-Agency and Expert Group on the Sustainable Development Goal (IAEG-SDGs) Working Group on Geospatial Information. *United Nations Publications.*

Damiano ABBATINI, Istat, abbatini@istat.it;
Tiziana CLARY, Istat, clary@istat.it;
Raffaella CHIOCCHINI, Istat, chiocchini@istat.it;
Davide FARDELLI, Istat, davide.fardelli@istat.it;
Angela FERRUZZA, Istat, ferruzza@istat.it;
Luisa FRANCONI, Istat, luisa.franconi@istat.it;
Fabio LIPIZZI, Istat, lipizzi@istat.it;
Stefania LUCCHETTI, Istat, lucchetti@istat.it;
Stefano MUGNOLI, Istat, mugnoli@istat.it;
Enrico ORSINI, Istat, enrico.orsini@istat.it;
Andrea PAGANO, Istat, andrea.pagano@istat.it;
Alberto SABBI, Istat, alberto.sabbi@istat.it;
Gianluigi SALVUCCI, Istat, salvucci@istat.it;
Assunta SERA, Istat, sera@istat.it;
Pina TICCA, Istat, ticca@istat.it

RE-ENGINEERING ENVIRONMENTAL DATA COLLECTION IN CITIES[1]

Domenico Adamo, Gianpiero Bianchi, Lucia Mongelli

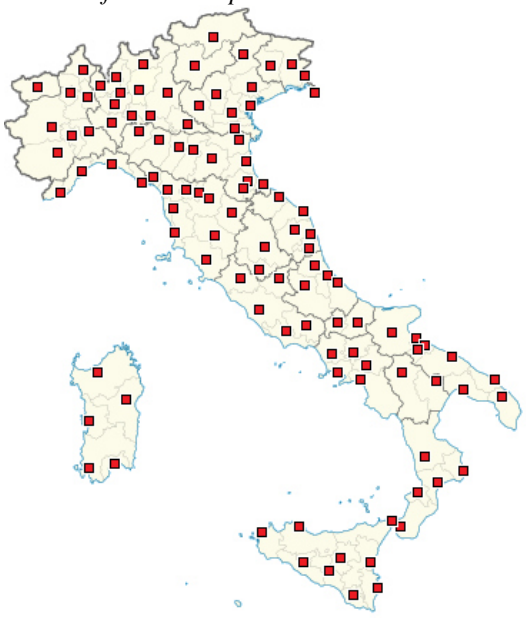
Abstract. The work provides an information framework to support the monitoring of the state of the urban environment and the activities carried out by administrations of provincial capitals to improve quality of the environment in cities. In particular, the "Survey on urban environmental data" is carried out annually by Istat, is included in the National Statistical Program in force and collects environmental information about all Italian capital municipalities. The work describes the study and design of a new validation process, according to a generalized perspective, which includes automatic procedures for checking the consistency of the data collected, monitoring the processing and interaction with the Municipal Statistics Offices. A representation of the rules in formal logic will be adopted, through a metalanguage in order to support an automatic approach. The result is an integrated system of generalized services that works formally and therefore can be used in different contexts. In order to maintain the quality standards of the data disseminated by the survey, a study on the administration of questionnaires on different editions of the survey was designed.

1. A brief overview of the Survey on urban environmental data

Data on the urban environment is a multi-source statistical process, organized in 8 thematic modules: air quality, urban waste, mobility, noise, energy, urban green, water, eco-management, which produces environmental indicators for 110 Italian cities (95 provincial capitals, 14 metropolitan city capitals and the municipality of Cesena, which participates on a voluntary basis), Figure 1.

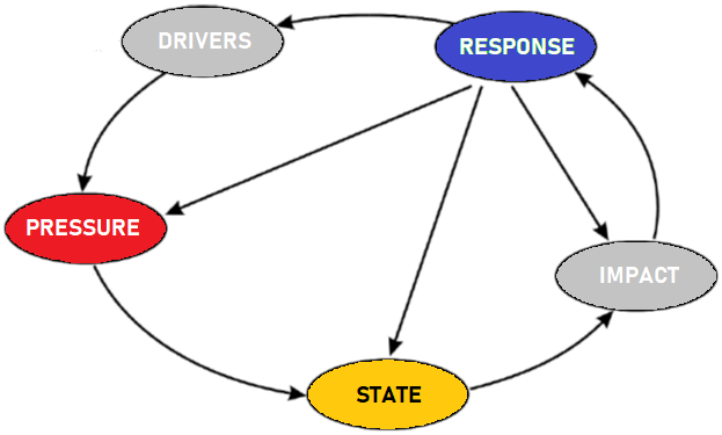
[1]The paper is the result of the common work of the authors. In particular, paragraph 1 is attributed to Domenico Adamo, paragraphs 2 and 3 are attributed to Lucia Mongelli, and paragraphs 4 and 5 are attributed to Gianpiero Bianchi. The conclusions (ph.6) are a joint work of all the authors.

Figure 1 - *Spatial distribution of the municipalities involved in the survey.*



The process provides a comprehensive information framework for monitoring the quality of the urban environment, status and pressure indicators (Adamo et al., 2020), according to the DPSIR model, developed by the European Environment Agency (Fig.2, Bosch et al., 1999) and environmental policies implemented by local governments (so-called response indicators, such as directives, plans, technology development).

Figure 2 – *The DPSIR model.*



The DPSIR model consists of: determinants (agriculture, population, and transport), pressures (waste, emissions), state (water, air), impact (costs, pathologies), responses (directives, policies, technology development); it represents a tool capable of evaluating the causal chain leading to environmental alteration, (measured through environmental indicators).

The urban environment survey is part of the National Statistical Programme (NSP), managed by SISTAN and updated every 3 years. Being included in the NSP as a public interest investigation, data collection is carried out by law and the response is mandatory for reporting units.

The NSP also provides the legal basis for the use of administrative data for statistical purposes. Specific agreements are concluded between Istat and the data controllers to define the characteristics and timing of the provision of data, within the SISTAN regulatory framework and in accordance with the rules for the protection of personal data.

The use of administrative data reduces costs and burden for respondents.

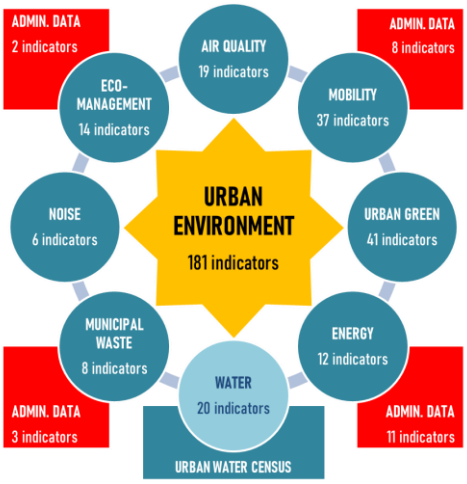
The survey data are collected through thematic questionnaires: Air quality, Eco-management, Noise, Urban waste, Water, Energy, Urban green, Mobility. The theme Water from 2018 is taken from the urban water census, which covers the entire national territory.

The survey data are then supplemented with four modules on particular topics and provided separately by the data controllers.

The whole process handles about 500 elementary variables, to produce the 181 indicators (2020 version), 13% of which are based on administrative data.

Figure 3 summarizes the indicators by theme.

Figure 3 - *Diagram of the data source.*



All indicators are disseminated by municipality, and aggregate estimates are provided by geographical area.

Some indicators are part of the set of statistical measures of Istat for the monitoring of SDGs (Sustainable Development Goals) in Italy, consisting of 17 points, identified by the UN in 2015 that aim to safeguard the planet and the welfare of its inhabitants, with a horizon that reaches up to 2030.

2. Data collection

Since 2008, Istat has introduced important methodological innovations for the "Survey on urban environmental data" with the aim of improving, standardizing the encodings and formats of variables, and simplifying the data collection process.

According to the provisions of the Code of Digital Administration in 2005 (d.lgs 82/2005 and subsequent additions and changes) which provides that data must be transmitted to Istat in computerised mode, the CAWI technique (Computer Assisted Web Interviewing) has been introduced for data acquisition in electronic format, through the Gino++ (Gathering information Online) portal of Istat (Torelli, R. 2011).

GINO++ is a generalized software that allows not only the collection of data but the complete management of surveys via web, creation of web questionnaire, controlled acquisition of data online and/ or file upload, custom site preparation for the survey, monitoring of the status of questionnaires and records, contacts for reminders and reminders, reports.

In addition, on the home page of the Gino Portal (https://gino.istat.it/amburb/) respondents find the support material to fulfill all the obligations provided by the survey: the description of the survey, the detail of the law for the response obligation, instructions for accessing and filling in the questionnaire, IT requirements, FAQs.

The data are collected by the Municipal Statistics Offices, which, through a pre-survey (limesurvay), identify in the Administrations to which they belong a coordinator and one or more persons referencing the survey topics, which are provided with personal credentials to access, enter, modify and save data.

Depending on the topics, the reference persons collect the data directly from the municipalities, or request them from other local authorities (e.g. public transport companies).

Through GINO++ the Municipal Statistics Offices, the coordinators and the referents of the different topics, can send the data through the direct compilation of web questionnaires (CAWI). To improve the completeness and consistency of data entered in the Gino++ data acquisition system, automated checks have been implemented to report anomalies, to prevent inconsistent or invalid or out-of-range data entering and sending questionnaires with missing answers.

An additional monitoring function allows to constantly monitor the activity of respondents, from recording to sending data, including reporting any violations of consistency rules.

3. Process innovations. Validation automation

In line with the objectives of Istat to provide the country with correct statistical information and to innovate the various processes of production of statistical information, consistent with the progressive digitalization of data collection processes, it became necessary to design the use of innovative solutions by re-engineering the validation phase of the questionnaires, with the implementation of additional automatic control rules different from those already provided for by the validity and internal consistency checks of the Gino ++ system.

During this experimental phase, in order to ensure the regular conduct of the survey while maintaining the quality standard of the data collected, a different frequency is expected for some thematic questionnaires, which do not produce indicators intended for institutional dissemination.

For the 2023 edition the thematic of the questionnaire are questionnaires: Air, Mobility, Municipal waste, Noise, Urban green. For the 2024 edition, however, the thematic of the questionnaire will be questionnaires: Air, Eco-management, Energy, Mobility, Urban green.

This innovation could be completed during the two editions of the survey, 2023 and 2024.

The new rules manage, at least in part, the aspects so far entrusted to the review by monitors: in particular the interception of measurement errors, discontinuity of the time series and other anomalous values. These rules would work on a dynamic basis, by comparing the data collected in the current edition with those validated and disseminated in previous editions.

In table 1 the phases of the investigation after the re-engineering.

Table 1 - *Survey stages. After re-engineering.*

		MUNICIPAL OFFICES	STAGES	ISTAT	
				Data collection	Environmental statistics
			Survey design	Survey organization, Implementation of CAWI questionnaires	Information contents and metadata management
Questionnaire states	**Initial** - before taking over by reference person	Registration	**Data collection**	Controlled acquisition through Gino electronic questionnaire with rules Automatic control of: -) measurement errors, -) historical series discontinuities -) other abnormal values Monitoring of survey operations	Assistance to respondents and to data collection staff
	In process - after first opening by reference person	Data entry			
	Sent - after completion by reference person				
	Checked - after preliminary check			Automated return on respondents	
			Data processing		Data editing and validation
			Data dissemination		Data analysis and reporting

4. A generalized data editing for error detection

A generalized editing system allows checking the consistency of the data collected with respect to the check plan for survey data, with intra-record and inter-record rules. Furthermore, the editing system identifies the inconsistency and redundancy in the rules set.

The application classifies exact and incorrect questionnaires, identifying the collected units involved in violated edits together with the fields involved in the violation of the rules.

The application uses a customizable metadata table to apply the editing plan.

This table contains the following information:

- The type of rule: Validity, Logic, Mathematics and Logical-Mathematics;
- The textual description of the rule;
- The representation in formal logic, that is through a meta-language understandable to the editing application;
- Typology of rule between hard (blocker rule) or soft rule (non-blocker rule);

- A hierarchy of rules, to indicate to the application the relationship and the order of control of the rules. In the case of a violated rule, all the related rules (which are a specialization of the rule itself) of a subsequent order can be put directly to violated.

This section provides some useful concepts for the representation of rules in formal logic. In particular, it provides the definition and representation of a list of check rules and for understanding how to transform the textual rules, defined in the examples described below, in compatibility or incompatibility rules when they are translated in a formal language (Bruni and Bianchi, 2012). Rules are expressions typically used to detect, among a possibly large set of elements, the ones verifying some conditions. It is convenient, in order to verify a set of checking rules, to express them using a structure based on propositional logic.

Propositional logic, sometimes called sentential logic, can be considered a grammar for exploring the construction of complex sentences using atomic statements as building blocks connected by logical connectives. In this type of logic, logical formulas (sentences, propositions) are built up from atomic propositions that are unanalysed. The meaning of these atomic propositions will be known for the specific domain of application. A truth assignment to such atomic propositions will determine the truth value of the whole formula according to the truth rules of the logical connectives. The traditional (symbolic) approach to propositional logic is based on a clear separation of the syntactical and semantical functions.

The syntax deals with the laws that govern the construction of logical formulas from the atomic propositions and with the structure of proofs. Semantics, on the other hand, is concerned with the interpretation and meaning associated with the syntactical objects. A basic aspect of propositional calculus is that inferences are obtained as purely syntactic and mechanical transformations of formulas. The set of primary logic connectives $\{\neg, \vee, \wedge\}$, together with the brackets () to distinguish start and end of the field of a logic connective.

- The set of proposition symbols, such as x1, x2, ..., xn.
- The only significant sequences of the above symbols are the well-formed formulas (WFFS). An inductive definition is the following:
- A propositional symbol x or its negation ¬x.
- Other WFFS connected by binary logic connectives ($\vee$, $\wedge$) and surrounded, in case, by brackets.

Both propositional symbols and negated propositional symbols are called literals. Propositional symbols represent atomic (i.e. not divisible) propositions, sometimes called atoms. An example of WFF is the following:

$$(\neg x_1 \vee (x_1 \wedge x_3)) \wedge ((\neg(x_2 \wedge x_1)) \vee x_3) \tag{A.1}$$

A formula is a WFF if and only if there is no conflict in the definition of the fields of the connectives. In order to simplify the exposition, we will henceforth assume that all our formulas are well formed unless otherwise noted.

The calculus of propositional logic can be developed using only the three primary logic connectives above. However, it is often convenient to introduce some additional connectives, such as ⇒ which is called *implies*.

They are essentially abbreviations that have equivalent formulas using only the primary connectives. In fact, if S_1 and S_2 are formulas, we have:

$(S_1 \Rightarrow S_2)$ is equivalent to $(\neg S_1 \vee S_2)$.

The elements of the set {T, F}(or equivalently{1,0}) are called truth values with T denoting True and F denoting False. When all the proposition symbols of a formula receive truth values, the truth or falsehood of that formula is obtained according to the truth rules of the logical connectives (considering their appropriate meaning of "not", "or", and "and"). As an illustration, consider the formula (A.1).

Let us start with an assignment of true (T) for all three atomic propositions x_1, x_2, x_3. At the next level, of sub formulas, we have $\neg x_1$ evaluates to F, $(x_1 \wedge x_3)$ evaluates to T, $(x_2 \wedge x_1)$ evaluates to T, and x_3 is T. The third level has $(\neg x_1 \vee (x_1 \wedge x_3))$ evaluating to T and $((\neg (x_2 \wedge x_1)) \vee x_3)$ also evaluating to T. The entire formula is the "and" of two propositions both of which are true, leading to the conclusion that the formula evaluates to T. This process is simply the inductive application of the rules:

- S is T if and only if ¬S is F.
- $(S_1 \vee S_2)$ is F if and only if both S_1 and S_2 are F.
- $(S_1 \wedge S_2)$ is T if and only if both S_1 and S_2 are T.

Such a truth evaluation approach can be the basis for developing *control rules*, which are rules that allow the individuation of inconsistent or erroneous data records into a large set of similar records. We denote by P a *record schema*, that is a set of *fields* f_i, with $i = 1...m$, and by p a corresponding *record instance*, that is a set of values v_i, one for each of the above fields.

$$P = \{f_1, \ldots, f_m\} \quad p = \{v_1,\ldots,v_m\} \tag{A.2}$$

Each field f_i, with $i = 1...m$, has its *domain* D_i, which is the set of every possible value for that field. Examples of fields f_i are age or marital status, and corresponding examples of values v_i are 18 or single.

18. A control rule should be applied to a generic record and provide a binary value. Therefore, each rule can be seen as a mathematical function r_k from the Cartesian product of all the domains to the Boolean set {0,1}, as follows (see also Fellegi and Holt, 1976).

$$r_k : \begin{array}{ccc} D_1 \times \ldots \times D_m & \rightarrow & \{0,1\} \\ p & \mapsto & 0,1 \end{array} \qquad \text{(A.3)}$$

The problem of error detection can be approached by formulating a set of rules $R = \{r1, ..., r_t\}$ that are verified by consistent, or correct, records, and are not verified by inconsistent, or erroneous, records. These rules are called compatibility rules, they are such that a generic record p is recognized as a correct record if and only if $r_k(p) = 1$, for all $k = 1, ..., t$. On the other hand, incompatibility rules are verified by erroneous records and not verified by correct records. The detection of erroneous records into a large set of records is a very relevant problem in the field of data E&I.

Compatibility and incompatibility rules can be expressed as disjunction (∨) and/or conjunction (∧) of conditions (also called propositions), hence with the structure of propositional logic formulas. Like to the truth evaluation technique described above, the value of each field of a record under analysis provides a truth assignment for those propositions. The truth/falsehood of the formula constituting the rule provides now the detection of inconsistent or erroneous data records.

However, differently from the case of pure propositional logic, conditions may have an internal structure.

It is necessary to distinguish between two different types of structures for the conditions:

- A condition involving values of a single field is called a logical condition, and corresponds to an atomic proposition of propositional logic. For instance, ($age < 14$) is a logical condition.
- A condition involving mathematical operations between values of fields is called mathematical condition. For instance: (age - $years\ married \geq 14$) is a mathematical condition.

We call *logical rules* the rules expressed only with logical conditions, *mathematical rules* the rules expressed only with mathematical conditions, and *logic-mathematical rules* the rules expressed using both types of conditions. For instance, a logical rule expressing that "*if PM10 number of exceedances of the daily average of 50 µg/m3 isn't less than 0, then PM10 annual average concentration value should be not less than 0*" is:

$$PM10\_SUP\_CENTR\_ARIA \geq 0 \Rightarrow PM10\_MEDIA\_CENTR\_ARIA\_T1 \geq 0$$

This rule can be represented by the following compatibility rule:

$$(PM10\_SUP\_CENTR\_ARIA \geq 0) \vee PM10\_MEDIA\_CENTR\_ARIA\_T1 \geq 0$$

or, equivalently, by the following incompatibility rule:

$$PM10\_SUP\_CENTR\_ARIA \geq 0 \wedge \neg (PM10\_MEDIA\_CENTR\_ARIA\_T1 \geq 0).$$

A formal definition of the structure of the rules allows solving by means of automatic formal methods a number of difficult and computationally demanding problems arising in the different steps of E&I procedures. Examples are:

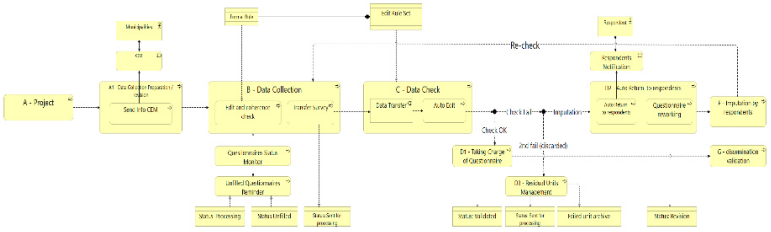
- Problems of error localization (the determination of the erroneous fields of a record).
- Problems of imputation (the determination of the correct values for the erroneous fields of a record. This can be done according to a minimum change principle or by means of a data driven approach).
- Problems of finding contradictions into the set of rules itself (the determination of a (sub)set or rules determining a logical inconsistency).

Note, in particular, that very effective solution approaches are available when encoding rules into linear inequalities. Indeed, a parallelism can be established between logic formulas and linear constraints, and between atomic propositions and 0-1 variables (see Chandru and Hooker, 1991). The above problems are converted into linear or integer linear programming problems and solved by means of efficient optimization solvers. For further details on those techniques see e.g. Bruni and Bianchi, 2012; Bianchi *et al*., 2020; Bianchi *et al*., 2008.

5. New validation process

A sketch of the new validation process is shown in figure 4.

Figure 4 – *Process outline.*



The Process Spans into several Phases

- Survey Project and Preparation (A – A1)
- Data Collection (B)
- Data Check (C - D - E)
- Data transfer to archieves and to production units (G)

A Monitor control the data flow and records questionnaires status and overall completion status

Questionnaires check can fail only once, so unfilled and double failing units are separated and stored elsewhere. After that, validated questionnaires are ready for dissemination.

6. Conclusion

Whereas the ambitious project described above aims to improve data accuracy and consistency through the following progressive design innovations:

1. The design and implementation of a generalized validation process with automated error and analysis reports aimed at increasing the efficiency of the process, increasing the quality of the data collected and reducing the resources employed.
2. The definition of the methodologies and algorithms needed to perform the automatic checks required by the validation process. A main advantage is that this procedure works only at the formal level, so it can be performed without the need of going into the semantic meaning of the validation rules.
3. The design and development of application components and databases, ensuring the integration of the validation process with the acquisition and production environment.
4. The analysis and validation of the implemented tools and the results of the new validation process, through the definition of test cases and continuous experimentation on the 2023 survey, allow the generalised model to be applied to highly differentiated and technically complex situations related to the themes identified by the urban environment survey. For example, for the urban green topic, the test concerns the green management tools used by municipal administrations (qualitative variables); for public transport, on the other hand, the model is applied to the demand and supply of the service (quantitative variables).

The effectiveness in introducing these new generalized solutions that adopt standard methodologies based on technological innovations have the stated goal of applying a new strategy for the pre-validation of data sent by municipalities that by standardizing and automating the recall of incongruents, they want to reduce to a few cases those necessary for in-depth examination by Istat's thematic experts.

References

ADAMO D., COSTANZO L., BUZZI L., LAGANÀ A., GAROZZO S., GRECO V. 2020. *Principali fattori di pressione sull'ambiente nelle città italiane– Anno 2018,* In ADAMO D. and COSTANZO L. (Eds.), Istat, Territori, Letture statistiche.

BIANCHI G., BRUNI R.A. 2012. Formal Procedure for Finding Contradictions into a Set of Rules, *Applied Mathematical Sciences,* Vol. 6, No. 126, pp. 6253-6271.

BIANCHI G., MANZARI A., REALE A. 2008. An overview of editing and imputation methods for the next Italian censuses. In *Proceedings of the Conference of European statisticians,* Geneva.

BIANCHI G., MANZARI A., PEZONE A., REALE A., SAPORITO G. 2005. New procedures for editing and imputation of demographic variables, In *Proceedings of the Conference of European statisticians,* Ottawa.

BOSCH P., BÜCHELE M., GEE D. 1999. Environmental Indicators: Typology and Overview, *European Environment Agency*, pp. 19.

CHANDRU V., HOOKER J.N. 1991. Extended Horn Sets in Propositional Logic, *J. ACM,* Vol. 38, pp. 205-221.

FELLEGI I.P., HOLT I.P.D. 1976. A Systematic Approach to Automatic Edit and Imputation, *Journal of the American Statistical Association,* Vol. 71, pp. 17-35.

TORELLI R. 2011. A generalized system for web surveys, In *Proceedings of Statistics Canada Symposium.*

Domenico ADAMO, Istat, adamo@istat.it
Gianpiero BIANCHI, Istat, gianbia@istat.it
Lucia MONGELLI, Istat, mongelli@istat.it

EVALUATING COMPUTER-ASSISTED QUESTIONNAIRE USABILITY: THE CASE OF PERMANENT CENSUS OF THE POPULATION AND HOUSING[1]

Sabrina Barcherini, Katia Bontempi, Manuela Bussola, Barbara Maria Rosa Lorè, Simona Rosati

Abstract. This paper examines the optimization of a questionnaire, aiming to simplifying the respondent's task and improving data quality. Various methods, including pre-testing techniques and paradata analysis, are utilized to evaluate the survey tool's content, usability, and functionality. Focusing on the Census of the Population and Housing questionnaire, the study analyzes data from 2019 and 2021 to assess the impact of usability optimization on reducing completion difficulties. Findings demonstrate the effectiveness of the usability solutions implemented in alleviating critical issues and addressing respondents' difficulties in completing the questionnaire. An analysis based on regression methods has been used to classify households according to their dependence on external help to complete the questionnaire.

1. Introduction

Usability is a fundamental concept in user-centered design. It is a measure of how easy something is to use and refers to the interaction between product and user. The International Organization for Standardization (ISO) defines usability as the "*extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency, and satisfaction in a specified context of use*" (ISO 9241-11:2018). In this perspective, a product is usable when users can fulfil their task with minimal effort, stress and errors, while feeling satisfied with their interaction with it. This necessitates the design of interfaces that are intuitive, easy to use, and capable of supporting users in their tasks (Nielsen, 1993).

In the context of questionnaire design, usability focuses on creating a smooth and effortless user experience. A well-designed questionnaire should function as a virtual

[1] This article is the result of the collaboration between the authors. In particular: paragraphs 1 and 7 have been written by Barbara M. R. Lorè, paragraph 2 has been written by Manuela Bussola, paragraphs 3 and 4 have been written by Simona Rosati, paragraph 5 has been written by Katia Bontempi, paragraph 6 has been written by Sabrina Barcherini.

assistant, interacting with respondents, performing some operations on their behalf, reminding them of their progress, and providing instructions for the next steps. Taking usability principles into account when designing a questionnaire will result in an easy and stress-free completion experience, leading to low burden, promoting respondent fidelity, and ultimately yielding satisfactory response rates and accurate data. As usability is a key factor in creating a successful product, it is important to place the user at the centre of the product design and development process. When designing a questionnaire, this entails understanding what features of the instrument might make users uncomfortable. A number of methods are available to the researcher to explore the respondents' difficulties with the questionnaire. Some of these are specifically designed to gather information on critical issues, while others serve as secondary sources. The first category encompasses cognitive interviewing, respondent feedback, interviewer debriefing, pilot surveys and experiments, while paradata and help desk tickets analysis fall into the second group (Barcherini et al, 2022).

This paper discusses the results of respondent feedback, interviewer debriefing and helpdesk ticket analysis from the census of the population and housing. In 2018, when the Italian census became permanent, a brand-new electronic questionnaire was developed, for both CAPI (computer-assisted personal interviewing) and CAWI (computer-assisted web interviewing). The questionnaire consists of three parts: the list of the household members, pre-filled with administrative data and requiring verification and potential edits by the respondents; an individual questionnaire for each family member; and a questionnaire for the household as a whole, collecting information on the dwelling. During the initial two years, evidence of respondents' difficulties with the new questionnaire was reported both by field staff and respondents themselves. Consequently, in 2020 layout and functionality changes were implemented to address these issues.

In the following paragraphs, a comparison between the new and original versions of the questionnaire is provided to evaluate the extent of improvement.

2. Critical points in the use of the electronic questionnaire and the need for improvement

In 2018, a debriefing with the Heads of the Municipal Census Offices (HMCOs) was conducted through an electronic questionnaire a few weeks after the conclusion of the census survey. It provided a valuable opportunity to gather feedback from the managers. A specific section of the debriefing questionnaire aimed to collect the managers' opinions on the usability and navigation of the electronic questionnaire. The managers' views were based on their interactions with households that had

received assistance from to Census Office (UCC) and on the experiences reported by fieldwork staff who conducted face-to-face or phone interviews the households using the computer-assisted (CA) questionnaire.

The HMCOs highlighted some recurring issues. The majority of them reported that both the field staff (88.9%) and the households (79.5%) had experienced difficulties, either occasionally or frequently, in using the questionnaire. Only a residual percentage of managers stated that neither the households (20.5%) nor the fieldworkers (15.1%) had found any difficulties. The HMCOs were then asked to clarify the nature of these problems. In addition to issues related to the application itself, such as network slowness and access problems, 38.1% of HMCOs received reports from households of difficulties during the final submission stage. Fieldworkers also encountered similar problems: in this case, the percentage of HMCOs who collected at least one report of this type increased to 41.1%. Furthermore, 13.3% of HMCOs reported feedback from households about difficulties in navigating the electronic questionnaire. However, this percentage decreased to 9.7% when it was the fieldwork staff who reported the problem to their HMCO. These results highlighted the existence of some critical issues in the original version of the electronic questionnaire that required attention. Improvements needed to be made in order to enhance the efficiency and usability of the system and to ensure a better overall experience for respondents during the data collection process.

In the second year of the census (2019), requests for assistance from households received by the dedicated Contact Center were analysed. A total of 2,276 assistance requests out of approximately 43,000, were specifically related to the navigation and usability of the electronic questionnaire (Istat, 2021). The results show that three out of four requests were related to difficulties encountered during the final submission of the questionnaire. This high percentage highlights the importance of focusing on the difficulties encountered by users in completing and submitting the electronic questionnaire. It is therefore essential to allocate resources and efforts towards improving this phase of the census process, to help users submit the questionnaire correctly and timely. Less common but still noteworthy, were difficulties in accessing individual questionnaires, which accounted for 13.2% of the requests for assistance. Similarly, 5.7% of the assistance requests were related to adding family members to the initial list or completing editable fields. Although these figures are lower than the previous ones, they still require attention in order to guide subsequent improvement measures. Finally, 4.4% of the requests for assistance revealed problems related to the partial saving of data. This issue is relevant as inefficient partial saving can lead to data loss and to an overall increase in the burden on respondents.

In conclusion, the analysis of the requests for assistance, which focused on the navigation and usability of the electronic questionnaire, confirmed the HMCO's

opinions. The main issues identified were related to the final submission process, and to a lesser extent, to the access to the individual questionnaires, the use of editable fields, and the partial data saving.

3. Usability improvements implemented in the computer-assisted questionnaire

Based on the issues identified in 2018 and 2019, and taking advantage of the interruption of the census in 2020, a number of improvements have been implemented for the 2021 edition of the computer-assisted questionnaire (CA questionnaire) in order to address the identified problems and eliminate, or at least reduce, their impact.

As mentioned above, it is important to consider the usability of a CA questionnaire within the broader context of websites usability. Following the definition of usability as "the extent to which specified users achieve specific goals with effectiveness, efficiency, and satisfaction in a given environment" (ISO, 2018), it is evident that in the context of a CA survey, users are represented by respondents in CAWI surveys and interviewers in CAPI surveys, while the website is represented by the questionnaire itself (Gabbiadini, Mari, Volpato, 2011).

According to Gabbiadini, Mari, and Volpato, it is possible to apply the five qualitative components that define web usability to the essential attributes that a CA questionnaire should possess to be deemed usable. Adhering to these principles involves creating a CA questionnaire that "is appropriately interpretable, thus reducing the cognitive load on respondents and minimizing errors stemming from the inherent design characteristics of the web artefact" (Gabbiadini, Mari, Volpato, 2011, p. 250). These attributes include: a) Learnability. The ease with which users can successfully complete all required operations; b) Efficiency. The speed at which users can perform various response operations to the items; c) Memorability. The ease with which users can acquire and remember the necessary operations for subsequent completion sessions; d) Errors. The ability to proactively reduce the likelihood of errors in responding or to guide respondents in resolving any potential completion errors (reducing the number of errors, dropout rates, and partial completions); e) Satisfaction. The extent to which the system is perceived as enjoyable and reduces the cognitive load on respondents.

Based on these principles, improvements have been made to several aspects of the census CA questionnaire. Specifically, interventions have been conducted in the following three main areas: 1) graphic restyling; 2) simplified navigation; 3) guided completion.

Graphic restyling. Screen reading has been proven to take longer than paper reading (Nielsen, 2000). Therefore, web pages should be designed to facilitate reading and responding to questions. In general, in adherence to usability principles, text should be concise, appealing, and visually light.

In this perspective, the colour scheme of introductory text and completion instructions has been revised to align with the broader communication standards of the population and housing census (Gabbiadini, Mari, Volpato, 2011). For instance, the colour red has been removed from the text and reserved only for structural components of the screen (e.g., headings) and institutional titles, since red is the distinctive colour of the population and housing census. Additionally, certain keywords have been highlighted in a brighter colour to guide reading and focus the users' attention on the most relevant concepts for navigation and completion support.

Simplified questionnaire navigation. Providing users with a tool that guides and orients them throughout the completion process is crucial in the design of a CA questionnaire. For instance, the presence of a progress indicator allows users to track their progress and know their current position within the web system. To achieve this, the navigation menu has been simplified. Its colour scheme has been revised to better differentiate completed sections (indicated in blue) and sections that are yet to be filled out (indicated in gray). Moreover, efforts have been made to reduce call-to-action items and to make the buttons more intuitive. For example, the labels on the buttons that enables partial saving and advancement to subsequent screens have been made more descriptive. This is important to reassure users that they can progressively save their answers, in respect of the effort made up to that point (Nielsen, 2000; Polillo, 2006).

Guided completion. Reducing the cognitive load on respondents is a priority that a usable web system should ensure. In this regard, interventions have been made in three aspects of the functionality of the census CA questionnaire.

The first aspect addressed the verification and modification of the list of family members. This operation was initially confusing and redundant, requiring multiple actions within the same table, such as "edit," "confirm," and "complete." To streamline the process, pre-filled fields have been made immediately editable by simply placing the mouse cursor in the field, eliminating the need to click the "edit" button. The graphical display of the list has been simplified, and now shows only the data of the family members that the respondent must verify or update if necessary. The "confirm list" button has been made more prominent, placed outside the table, and visually standardized with other call-to-action buttons.

The second improvement focused on optimizing the operations required to access individual questionnaires. The number of steps has been reduced, and clear labels identify the required actions, which are arranged sequentially within the screen. Access to individual profiles has been emphasized with a dedicated button that only

appears displayed once the list of family members has been confirmed. Previously, access to individual profiles was embedded in the list of family members.

The third aspect addressed a significant issue, which was the difficulty in completing the final submission of the questionnaire. A number of measures have been taken to overcome this challenge. The screen has been visually streamlined to highlight the submit button. The screen has been visually streamlined to highlight the submission button. Graphics have been aligned with previous screens, and text has been simplified to eliminate redundancy.

4. The interviewers debriefing and help desk tickets analysis to evaluate the impact of the changes to the questionnaire

With the completion of the first cycle of the census (2018-2021), the municipal fieldwork network was invited to respond to a debriefing questionnaire, administered using CAWI technique. The purpose was to gather opinions and suggestions on the entire census process. Approximately 50% of the 23,524 operators responded to the questionnaire.

Specific questions in the consultation questionnaire were aimed at gathering feedback on any difficulties related to the completion of the CA census questionnaire. A total of 9,852 network operators who conducted field interviews responded to these questions, representing 80.4% of those who completed the consultation questionnaire. The results were overwhelmingly positive. According to the majority of network operators, the CA questionnaire works very well. Over 90% of operators reported no difficulties for every aspect of functionality and usability considered: filling in the family members list (98.3%); access to the individual forms (98.5%); navigation menu (97.3%); questionnaire navigation (95.3%); warning prompts (96.8%); tooltip use (98%); visualization of the sections summary (98.6%); access to the preview of the questionnaire (98.5%); final submission (97.9%).

This result is further supported by the significant decrease in the number of tickets received by the Contact Center in 2021 regarding the functioning of the CA questionnaire. Only 500 out of the 100,000 tickets collected were associated with completion difficulties.

In conclusion, these data demonstrate the effectiveness of the developed CA questionnaire and the progressive reduction of problems encountered during the completion process by the end users: respondents and interviewers.

5. The analysis of respondents' feedback to detect users' difficulties

The census questionnaire includes a final section of standardized questions aimed at collecting feedback on the respondents' completion experience. This feedback encompasses aspects such as the mode of completion, the difficulties encountered, requests for assistance, and more. The analysis of this feedback can help improve the questionnaire design and, consequently, make it more user-friendly.

In order to determine the impact of the usability improvements introduced in 2021 on reducing respondents' completion difficulties, the respondent feedback collected in 2019 and 2021 was analysed[2]. In 2019, CAWI was chosen by 51.4% of respondents, which increased to 53.1% in 2021. The percentage of questionnaires filled in by a member of the household rose from 86.9% in 2019 to 87.7% in 2021. Furthermore, the percentage of households that did not need any help to complete the questionnaire (not from friends or relatives, not from the help desk, not from the municipal census offices, etc.) increased from 78.2% to 79.9%.

When examining households where all members are elderly (65 years or older), the percentage of CAWI usage increased from 44.5% in 2019 to 45.7% in 2021. The percentage of questionnaires filled in by a household member also rose from 57.9% to 60.5%, while the percentage of households not needing any assistance increased from 50% to 52%. Therefore, the analysis of respondents' feedback seems to confirm that the usability improvements have succeeded in simplifying the completion of the questionnaire, especially for elderly users.

In 2021, a new question asking for the reasons of the requests for help has been introduced. Figures show that about one household out of four reported experiencing usability difficulties such as navigation or submission problems. Among elderly households, this percentage increases to 28.2%, showing that this is a subpopulation still needing attention as for the questionnaire design.

6. A classifications of households who need assistance in filling out the questionnaire

In order to better identify households experiencing difficulties, it has been necessary to classify households according to the degree of difficulty they encountered in completing the questionnaire. Based on the respondents' feedback, a simple way to assess the level of difficulty is to calculate the number of different types of assistance required by respondents.

[2] Unweighted raw data have been used.

Some of the indicators calculated for the respondents have been used in a tree regression model to classify households according to their requirement for assistance in completing the questionnaire, whether it be from friends or relatives, the help desk, the municipal census offices, or others. The regression tree allows for finding a relationship between a quantitative variable and a set of independent variables. The dependent variable was the number of different types of help requested, while the independent variables were household socio-demographic indicators: geographical distribution of residence, level of education (represented by the highest educational title within the household), number of household members, presence of elderly members (determined by an indicator that distinguishes households where all members are 65 years or older from those with at least one member below 65), employment status (households with at least one employed person versus those without), citizenship (households with all foreigners versus households with at least one Italian citizen), and ownership or tenancy of the dwelling.

A regression tree is an iterative binary recursive partitioning method that divides data into groups, which are then divided into smaller groups (Breiman, 1984).

The tree regression model has been applied to census data[3] from 2019 and 2021. The results of the classification method are interesting, and a very high goodness-of-fit of the model is observed, as evidenced by a high chi-square value.

According to the model, in 2019 nearly 22% of the respondents needed help to fill out the questionnaire, and the first variable that characterized the analysis was the level of education (Figure 1). The only variable that did not help explain the model was the one related to rental or ownership of the dwelling in which the family lives.

The propensity to require assistance decreases as the level of education increases. For 11 percent of households with no or very low levels of education (nodes 3 and 4) the percentage requesting help is more than 63%. Request for assistance drops below 15 percent for the most educated households, which account for three-quarters of all households (nodes 1 and 2).

The next partition is determined by employment status: the request for assistance is higher when there is no employed member in the household. In node 13, characterized by households with a primary education level and no employed member (9% of all households), the demand for help is 64.2%, compared to 7.3% observed in node 6 represented by graduated households with at least one employed member (25% of all households).

The absence of foreign members in the household also contributes to a further decrease in the need for help. In nodes 16 and 20, representing 63% of households, the absence of foreign members reduces the need for assistance to 7% when there is

[3] Unweighted raw data have been used.

at least one employed and graduated member in the household, and to 11.2% when there is at least one member under 65 with an upper secondary education.

Finally, the geographic breakdown of residence is another variable that explains a high propensity to request assistance in filling out the questionnaire. In node 28, represented by the households of low-educated elderly people residing in Southern Italy or in the Islands, the percentage of requesting assistance is close to 70%.

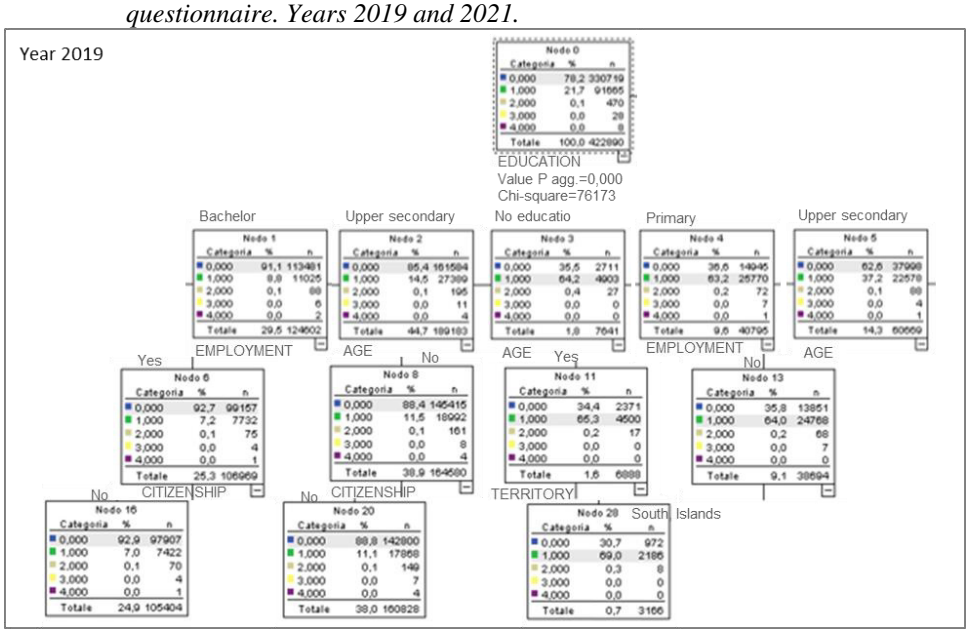
The model applied in 2021 also shows a good fit to the data (Figure 1). The percentage of people demanding assistance is 20%, lower than in 2019. The level of education continues to be the most representative variable, but the geographical distribution is no longer significant. As in 2019, being a homeowner or a renter does not contribute to the model.

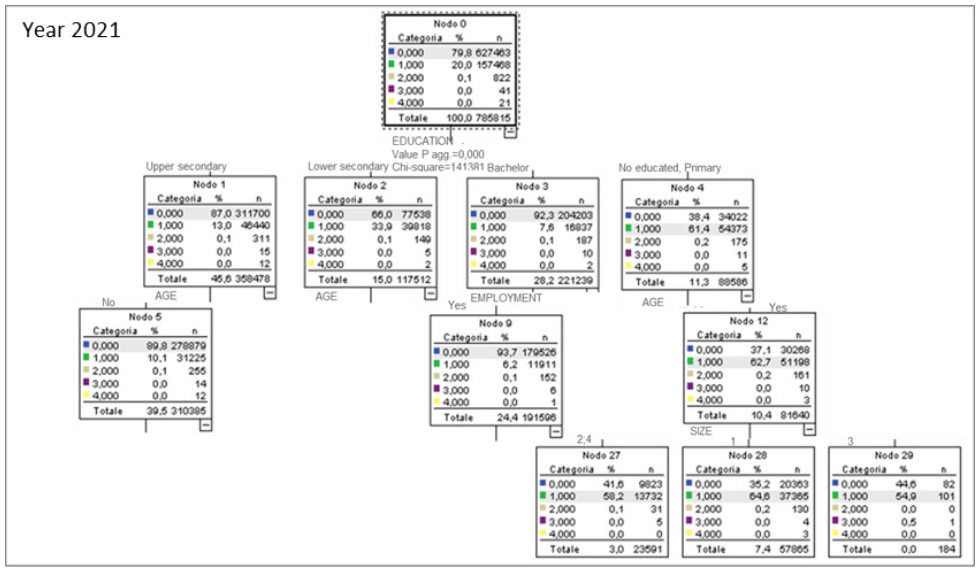
More than 45% of households are represented by node 1, which comprises respondents who have reached a high school level of education (upper secondary education), for whom the need for assistance is only 13%. The need for assistance drops to 7.6% at node 3 (28.2% of households), where there is at least one member who has attained a bachelor's degree. The highest rates of assistance continue to be observed among those with the lowest levels of education: in 2021, the absence of education or primary education (node 4), is represented by 11.3% of households, who need assistance in 61.4% of cases (percentage still lower than in 2019).

In 2021, the next partition is determined by age: the demand for assistance increases with age. Having an upper secondary education, accompanied by the absence of elderly members, reduces the demand for assistance to 10%, as observed at node 5, which includes about 40% of households. Conversely, the request for assistance reaches almost 63% when the lack of education is combined with the absence of members younger than 65 (node 12 represented by 10.4% of households).

The presence of employed members also contributes to a reduction in the demand for help: at node 9, represented by a quarter of the responding households, the percentage of assistance decreases to 6.3% for households with employed graduates. If focusing on the respondents facing the greatest difficulty in filling out the questionnaire, the most vulnerable group are uneducated elderly individuals living alone (7.4%), for whom the need for assistance reaches 65% (node 28).

Figure 1 – *Regression Trees of the households requests for assistance in completing the questionnaire. Years 2019 and 2021.*





Elaborations from unweighted raw census data.

7. Discussion

Completing an electronic questionnaire is a process that entails different skills and abilities. It is essentially a cognitive task, in which respondents find themselves alone facing an object they are unfamiliar with. It is a task that demands considerable cognitive effort, wherein both linguistic proficiency (knowledge and comprehension of the language used) and problem-solving skills come into play.

Hence, it is not surprising that education plays a pivotal role in determining respondents' capability to perform the task independently. Moreover, this phenomenon is not binary in nature. Respondents' autonomy seems to be associated with the amount of education they have received. When the level of education is lacking or at a minimal level, the reliance on external help is four times higher compared to more educated households.

The impact of education on questionnaire completion capability is reinforced by the presence of employed individuals within the family. Although this association may be attributed to the increased employability of educated individuals, resulting in a more favourable employment status of highly educated households, it may also be due, to some extent, to a greater problem-solving ability possessed by those who are employed. In fact, whatever the occupation, it undoubtedly requires a constant effort to analyze situations, identify problems, and seek and apply solutions. The parallels with the demands of an electronic questionnaire are strong.

However, human capital alone is not everything when the problem that needs to be solved is completing an electronic questionnaire. Social support also matters. The analysis reveals that the need for assistance intensifies among the most vulnerable segment of the population, namely uneducated elderly individuals living alone, who constitute the 7.4% of the population. Nonetheless, with appropriate support, these respondents still manage to complete and submit the questionnaire, demonstrating the significance of social capital. The good news is that in 2021 the territorial gap has reduced, probably as a consequence of the digitalisation process of the elderly during the pandemic. Thus, residing in the southern regions or on islands no longer seems to be a factor that worsens the autonomy of this target group.

As questionnaire designers, our focus must be on the needs of respondents, particularly those with limited resources. The comparison between 2019 and 2021 demonstrates that thoughtful actions yield prompt results. The crucial point is to find solutions that will allow the questionnaire to replicate as closely as possible the supportive behaviour provided to the respondents. To achieve this goal, we need to employ targeted methods, specifically designed for this population segment. These could entail a combination of usability testing and cognitive interviewing. Ultimately, this approach will enable us to offer respondents a stress- and frustration-free experience, which should serve as the guiding principle of our work.

References

BARCHERINI S., BONTEMPI K., FAZZI G., LIANI S., LORÈ B., PIETROPAOLI S., ROSATI S. 2022. The respondent as the focus of the questionnaire design. In *Proceedings of the UNECE Expert Meeting on Statistical Data Collection*, Rome.

BREIMAN L., FRIEDMAN J., OLSHEN R., STONE C. 1984. *Classification and regression trees*. CRC press.

GABBIADINI A., MARI S., VOLPATO C. 2011. Internet Come Strumento di Ricerca: Linee Guida per la Creazione di Web Survey, *Psicologia Sociale*, Vol. 2.

ISTAT. 2021. *La conduzione della raccolta dei dati del Censimento permanente della popolazione e delle abitazioni 2019.* Roma: Istituto Nazionale di Statistica.

NIELSEN J. 1993. *Usability engineering*. Burlington: Morgan Kaufmann.

NIELSEN J. 2000. *Design web usability: the practice of simplicity*. Indianapolis: New Riders Publishing.

POLILLO R. 2006. *Plasmare il web: road map per siti di qualità*. Milano: Apogeo.

Sabrina BARCHERINI, Istat, sabrina.barcherini@istat.it
Katia BONTEMPI, Istat, katia.bontempi@istat.it
Manuela BUSSOLA, Istat, bussola@istat.it
Barbara Maria Rosa LORÈ, Istat, lore@istat.it
Simona ROSATI, Istat, srosati@istat.it

WEIGHT OPTIMIZATION FOR COMPOSITE INDICATORS BASED ON VARIABLE IMPORTANCE: AN APPLICATION TO MEASURING WELL-BEING IN EUROPEAN REGIONS

Viet Duong Nguyen, Chiara Gigliarano

Abstract. Composite indicators are widely recognized as effective tools for representing complex assessments in the form of a one-dimensional measure. The proliferation of related theoretical frameworks and methodologies has been accompanied by a growing debate around the determination of optimal weights in developing composite indicators. This paper introduces two weighting procedures aimed at assisting developers in attaining the most plausible solution, which closes the disparity between the importance of input features and their corresponding weights. The first technique involves utilizing variance-based sensitivity analysis and calibrating the weights in accordance with the contribution of each input to the output uncertainty. Alternatively, the second approach employs a combination of cluster analysis and predictive modeling to evaluate the relative capability of individual features in differentiating observations within the multidimensional context, thereby informing a proper weight assignment. To demonstrate the practical application of these weighting procedures, a composite indicator has been developed to assess the level of well-being in large European regions during the ten-year period from 2010 to 2019. Despite differences in the weighting schemes used to calculate the final index values, the empirical results indicate a general consensus regarding the allocation of welfare across the territories.

1. Introduction

Composite indicators are basically models used to measure the performance of objects in complex concepts which are not able to judge based on a single aspect. The role of composite indicators is to provide a proper aggregation that combines the conduct of objects in different dimensions into only one scalar. On the one hand, composite indicators are useful to support decision makers in capturing multidimensional realities and comparing object performance straightforwardly. On the other hand, they might provide incorrect benchmarks and misleading policy messages if they are poorly constructed, induced by inefficient input selection or misinterpreted model configuration (Nardo *et al*., 2005b).

Whereas the selection of inputs is primarily contingent upon the definition of the interested phenomenon, the configuration of weights and aggregation functions largely falls within the purview of modelers. A multidimensional problem entails many possible measurement approaches, leading to a certain degree of subjectivity when imposing judgments on its constituent components. Consequently, weights can be attained from any consideration such as statistical models, participatory methods, or expert opinions. Since weights highly impact the result of composite scores and the ranking of units in benchmarking exercises, it is imperative that the assumptions and implications of the employed weighting scheme are transparent and rigorously tested for robustness (Nardo *et al.*, 2005a).

In this study, we introduce two data-driven weighting methods for constructing composite indicators. The first approach involves an optimization procedure based on variance-based sensitivity measures. The application of variance-based sensitivity analysis to composite indicators has been pioneered in a number of studies that primarily focuses on assessing uncertainty in model output (Grupp and Mogee, 2004; Saisana *et al.*, 2005). Nevertheless, this technique has not gained widespread adoption in weight elicitation due to the necessity of having a predefined set of weights prior to conducting the analysis. Becker *et al*. (2017) has devised a strategy for calibrating weights to ascertain the empirical significance of each input so that it is aligned with the value recommended by expert opinions. Despite the merits, this method tends to generate gaps between the estimated importance of variables and their corresponding weights, diminishing the transparency when interpreting the composite index. Our proposed method closes the gaps by seeking the most effective configuration for the multidimensional context, wherein a set of weights is tuned to achieve no difference between the weights and the normalized sensitivity score of inputs.

The second approach, adopting an alternative perspective, leverages valuable information garnered from unsupervised learning techniques to derive appropriate weights for a composite indicator. Among these techniques, principal component analysis (PCA) and factor analysis are two prominent candidates thanks to their ability of dimensionality reduction. Some noteworthy instances of PCA weighting can be found in the works of Klasen (2000) and Nicoletti *et al.* (2000). However, PCA and factor analysis exhibit discernible limitations, such as inapplicability to low-correlated data, susceptibility to outliers, and the potential to produce negative weights (Nardo *et al.*, 2005b). As a viable substitute, we advocate employing a combination of cluster analysis and predictive modeling to gauge the importance of input variables in distinguishing objects in the multidimensional space. Subsequently, this information serves as the basis for defining the weights based on the rationale that variables of higher significance in classification ought to be more amplified within the composite measure.

For a practical application, we present a composite indicator designed to measure well-being in 198 large European regions over the decade from 2010 to 2019. The equal weighting, the PCA weighting, and the two proposed methods were applied to provide a comprehensive view of welfare allocation across the European territories at both regional and national levels.

2. Measuring Variable Importance

2.1. Sensitivity Analysis Approach

Sensitivity analysis involves the study of how uncertainty in the output of a model can be apportioned to different sources of uncertainty in the model input (Saltelli, 2002). This approach allows for identifying which variables have the greatest influence on the composite indicator score, thereby providing insights into the model's validity and reliability. In this paper, we focus on the technique of measuring sensibility using conditional variances. The foundation of this approach was introduced by Sobol' (2001), who devised a measure that bears his name. The Sobol' method is based on an essential assumption that all the input features are mutually independent, which might be unrealistic in practice. Mara *et al.* (2015) developed a methodology to overcome this issue by proposing a strategy of estimating importance indices that account for the dependency of input factors. Let Y denote a composite indicator obtained from a square integrable function $f(X)$ where the input $X = (X_1, X_2, \dots, X_n)$ is a random vector, the authors provided an improvement of the original Sobol' indices:

$$
\begin{aligned}
S_i^{full} &= \frac{\mathrm{Var}(\mathrm{E}_{X_{\sim i}}(Y|X_i))}{\mathrm{Var}(Y)}, \\
ST_i^{full} &= \frac{\mathrm{E}(\mathrm{Var}_{X_i}(Y|(X_{\sim i}|X_i)))}{\mathrm{Var}(Y)}, \\
S_i^{ind} &= \frac{\mathrm{Var}(\mathrm{E}_{X_{\sim i}}(Y|(X_i|X_{\sim i})))}{\mathrm{Var}(Y)}, \\
ST_i^{ind} &= \frac{\mathrm{E}(\mathrm{Var}_{X_i}(Y|X_{\sim i}))}{\mathrm{Var}(Y)}.
\end{aligned}
\tag{1}
$$

The measures S_i^{full} and ST_i^{full}, called the full Sobol' indices, reflect the main and the total contribution of X_i to the output variance, taking into account its dependency with the other inputs. On the other hand, S_i^{ind} and ST_i^{ind}, called the independent Sobol' indices, respectively measure the main and total contributions

of X_i that does not account for its mutual dependence on all the other inputs. With respect to the variable X_i, denote w_i as the weight and I_i as the variable importance measured by one of the four indices. The importance measures for all the variables are normalized by $\tilde{I}_i = I_i / \sum_{k=1}^{n} I_k$ to make them comparable to the value of weights. Denote a loss function

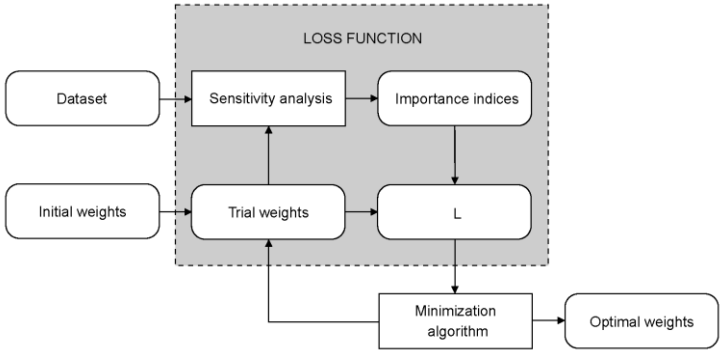
$$L = d^2(w, \tilde{I}) = \sum_{i=1}^{n} \left(w_i - \tilde{I}_i\right)^2, \tag{2}$$

which is the squared Euclidean distance between two vectors $w = (w_1, \dots, w_n)$ and $\tilde{I} = \left(\tilde{I}_1, \dots, \tilde{I}_n\right)$. The optimal set of weights is defined by

$$w^* = \underset{w_1,\dots,w_n}{\operatorname{argmin}} L \quad \text{s.t.} \quad w_i \in (0,1),\ \sum_{i=1}^{n} w_i = 1. \tag{3}$$

At $L_{\min}$, the distance between the two vectors is minimal and hence we attain the set w^* as close as possible to $\tilde{I}$. In case $L_{\min} = 0$ that is equivalent to $w^* \equiv \tilde{I}$, the weights obtained are exactly proportional to the measures of importance. Figure 1 gives an illustration of the optimization procedure. Please note that the rounded boxes indicate the inputs/outputs while the rectangle boxes demonstrate the functions. At the beginning, a sample of X and an initial set of weights are fed into the loss function to estimate the distance L. The trial weights are then updated using a minimization algorithm based on the estimated values of L until the loss function achieves its minimum, which indicates the best course of action.

Figure 1 - *Weight optimization procedure based on sensitivity analysis.*



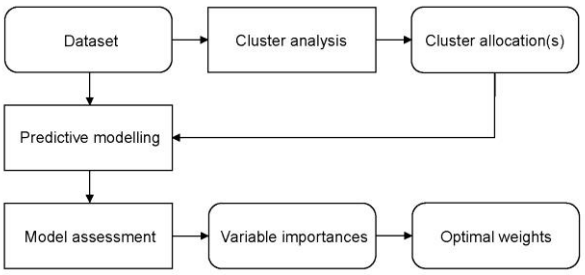
2.2. Unsupervised Learning Approach

The techniques such as PCA or factor analysis encompass extracting a small number of principal components (factors) that capture sufficiently large variation in the data and evaluating the relative correlation of each variable with the identified components. In this context, the importance of a variable reflects how well it might help in explaining the major variance in the multidimensional data space. A standard procedure when employing PCA/factor analysis to build composite indicators is using the loadings on the first component as the weights assigned to the variables (Klasen, 2000; Greyling and Tregenna, 2017).

Despite of popularity, the use of PCA and factor analysis is limited in the case of low correlation among the feature dimensions. To handle this problem, we suggest using cluster analysis as an alternative solution in measuring variable importance. The notion of variable importance in clustering refers to the extent to which individual features contribute to the formation of distinct clusters. Given a cluster structure, it is possible to fit a model to predict the cluster labels from the input features, and variable importance can be calculated as the mean decrease in prediction accuracy when a particular feature is permuted (Breiman, 2001). This value provides information regarding the variable's efficacy in discriminating data points based on common patterns. The weights derived from the cluster-based technique thus reflect the ability of input features to distinguish observations in a multidimensional context.

The illustration of the weighting procedure based on cluster analysis is depicted in Figure 2, with the inputs/outputs presented in the rounded boxes and the used functions given in the rectangle boxes. The key idea behind this approach is to train a mapping function employing the variables in the dataset to predict the cluster membership, which is derived from the clustering process. By evaluating the mean decrease in prediction accuracy by each variable, the importance measures can be defined and then normalized to obtain the weights for the composite indicator.

Figure 2 - *Weight optimization procedure based on cluster analysis.*



3. Data

Table 1 provides a concise overview of the dimensions utilized to create the composite indicator for measuring well-being in European regions. According to the OECD well-being framework (OECD, 2020), the dimensions of income and jobs pertain to material conditions that shape people's economic sustainability. Whereas the aspects of health, education, environment, and safety refer to the fundamental measures of life quality. Civic engagement reflects the degree of public trust in government and of voters' participation in the political process. The society dimension measures the severity of social exclusion, which is represented by the share of young people not in employment, education, or training (NEET). Lastly, digital accessibility is an essential topic to be considered, as it promotes social inclusion, access to information resources, economic opportunities, and personal empowerment for individuals with disabilities.

Table 1 – *Dimension structure for measuring regional well-being.*

Dimension	Measuring indicator	
Income	*income:*	household disposable income per capita (real USD PPP)
Jobs	*emp_rate:*	employment rate (%)
	unemp_rate:	unemployment rate (%)
Health	*life_exp:*	life expectancy at birth (years)
	mort_rate:	age adjusted mortality rate (per 1000 population)
Education	*sec_edu*:	share of population from 25-64 years old with at least secondary education (%)
Environment	*air_pol:*	air pollution in $PM_{2.5}$ (average level in $\mu g/m^3$ experienced by the population)
Safety	*hom_rate:*	intentional homicide rate (per 100 000 population)
Civic engagement	*vote:*	voter turnout to general elections (%)
Society	*soc_exc:*	share of population from 18-24 years old not in employment and not in any education and training (%)
Digital accessibility	*bb_acc:*	share of households with broadband access (%)

Owing to the fact that the eleven indicators are gauged in different units, it is required to make them comparable by converting all the features into the same scale $[0,1]$ using the formulas:

$$\bar{x} = \frac{x - \min(x)}{\max(x) - \min(x)}, \text{ or} \tag{4}$$

$$\bar{x} = \frac{max(x) - x}{\max(x) - \min(x)} \tag{5}$$

The min-max normalization (4) is applied to the transformation of features that are positively correlated with well-being, including income, employment rate, life expectancy, education attainment, voter turnout, and broadband access. On the contrary, the max-min normalization (5) is implemented for the remaining features, which are considered to exert a negative effect on well-being. If a dimension is constituted by a pair of indicators, such as jobs and health, the dimension score is computed by averaging the normalized values of both components, then applying the min-max normalization again to ensure conformity to the identical scale.

The data for the well-being features is collected from the OECD Regional Statistics database (OECD, 2023), which encompasses yearly time-series for the variables of demography, economy, labor market, social and innovation themes in the OECD member countries. The original dataset comprises 2250 observations, measuring the eleven well-being factors for 225 large (TL2) regions in 29 European countries over a ten-year period from 2010 to 2019. We performed few data replacement for the Netherlands and Greece, where the features of homicide rate and broadband access are not available at the regional level in some time ranges, by utilizing the figures at their national level. We decided to remove from the original dataset the regions that are subjected to the following two conditions: the number of missing values is greater than 15% of the total data cells; and the information in at least one variable is completely unavailable. With these criteria, 27 regions from 11 countries were disposed of, leading to the absence of five countries including Bulgaria, Ireland, Iceland, Lithuania, and Malta.

To thoroughly address the problem of missing data, we applied the *k*-nearest neighbors (*k*-NN) imputation technique with the Euclidean distance metric. An encoding technique was also employed to take into account the regional and temporal effects. The factor variables of time and location were first converted using dummy encoding, subsequently fed into the *k*-NN algorithm along with all the other indicators to define the proximate data points of observations with missing information. Following this, the unknown cells were filled by a distance-weighted average of the values from their closest neighbors. Finally, the complete features in the imputed dataset were used to compute the nine well-being dimensions.

4. Well-being Scores by Composite Indicators

To establish the composite indicator for regional well-being, we used the weighted arithmetic mean function

$$C^{(r,t)} = \sum_{i=1}^{n} w_i X_i^{(r,t)}, \tag{6}$$

where C is the composite score, X_i is the normalized score in dimension i, w_i is the weight assigned to dimension i, and the term (r, t) denotes the region and time allocated to the observation. Table 2 presents the sets of weights for the well-being composite indicator derived from four different weighting methods. The first column simply contains the equal weights, which can be used as a baseline for comparison. The second column shows the weights calculated from the standard weighting approach using PCA. The values in the third column are the result from the sensitivity-based weighting procedure using the full main Sobol' index (S_i^{full}). The last column displays the clustering-based weights utilizing permutation importance estimated by a random forest classifier for the three-cluster allocation.

Table 2 – *Weights by different weighting methods.*

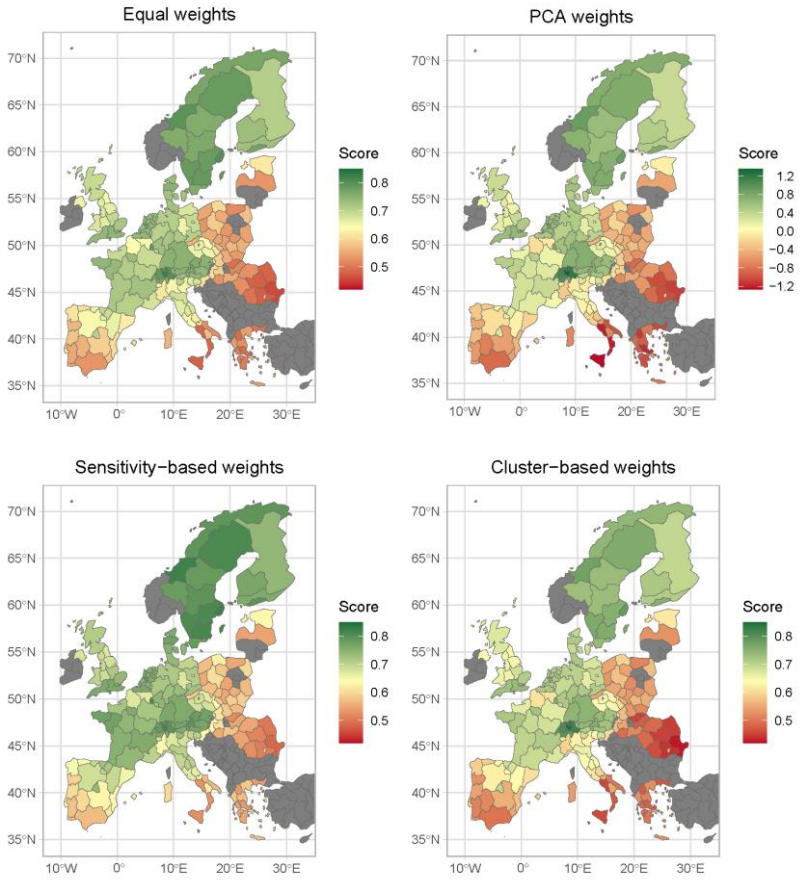
	Equal	PCA[1]	Sensitivity-based[2]	Cluster-based[3]
Income	0.111	0.128	0.067	0.128
Jobs	0.111	0.161	0.091	0.118
Health	0.111	0.079	0.133	0.176
Education	0.111	0.103	0.138	0.165
Environment	0.111	0.084	0.130	0.101
Safety	0.111	0.047	0.155	0.048
Civic engagement	0.111	0.081	0.117	0.067
Society	0.111	0.161	0.090	0.093
Digital accessibility	0.111	0.156	0.081	0.104
Sum	1.000	1.000	1.000	1.000

There are differences in the level of importance each method assigns to the variables, as manifested through the corresponding weights. This inconsistency is the result of the distinct mechanism employed by each method. However, a common pattern in the allocation of well-being composite scores is found (see Figure 1) due to the compensation between the features in the aggregation process. Northern Europe is the area that displays the highest average scores in well-being, followed by Western European regions and the British Isles. In Eastern Europe, most territories show a welfare level below the regional average, and places in the southeast notably exhibit the lowest scores throughout the entire observed regions. In Southern Europe, a north-south gradient in well-being is visible, where the northern part produces mostly above-moderate scores while the southern part

[1] Weights are the normalized factor loadings of the first principal component (37% of total variance).
[2] Weights are the normalized S_i^{full} estimated by penalized cubic splines via generalized cross-validation. Note that $S_i^{full} = ST_i^{full}$ in purely addictive models.
[3] Weights are the normalized permutation importances calculated by a random forest ensemble that grows 500 trees and randomly samples three features as candidates at each split. The allocation in the three-cluster solution by k-means clustering is chosen as the response based on the elbow method.

predominantly records subpar status. At the TL2 regional details, the regions in Switzerland, Sweden, and Norway consistently show a superior level of well-being while the most deprived ones are in Eastern Romania, Southern Italy, and Greece.

Figure 1 – *Average well-being composite scores for European regions in the period 2010-2019.*



Note: the grey areas denote territories with missing in- formation or territories not in European regions.

With respect to national well-being, Figure 2 shows the alterations in rankings for 24 European countries in the observed period. A country's score is computed by taking the population-weighted average of the scores from all its constituent regions. Switzerland, Sweden, Norway, and Luxembourg consistently rank at the highest positions in the charts. These nations tend to display robust performance on the composite indicator regardless of the weighting schemes used, proving their

status as the most-welfare countries in Europe. On the other hand, the countries of Romania, Latvia, Hungary, and Greek frequently appear at the lower end of the rankings. This evidence implies that these nations face challenges in various well-being dimensions, which diminishes their overall prosperity compared to the other European members.

Figure 2 – *Annual rankings by national average scores in well-being.*

5. Conclusion

This paper introduces two innovative weighting procedures for composite indicators, focusing on the quantification of variable importance across diverse conceptual frameworks. The first approach, sensitivity-based weighting, enables researchers to derive a solution wherein the magnitude of weights corresponds to the contribution of input features to the variance in composite scores. This method is designed to work compatibly with any single-valued function, independent of its complexity and parameter configuration, making it applicable to all models that return a scalar output. The second approach, cluster-based weighting, addresses multidimensional challenges by investigating the underlying cluster structure in data and estimating the impact of each dimension on predicting cluster membership. The optimal number of clusters can be determined through clustering validation indices or by examining the association between various clustering schemes and the performance of prediction models. The cluster-based weights obtained from this process can serve as a measure of each variable's ability to differentiate observations in the multidimensional space representing the phenomena of interest.

We have developed a composite indicator for measuring the well-being of inhabitants in large European regions. An imputed dataset containing information on nine well-being dimensions for 198 regions during the 2010-2019 period was used for computing the composite scores. Four weighting methods were employed, including equal weighting, PCA weighting, and the two novel techniques proposed in our study. Despite the variations in weighting schemes and score outcomes, all four methods collectively reveal similar patterns of welfare allocation throughout the regions under evaluation. Regions in Northern Europe exhibit the highest average well-being scores, followed by their Western European counterparts and those within the British Isles. Southern Europe holds the third position with a clear north-south differentiation while Eastern European regions experiences the lowest levels of well-being. At the level of national well-being, Switzerland, Sweden, Norway, and Luxembourg maintain their prominence by consistently securing top positions in the annual ranking charts, whereas Romania, Latvia, Hungary, and Greek frequently appear as the most deprived nations in these figures.

References

BECKER W., SAISANA M., PARUOLO P., VANDECASTEELE I. 2017. Weights and Importance in Composite Indicators: Closing the Gap, *Ecological Indicators*, Vol. 80, pp. 12-22.

BREIMAN L. 2001. Random Forests, *Machine Learning*, Vol. 45, pp. 5-32.
GREYLING T., TREGENNA F. 2017. Construction and analysis of a composite quality of life index for a region of South Africa, *Social Indicators Research*, Vol. 131, pp. 887-930.
GRUPP H., MOGEE M. E. 2004. Indicators for National Science and Technology Policy: How Robust are Composite Indicators? *Research Policy*, Vol. 33, No. 9, pp. 1373-1384.
KLASEN S. 2000. Measuring Poverty and Deprivation in South Africa, *Review of Income and Wealth*, Vol. 46, No. 1, pp. 33-58.
MARA T. A., TARANTOLA S., ANNONI P. 2015. Variance-Based Sensitivity Indices for Models with Dependent Inputs, *Environmental Modelling & Software*, Vol. 72, pp. 173-183.
NARDO M., SAISANA M., SALTELLI A., TARANTOLA S. 2005a. Tools for Composite Indicators Building. *Technical Report EUR 21682 EN*, JRC-EC, Ispra, Italy.
NARDO M., SAISANA M., SALTELLI A., TARANTOLA S., HOFFMAN A., GIOVANNINI E. 2005b. Handbook on Constructing Composite Indicators: Methodology and User Guide. *OECD Statistics Working Paper No. 2005/03*, OECD Publishing.
NICOLETTI G., SCARPETTA S., BOYLAUD O. 2000. Summary Indicators of Product Market Reg- Ulation with an Extension to Employment Protection Legislation. *ECO Working Papers No. 226*, OECD.
OECD. 2020. *How's Life? 2020: Measuring Well-being*. OECD Publishing.
OECD. 2023. OCED Regional Statistics (database). https://doi.org/10.1787/region-data-en, accessed on 01 march 2023.
SAISANA M., SALTELLI A., TARANTOLA S. 2005. Uncertainty And Sensitivity Analysis Techniques as Tools For The Quality Assessment Of Composite Indicators, *Journal of the Royal Statistical Society: Series A*, Vol. 168, No. 2, pp. 307-323.
SALTELLI A. 2002. Sensitivity Analysis for Importance Assessment, *Risk Analysis*, Vol. 22, No. 3, pp.579-590.
SOBOL' I. M. 2001. Global Sensitivity Indices for Nonlinear Mathematical Models and their Monte Carlo Estimates, *Mathematics and Computers in Simulation*, Vol. 55, No. 1-3, pp. 271-280.

Viet Duong NGUYEN, University of Insubria, vdnguyen@uninsubria.it
Chiara GIGLIARANO, University Carlo Cattaneo, cgigliarano@liuc.it

SUBJECTIVE WELL-BEING AND HETEROGENEITY IN CULTURAL CONSUMPTION IN AGING POPULATIONS

Maria Carella, Roberta Misuraca

Abstract. This study investigates the relationship between cultural consumption patterns and well-being in the older population. Using data from the 2018 Italian Multipurpose Survey on Households *"Aspects of daily life*", we employ Latent Class Analysis to identify distinct profiles of cultural consumers based on their attendance and engagement in various cultural and art activities. We then investigate the effects of these cultural consumption profiles on life satisfaction and other domains of well-being, including leisure and friend satisfaction. Our findings reveal a positive association between cultural engagement and subjective well-being across different domains. Specifically, individuals who allocate more time to diverse cultural experiences show higher levels of well-being. We also observe gender differences in well-being outcomes. These results highlight the importance of promoting cultural participation to enhance older adults' well-being and inform the development of targeted welfare policies.

1. Introduction

In recent decades, there has been an increasing awareness that economic well-being does not fully capture the multidimensional nature of individual well-being. As a result, social science researchers have turned their attention to subjective well-being (SWB) and related concepts like life satisfaction and happiness. Along this line, a crescent body of literature has consistently shown that both active and passive engagement in arts and cultural activities can have a positive impact on individual well-being (Fancourt and Finn, 2020; Bertacchini et al., 2023). Furthermore, the growing number and share of older people in the population has generated increased interest in studying experiences or activities that can improve life satisfaction during the later stages of adulthood. A crucial aspect of successful aging is the social integration of the older population: by participating in cultural activities, older adults have the chance to engage with their peers, expand their social networks, and combat feelings of loneliness and marginalization. Research has shown that cultural activities contribute to a sense of community and belonging, promoting social interaction and

meaningful connections among older adults. However, despite the growing relevance of this issue, it is still insufficiently explored. Moreover, the existing studies tend to consider only a single cultural activity, neglecting to consider the effect of a simultaneous combination of variety and intensity (frequency) of cultural participation on subjective well-being, in general, and specifically on populations differentiated by age groups. Our study aims to complement and update the knowledge on this topic, by exploring the relationship between cultural consumption patterns and various domains of subjective well-being, including life satisfaction, leisure, and friendship satisfaction, among the aging population. Specifically, using data from the 2018 Italian Multipurpose Survey on Households "*Aspects of daily life*" by the Italian National Institute of Statistics (ISTAT) on a sample of approximately 16,550 individuals, we seek to identify different profiles of cultural consumers based on their patterns of attendance and engagement in various cultural activities (*Culturally Inactive, Culturally Omnivore, Heritage Lovers* and *Culturally Voracious*) (Katz-Gerro, 2004; Sullivan and Katz-Gerro, 2007). To achieve this purpose, we employed Latent Class Analysis (LCA). After controlling for individual socio-demographic characteristics and territorial variables at the regional level, we investigate how heterogeneity in cultural participation, due to different simultaneous combinations of variety and frequency of engagement at the individual level, impacts life satisfaction and friend and leisure satisfaction of the older population. In doing so, our findings show a positive relationship between cultural consumption and subjective well-being and demonstrate that more time dealing with diverse cultural experiences is associated with higher levels of life satisfaction and its other components, particularly among older adults. In line with many aging studies (Mendes de Leon 2005; Fancourt and Steptoe, 2018) this result confirms that cultural participation can play a crucial role in active aging and highlights the importance of promoting activities in this field among older adults to enhance their quality of life. Finally, we analyze by gender to verify whether there are significant differences in culture consumption patterns when comparing men and women and to evaluate how the impact of cultural activities on the distinct components of SWB relates to gender.

2. Theoretical Background

Currently, the literature on well-being is vast and implies diverse disciplines or research fields nevertheless a consensus on its comprehensive definition remains elusive (Brown *et al.*, 2015; Galloway, 2006). The starting point in the definition of well-being is the differentiation between objective and subjective components. While the first considers “outer” qualities, such as living in a good environment or an acceptable state of physical and mental health, the second captures the personal

satisfaction of life, in terms of subjective assessment of individual life circumstances. Within the field of happiness economics, subjective well-being is often associated with life satisfaction (Christoph and Noll, 2003). There is a lack of consensus on the metrics of subjective well-being measurement scales and about the correct and best way to analyze them. In conceptualizing subjective well-being, Diener and Suh (1997) refer to three interrelated components: satisfaction with life, pleasant and unpleasant affect, further taking into account the central role played by the presence of negative experiences. In our work, we embrace a multidimensional perspective of subjective well-being (Diener and Suh, 1997) that includes individual evaluation of overall life satisfaction and relevant subdomains (leisure, friendship relation). An extensive body of research has documented the effects of various aspects of life on well-being, such as household income (Diener *et al.*, 2013), employment conditions (Bonanomi and Rosina, 2022), and health status (Galloway, 2006). Recently, non-monetary factors have gained attention, highlighting the multidimensional nature of life satisfaction. Research on the relationship between leisure activities, including arts and culture, and subjective well-being is limited. Empirical evidence suggests a positive association between engagement in arts and cultural activities and well-being, including cognitive enhancement, increased happiness, and the development of pro-social attitudes (Brown *et al.*, 2015; Fancourt and Finn, 2020). The existing studies on this topic, however, focused on samples of the overall population (Brown *et al.*, 2015; Graham and Pozuelo, 2017), neglecting to look into broader phenomena and deepen the interactions between greater engagement in arts and cultural activities and specific domains of life satisfaction. Moreover, they tend to examine the impact of a single cultural activity or treat them as additive factors (Bertacchini et al., 2023) disregarding the potential combined effect arising from the simultaneous interaction of variety and frequency of engagement in different cultural activities. This calls for a comprehensive understanding of how subjective well-being is connected to cultural consumption according to this interaction. To address this research gap, we draw the concept of cultural consumption profiles from the sociological literature (Katz-Gerro, 2004). The pivotal study on this subject is the work of Bourdieu (1984) which introduced the notion of highbrow and lowbrow cultural goods to grasp the different inclinations of individuals toward cultural participation. However, the concept of cultural consumption is broader and encompasses a wide range of activities and experiences. In this direction, Peterson (1992) proposed the concept of (cultural) “omnivores” and “univores”. “Cultural omnivores” are individuals who exhibit a broad and diverse cultural “appetite”, engaging in various cultural activities encompassing both highbrow and lowbrow cultural forms. In contrast, “cultural univores” tend to have a more limited range of cultural preferences and participation. Sullivan and Katz-Gerro (2007) further expanded these categories by introducing "voracious omnivores" who actively engage in multiple cultural activities. Understanding how these patterns

influence subjective well-being and cultural engagement is crucial. In a recent work, using data from an Italian survey, Bertacchini et al. (2023) investigate the association between the heterogeneity in cultural profiles, and overall life satisfaction, as well as specific domains such as health, leisure, and friendship relations. Findings from their work reveal a positive relationship between cultural participation and some components of SWB that, moreover, tend to increase about the diversity and intensity of cultural practices expressed in the profiles of cultural consumers. In our work, we attempt to complement these findings by investigating the effect of engagement in cultural activities and arts on older adults' subjective well-being. On this matter, a stream of literature concerning older people has documented that engagement in cultural activities provides opportunities for them to interact with others, fostering social connections and a sense of belonging. By serving as a driver for reducing social isolation, cultural activities offer opportunities to engage with their peers, expand their social networks, and combat feelings of loneliness and marginalization (Findlay, 2003). Overall, findings revealed different patterns of cultural consumption according to aging (Goulding, 2018) focusing on the role of sociodemographic factors, such as gender, ability/disability condition, and education level, as important drivers of cultural engagement for older adults (Keaney and Oskala, 2007). Additionally, maintaining cultural traditions seems to have a further positive effect on life satisfaction (Bernardo and Carvalho, 2020), due to the strengthening of interpersonal ties and better social coexistence. However, the existing literature in this area is limited and mainly focused on the consequences of social inclusion/exclusion on well-being.

3. Research Hypothesis

The existing research has found evidence of a positive association between active and passive cultural participation and the individual well-being of older people suggesting that engaging in cultural activities, such as attending concerts, visiting museums, or participating in arts, contributes positively also to cognitive function favoring an active and socially integrated lifestyle (Fancourt and Finn, 2020). By relying on this literature, we hypothesize a positive relationship between cultural consumption and subjective well-being for older populations in Italy (H1). In addition, the omnivore/univore theory (Katz-Gerro, 2004) has emphasized the role played by the simultaneous combination of variety and intensity of engagement in cultural Activities in determining different levels of subjective well-being. From this perspective, we, therefore, hypothesize that the joint influence of engaging in a diverse range of cultural activities (variety) and the frequency of participation in those activities (intensity) impacts SWB (life satisfaction) and other relevant domains, such as leisure, and friendship satisfaction (H2). Finally, some studies recognize the

presence of gendered cultural taste and assume that the heterogeneity in cultural consumption behavior could increase or diminish when controlling for the various groups' sociodemographic characteristics (Katz-Gerro, 2004). Following this approach, we, therefore, expect to find gender differences. More specifically, comparing men and women we suppose that the impact of their participation in cultural and art activities could be associated to different extents with the distinct components of the SWB (H 3).

4. Data and methodology

The data used in this study were derived from the 2018 Italian Multipurpose Survey on Households "*Aspects of daily life*," conducted by the Italian National Institute of Statistics (ISTAT) which contains a sample of approximately 42,000 individuals, Once selected participants aged over 55, we obtain a sample of 16,515 individuals. Given our research aims, we examine socio-demographic characteristics such as gender, age, work conditions, marital status, education level, physical limitations, and some items related to health satisfaction, and economic satisfaction. From the ISTAT data source, we derive information on regional cultural supply. Our main variable of interest is life satisfaction and other relevant subdomains, such as leisure and friendship satisfaction. Participants were asked to indicate how frequently they participated in outdoor cultural and leisure activities over the past twelve months, such as sports events, dancing venues, music concerts, classical music concerts, cinemas, theatres, museums, and monuments. Response options included "*never*," "*1-3 times*," "*4-6 times*," "*7-12 times*," and "*more than 12 times*" within the last twelve months. To grasp the effects of the simultaneous combination of variety and intensity of engagement in cultural activities on life satisfaction and other relevant subdomains, four cultural consumption profiles have been derived through Latent Class Analysis (LCA).

According to Bertacchini *et al.* (2023), the LCA approach is the following:

$$P(Y = y) = \sum_{c=0}^{C} \gamma_c \prod_{j=1}^{J} \prod_{r=1}^{R} \rho_{j,rj|c}^{I(yi=rj)}$$

Where $P(Y = y)$ is the probability of observing a vector of responses, conditional to $I(y_j = r_j)$ if the response to variable $j = rj$, 0 otherwise; γ_c is the probability of belonging to latent class *c*, while $\rho_{j,rj|c)}^{I(y_i=r_j)}$ is the probability of observing a specific response r_j for each individual *i.* The key parameters are γ, representing the latent class membership probabilities, and ρ, representing the item-response probabilities conditional on γ. The underlying idea of LCA is to endogenously form classes consisting of individuals with homogeneous responses. By employing LCA, we can

group individuals with similar preference structures in cultural consumption based on the diversity and intensity of their attendance. We will start by examining different models with varying numbers of classes, ranging from one class to six classes. The optimal number of classes will be determined based on the Akaike information criterion (AIC) and (Bayesian information criterion) BIC criteria[1]. Once the optimal number of classes is identified, we will assign each individual to the class they are most likely to belong to. Each specific cultural profile (class) will then be named based on the characteristics of cultural consumption (Chan and Goldthorpe, 2007) Subsequently, in order to test our main research hypothesis, we will include the cultural consumption profiles obtained through LCA as explanatory variables in our regression models. Given the scope of this study, a probit regression model using a binary choice approach is deemed more appropriate for addressing the cultural question. The main specification is the following:

$$SWB_{ird} = \alpha + \beta_1 CCP'_{ird} + \beta_2 X'_{ird} + \beta_3 Z'_r + u_{ird}$$

Where *SWB* is the subjective well-being of individual *i*, in region *r* for each domain *d* that assumes a value equal to 1 in case individual *i* in region *r* is satisfied, and 0 otherwise. Specifically, "Life Satisfaction" is assessed using an 11-point Likert scale, ranging from 0 (not satisfied at all) to 10 (completely satisfied). A dummy variable is created with a value of 1 for responses in the top four categories (7-10) and 0 for all other scores. Subdomains Friend Satisfaction and Leisure Satisfaction (time over the past twelve months) is measured using a 4-point Likert scale, with responses ranging from 1 (very happy) to 4 (completely unhappy). In this case, dummy variables are created, taking a value of 1 if the individual is either "very happy" or "quite happy," and 0 otherwise. Our main variable is CCP' which represents the four cultural consumption profiles for individual *i* in region *r* for each domain *d.* The cultural consumption profiles obtained through *LCA* implementation can be categorized as follows: *Culturally Inactive*: this category accounts for over 68% of individuals. These respondents have very high conditional probabilities (between 96-99%) of never participating in any of the cultural activities considered.

Culturally Omnivorous: they account for 11% of the sample. Individuals in this group demonstrate moderate probabilities (between 9-43%) of engaging in all the activities analyzed, with a frequency of 1-6 times; *Heritage Lovers*: this category accounts for about 17% of the sample. Individuals in this group exhibit a high preference (between 67-94% probability) for visiting heritage sites such as museums

[1] In latent class analysis (LCA), the Akaike information criterion (AIC) (Akaike, 1974) and the Bayesian information criterion (BIC) (Stone, 1979) are used for model selection. The AIC assesses in-sample fit and selects the model with the minimum value, while the BIC balances model fit and complexity. Lower AIC or BIC values indicate better fit in LCA.

and monuments. They engage in these activities with a moderate frequency of 1-6 times; *Culturally Voracious:* they represent a relatively small but distinct group, comprising approximately 3% of the sample. Individuals in this group participate in all the cultural activities with a high frequency of more than 7 times. The vector X' consists of individual-level observable characteristics. Specifically, it includes variables such as gender, marital status, and the presence of children in the household. Educational attainment is captured through dummy variables representing different levels of education, including low, upper-secondary, and tertiary levels (with low education serving as the reference group, encompassing up to the lower-secondary level). Labor status is represented by dummy variables, which partly capture differences in household income and the availability of leisure time. The ownership status of the individual's residence and a subjective assessment of economic conditions are also considered to understand the individual's propensity to spend on leisure activities. Additionally, subjective measures of health satisfaction and objective indicators of the presence of physical illness are included to account for the individual's health status. Z' consists of a vector of regional-specific characteristics. We include the per capita number of cinemas, concert halls, classical music concerts, theatres, dance floors, monuments, museums, and sporting clubs to capture the geographic variation in the local cultural supply at the regional level. In addition, to account for unobserved characteristics specific to the geographical area, dummies for the macro-region of residence are included (North-East, Centre, South, and Islands, with North-West as the reference group). Finally, u_{ird} is the error term. Table 1 reports the descriptive statistics of the key variables.

Table 1 - *Summary statistics-*

Variables	Obs	Mean	Std. Dev.	Min	Max
Life Satisfaction	16,394	.6541418	.4756618	0	1
Friend Satisfaction	16,425	.7802131	.4141148	0	1
Leisure Satisfaction	16,406	.6712178	.469785	0	1
Culturally Inactive	16,515	.6805934	.4662609	0	1
Culturally Omnivore	16,515	.1092946	.3120179	0	1
Heritage Lovers	16,515	.1772934	.3819283	0	1
Culturally Voracious	16,515	.0328186	.1781671	0	1

5. Results

Table 2 reports the results of probit regressions for all the variables of interest.

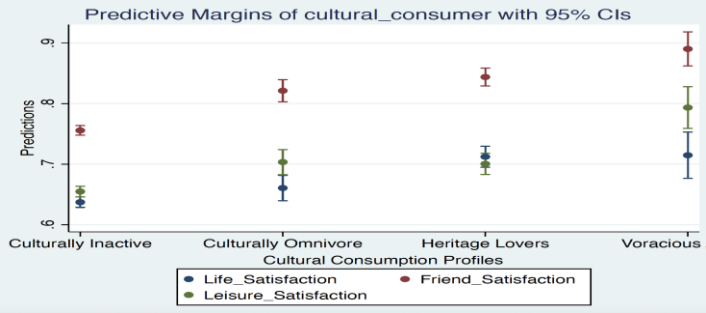
Table 2. *Probit estimations of the determinants of Life and Domain Satisfaction (Friend satisfaction and Leisure satisfaction) for the older population.*

	(1)	(2)	(3)
DV	Life Satisfaction	Friend Satisfaction	Leisure Satisfaction
Culturally Omnivore	0.0272**	0.0685***	0.0538***
	(0.0135)	(0.0106)	(0.0128)
Heritage Lovers	0.0863***	0.0916***	0.0505***
	(0.0116)	(0.00914)	(0.0114)
Culturally Voracious	0.0891***	0.137***	0.151***
	(0.0229)	(0.0147)	(0.0193)
Individual controls	YES	YES	YES
Territorial controls	YES	YES	YES
Regional cultural supply controls	YES	YES	YES
Observations	16,011	16,074	16,058

Note: The baseline category is Culturally Inactive. The sample includes individuals aged 55+. The table provides the results of the probit regression analysis examining the relationship between cultural consumption profiles and different domains of subjective well-being, specifically life satisfaction, friend satisfaction, and leisure satisfaction. Marginal effects displayed. Standard errors clustered at the individual level.

The findings from our study align with previous literature (Brown *et al.*, 2015) and confirm a positive correlation between satisfaction with life and the variety and breadth of cultural activities individuals engage in (column 1). Our results indicate that compared to those who are *Culturally Inactive*, significant improvements in life satisfaction can be achieved by engaging in various cultural activities. (H1). Furthermore, the results suggest that allocating more time to diverse cultural experiences (*Culturally Voracious*) is associated with higher levels of life satisfaction (H2). This finding confirms previous research based on preventing satiation effects (Galak *et al.*, 2011), indicating that increased cultural engagement contributes positively to subjective well-being. Furthermore, our study reveals that a specific category of cultural consumers, *Heritage Lovers*, displays a relatively high probability of being satisfied with life. This evidence is in line with recent studies carried out at the European level that show the positive contribution to engagement in heritage (Ateca-Amestoy *et al.*, 2021). Moreover, we find that variety and intensity in cultural consumption significantly impact friend and leisure satisfaction for older adults (columns 2-3). This underscores the importance of cultivating diverse cultural experiences to enhance social networks and combat feelings of loneliness and marginalization (Findlay, 2003). The analysis controls for individual-level factors, territorial factors, and regional cultural supply factors. The inclusion of these controls helps to isolate the specific effects of cultural consumption profiles on subjective well-being, accounting for other potential influences. The findings are further supported by the information depicted in Figure 1.

Figure 1. *Predictive Margins with 95% CI on the probability of being satisfied with life, friends, and leisure for cultural consumer profiles.*



As shown in Fig. 1, the probability of being satisfied with life consistently increases from *Culturally Inactive* to *Culturally Voracious* profiles. There is a notable increase in probability between *Culturally Inactive* and *Cultural Omnivores*, suggesting that even occasional engagement in cultural activities is positively linked to SWB domains. Furthermore, the predictive margins show that as individuals increase their consumption of cultural goods in terms of variety and intensity (to *Culturally Voracious*) their SWB increases. This implies that older adults experience higher levels of subjective well-being when they engage in a wider range of cultural activities and intensify their involvement. Subsequently, we explore to what extent cultural consumption patterns are associated with life satisfaction and different domains of well-being for the older population by gender (Table 2).

Table 2 suggests that the effects of cultural consumption profiles on subjective well-being may differ between genders. While *Culturally Omnivore* has a positive impact on both males' and females' subjective well-being even in the presence of slight differences across genders, the most significant positive effect is observed for the *Culturally Voracious* group, who engage frequently and diversely in cultural activities across all domains. The table indicates that this effect is particularly pronounced for women regarding friend and leisure satisfaction (columns 4 and 6). These results are in line with the previous literature that shows that women tend to report slightly higher subjective well-being than men (Meisenberg and Woodley, 2015). Women often excel in building and maintaining friendships, as well as investing time and effort into cultivating hobbies and leisure activities. They tend to establish and strengthen social connections more easily over time (Scanlon, 2000). Additionally, women are found to prioritize their health and engage in activities that promote well-being. Furthermore, we find a more positive impact of *Heritage Lovers* on leisure and friend satisfaction for women. A prominent role in this perspective is played by the presence of gendered cultural tastes (Katz-Gerro, 2004) which leads to a differentiated impact on SWB.

Overall, these findings underscore the potential benefits of higher levels of cultural consumption, in terms of variety and frequency, especially for older women.

Table 2. *Probit estimation of satisfaction with life and different domain satisfaction by cultural consumption profiles, heterogeneity across gender.*

	(1)	(2)	(3)	(4)	(5)	(6)
	Life Satisfaction		Friend Satisfaction		Leisure Satisfaction	
DV	M	F	M	F	M	F
Culturally Omnivore	0.0358**	0.0265	0.0733***	0.0612***	0.0587***	0.0526***
	(0.0180)	(0.0191)	(0.0136)	(0.0157)	(0.0166)	(0.0184)
Heritage Lovers	0.0863***	0.0844***	0.0767***	0.104***	0.0391**	0.0582***
	(0.0156)	(0.0164)	(0.0122)	(0.0131)	(0.0155)	(0.0161)
Culturally Voracious	0.0952***	0.0842***	0.127***	0.152***	0.139***	0.169***
	(0.0317)	(0.0313)	(0.0191)	(0.0208)	(0.0258)	(0.0268)
Individual controls	YES	YES	YES	YES	YES	YES
Territorial controls	YES	YES	YES	YES	YES	YES
Regional Cultural Supply covariates	YES	YES	YES	YES	YES	YES
Observations	7,273	8,738	7,297	8,777	7,297	8,761

Note: the baseline category is Culturally Inactive. The sample includes individuals aged 55+. The table provides the results for different groups: Males (M) and Females (F) for each of the three dependent variables. Marginaleffects displayed. Standard errors clustered at the individual level.

6. Conclusions

This work aimed to investigate how heterogeneity in cultural participation, due to different simultaneous combinations of variety and frequency of engagement at the individual level, impacts life satisfaction and other relevant domains (friend and leisure satisfaction) related to the subjective well-being (SWB) of older adults. In line with previous literature (Brown *et al.*, 2015), we find that increasing engagement in arts and cultural activities positively influences these aspects of SWB (H1). Furthermore, we derive distinct profiles of cultural consumers (*Culturally Inactive, Culturally Omnivore, Heritage Lovers,* and *Culturally Voracious)* and observe that individuals who allocate more time to diverse cultural experiences, as seen in the *Culturally Voracious* group, tend to have higher levels of subjective well-being. This supports our hypothesis that increasing the variety and intensity of cultural engagement contributes positively to subjective well-being (H2) and also emphasizes the crucial role played by social integration for successful aging. Engagement in cultural activities provides opportunities for older adults to interact with others, foster social connections, and combat social isolation. By participating in cultural activities, older adults can expand their social networks, develop a sense

of belonging, and experience a greater sense of well-being. Additionally, we uncover slight gender differences in subjective well-being and other relevant domains, suggesting that cultural and art participation may affect well-being differently for men and women (H3). This finding highlights the importance of considering gender-specific factors when designing policies and interventions aimed at promoting cultural engagement and enhancing well-being among older adults. In conclusion, our study contributes to the growing body of research on subjective well-being and cultural engagement among older adults.

Acknowledgments

We acknowledge funding from Next Generation EU, in the context of the National Recovery and Resilience Plan, Investment PE8 – Project Age-It: "Ageing Well in an Ageing Society" [DM 1557 11.10.2022]. The views and opinions expressed are only those of the authors and do not necessarily reflect those of the European Union or the European Commission. Neither the European Union nor the European Commission can be held responsible for them.

References

ATECA-AMESTOY V., VILLARROYA A., WIESAND, A. J. 2021. Heritage Engagement and Subjective Well-Being in the European Union. *Sustainability,* Vol. 13, No. 17, 9623.

BERTACCHINI E., VENTURINI A., MISURACA R., ZOTTI R. 2023. Exploring the relationship between subjective well-being and diversity and intensity in cultural consumption. *Working Paper* n.19/2023, Department of Economics and Statistics Cognetti de Martiis.

BONANOMI A., ROSINA A. 2022. Employment status and well-being: A longitudinal study on young Italian people. *Social Indicators Research*, Vol. 161, No. 2-3, pp. 581-598.

BOURDIEU P. 1984. *Distinction: A social critique of the judgment of taste*. Cambridge, Massachusetts: Harvard University Press.

BROWN J. L., MACDONALD R., MITCHELL R. 2015. Are people who participate in cultural activities more satisfied with life? *Social Indicators Research*, Vol. 122, No. 1, pp. 135-146.

CHAN T. W., GOLDTHORPE J. H. 2007. Social stratification and cultural consumption: The visual arts in England. *Poetics*, Vol. 35, No. 2-3, pp. 168-190.

CHRISTOPH B., NOLL H. H. 2003. Subjective well-being in the European Union during the 90s. In *European Welfare Production: Institutional Configuration and Distributional Outcome*, Dordrecht: Springer Netherlands, pp. 197-222.

DIENER E., SUH E. 1997. Measuring quality of life: Economic, social, and subjective indicators. *Social Indicators Research*, Vol. 40, pp. 189-216.

DIENER E., TAY L., OISHI, S. 2013. Rising income and the subjective well-being of nations. *Journal of personality and social psychology*, Vol. 104, No. 2, pp. 267-276.

FANCOURT D., FINN S. 2020. *What is the evidence on the role of the arts in improving health and well-being? A scoping review*. World Health Organization.

FANCOURT D., STEPTOE A. 2018. Cultural engagement predicts changes in cognitive function in older adults over a 10-year period: findings from the English Longitudinal Study of Ageing. *Scientific Reports*, Vol. 8, No. 1, 10226.

FINDLAY R. A. 2003. Interventions to reduce social isolation amongst older people: where is the evidence?. *Ageing & Society*, Vol. 23, No. 5, pp. 647-658.

GALAK J., KRUGER J., LOEWENSTEIN G. 2011. Is variety the spice of life? It all depends on the rate of consumption. *Judgment and Decision Making,* forthcoming.

GALLOWAY. S. 2006. Cultural participation and individual quality of life: A review of research findings. *Applied Research in Quality of Life*, Vol. 1, pp. 323-342.

GOULDING, A. 2018. The role of cultural engagement in older people's lives. *Cultural Sociology*, Vol. 12, No. 4, pp. 518-539.

GRAHAM C., POZUELO J. R. 2017. Happiness, stress, and age: How the U curve varies across people and places. *Journal of Population Economics*,Vol. 30, No. 1, pp. 225-264.

KATZ-GERRO T. 2004. Cultural consumption research: review of methodology, theory, and consequence. *International Review of Sociology*, Vol. 14, No. 1, pp. 11-29.

KEANEY E., OSKALA A. 2007. The golden age of the arts? Taking part survey findings on older people and the arts. *Cultural trends*, Vol. 16, No. 4, pp. 323-355.

MEISENBERG G., WOODLEY M. A. 2015. Gender differences in subjective well-being and their relationships with gender equality. *Journal of happiness studies*, Vol. 16, pp. 1539-1555.

PETERSON R. 1992. Understanding audience segmentation: From elite and mass to omnivore and univore. *Poetics*, Vol. 21, pp. 243-258.

SULLIVAN O., KATZ-GERRO T. 2007. The omnivore thesis revisited: Voracious cultural consumers. *European Sociological Review*, Vol. 23, No. 2, pp. 123-137.

Maria CARELLA, Department of Political Science, University of Bari Aldo Moro, maria.carella1@uniba.it

Roberta MISURACA, Department of Political Science, University of Bari Aldo Moro, roberta.misuraca@uniba.it

MEASURING THE SPATIAL CONCENTRATION OF THE MAIN FOREIGN COMMUNITIES RESIDING IN ITALY USING A NEW APPROACH

Massimo Mucciardi, Giovanni Pirrotta, Mary Ellen Toffle

Abstract. In the field of population studies the spatial concentration of population reflects many issues. Surely one of the most important is the residential segregation of the foreign populations which is closely linked to the settlement pattern of the population. Starting from previous research on spatial concentration proposed by Mucciardi and Benassi (2023) and subsequently by Benassi *et al.* (2023a), the present paper develops new insights into spatial concentration applied to the main foreign communities in Italy. The aims of the contribution are as follows: (i) to detect the level of spatial concentration of the main foreign communities residing in Italy in 2003, 2011 and 2021; (ii) to evaluate the role of country of citizenship in shaping such levels and dynamics. To achieve aims (i) and (ii) we applied a spatial version of the Gini index, called the Spatial Gini Index (SGI) and the relative curve called Spatial Lorenz Curve (SLC). The results show that the level of the spatial concentration for each foreign group remains almost stable over time (with rare exceptions). In contrast, the level of the spatial concentration for the main foreign communities is lower than the Italian population.

1. Introduction

As stated in a recent paper (Benassi *et al.*, 2023b), spatial concentration is an important aspect when analysing various distributions and patterns pertaining to immigrant communities. According to the important study conducted by Massey and Denton (1988), one of the most indicative elements of residential segregation is that of the actual concentration of the immigrant population. Furthermore, to a recent study conducted by Mucciardi and Benassi (2023), the concentration of the population in terms of space continues to be a strong indicator of urbanization. The phenomenon of immigration in Italy has demonstrated a rapid growth cycle in the last years. For example, there was a very small population of foreigners residing in Italy and most of them had illegal or irregular status (Strozza, 2004). After regularization procedures were put into force, the numbers of foreigners residing in

Italy increased rapidly to 1.89 million individuals in the first part of 2003 (Blangiardo, 2005). But migration intensified between 2003 and 2021: foreign residents increased from 4.32 million in 2011 to 5.03 million in 2021 with a net immigration of about 3.14 million people between 2003 and 2021 (ISTAT, 2023). In addition, there was a change in the countries of origin. East-west groups coming from the Middle East started replacing the south-north immigration pattern of low-income migrants from Asia and Africa (Benassi *et al.*, 2023b). In Italy foreigners are concentrated in certain areas of the country, especially in the large metropolitan cities. However, not all groups of foreigners tend to concentrate or settle in the same way: it is evident that the territorial distribution of the different nationalities is connected not only with the duration of stay (communities of recent immigration vs. communities of previous immigration) and with the stability of the communities on the territories but also with the different migration models. So, the foreign groups follow different settlement patterns and show different levels of spatial concentration. In the field of population studies, the spatial concentration of population reflects many other issues. Surely one of the most important is the residential segregation of the foreign populations which is closely linked to the settlement pattern of the population. Starting from previous research on spatial concentration proposed by Mucciardi and Benassi (2023) and subsequently by Benassi *et al.* (2023a), the present paper developed a new insight of spatial concentration applied to the main foreign communities in Italy. The aims of the contribution are the following: (i) to detect the level of spatial concentration of the main foreign communities residing in Italy in 2003, 2011 and 2021; (ii) to evaluate the role of country of citizenship in shaping such levels and dynamics. To achieve aims (i) and (ii) we applied a spatial version of the Gini index, called the Spatial Gini Index (SGI) proposed by Mucciardi and Benassi (2023) and subsequently by Benassi *et al.* (2023a). SGI is derived from the Lorenz curve (Lorenz, 1908), and it is based on comparing how the contribution in terms of "connectivity" and "variability" varies as the geographical distance between spatial units increases. The paper is organized as follows. In the next section we show a short review of the new procedure while in section 3 we present the results obtained. Some final remarks conclude the paper.

2. A new approach to measure spatial concentration: a short review

Certainly, in the study of concentration the best known and most widespread index is the Gini index (G). As already pointed out in Benassi *et al*. (2023b), despite the large dissemination of the index G, it has also been criticized. When applied to phenomena involving the study of variables in a territorial context, essentially this

index does not take into consideration the geographical nature of the phenomenon, being fundamentally an "aspatial" index (Reardon and O'Sullivan, 2004). To overcome this limitation some alternative measures and approaches have been continually proposed by scholars (Arbia and Piras 2009; Dawkins 2004; Rey and Smith 2013; Panzera and Postiglione 2020, Türk and Östh, 2023).

Here we briefly describe the method proposed by Mucciardi and Benassi (2023) and subsequently by Benassi *et al.* (2023a).

This approach is based on comparing how the contribution in terms of "*connectivity*" ($J_{(k)}$) and "*variability*" ($V_{(k)}$) varies as the geographical distance k between spatial units increases: if the variable observed is not dependent on space, the variations between the connectivity and variability components should not differ much from each other. The idea is to consider buffer or threshold distances (k) capable of progressively creating partitions of the territory (or territorial subsets). These partitions identify neighbouring and non-neighbouring units such that each partition is disjoint from the others and the sum of all the elements of all the partitions coincides with the number of all the possible pairs between the n spatial units (Mucciardi and Benassi, 2023; Benassi *et al.*, 2023). To satisfy these conditions we use the MaxMin distance method (Mucciardi, 2008). The properties of the territorial partitions of the MaxMin method makes the procedure compatible with the graphical representation of the Gini index according to the Lorenz curve approach (Lorenz, 1908). We called this index the "Spatial Gini Index" (SGI) and the relative curve "Spatial Lorenz Curve" (SLC).

Basically we can have three scenarios (Figure 1):

A) when the cumulative variability contribution in terms of variability $V_{(k)}$ is be larger than the cumulative connectivity contribution in terms of connectivity $J_{(k)}$ as the distance k increases, then $0 \leq SGI < 0.5$ (case of negative spatial autocorrelation, Figure 1A)[1].

B) when the cumulative variability contribution $V_{(k)}$ is equal to the cumulative connectivity contribution $J_{(k)}$ as the distance k increases, then SGI is perfectly equal to 0.5 (case of spatial no autocorrelation, Figure 1B)[2];

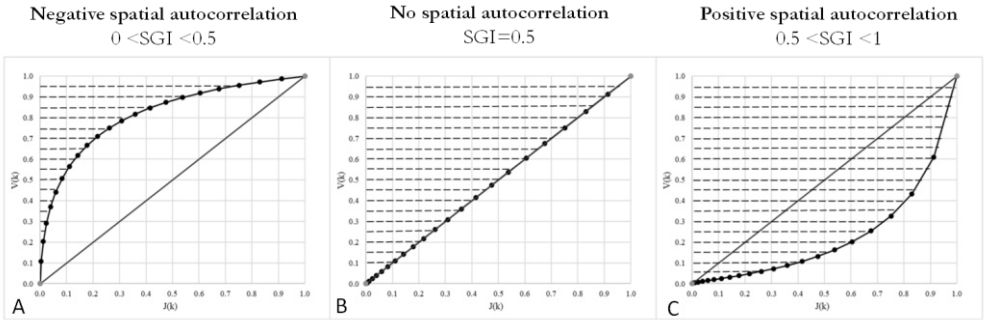
C) when the cumulative variability contribution in terms of variability $V_{(k)}$ is smaller than the cumulative connectivity contribution in terms of connectivity $J_{(k)}$ as the distance k increases, then $0.5 < SGI \leq 1$ (case of positive spatial autocorrelation, Figure 1C).

[1] For more details on the calculation of the SGI and on the construction of the SLC curve, see Mucciardi and Benassi (2023) and Benassi *et al*. (2023).

[2] We can think this case when in the Lorenz curve there is exactly equidistribution of a variable (for example income).

It is important to highlight the link between *spatial concentration* and *spatial autocorrelation*. Previous studies (Mucciardi and Benassi, 2023; Benassi *et al.*, 2023a) show that when a variable (e.g. human population) is uniformly distributed in the entire territory, the SGI tends to 0.5 and the SLC tends to the configuration of Figure 1B (case of no spatial autocorrelation). If the variable is not distributed uniformly throughout the entire territory, two sub-cases may arise: i) in the case of non-uniform distribution (or in any case with high variability) in the contiguous areas, SGI tends towards values lower than 0.5 and the SLC tends to the configuration of Figure 1A (case of negative spatial autocorrelation); ii) in the case of uniform distribution (or in any case with low variability) in the contiguous areas, SGI tends towards values greater than 0.5 and the SLC tends to the configuration of Figure 1C (case of positive spatial autocorrelation).

Figure 1 – *Three scenarios for SGI and SLC (dashed line): 1) case of negative spatial correlation – 0<SGI <0.5 (A); 2) case of no spatial correlation – SGI=0.5 (B); 3) case of positive spatial correlation 0.5<SGI <1 (C). The diagonal (solid line) represents the theoretical condition of spatial equality.*



Source: Benassi et al., 2023a

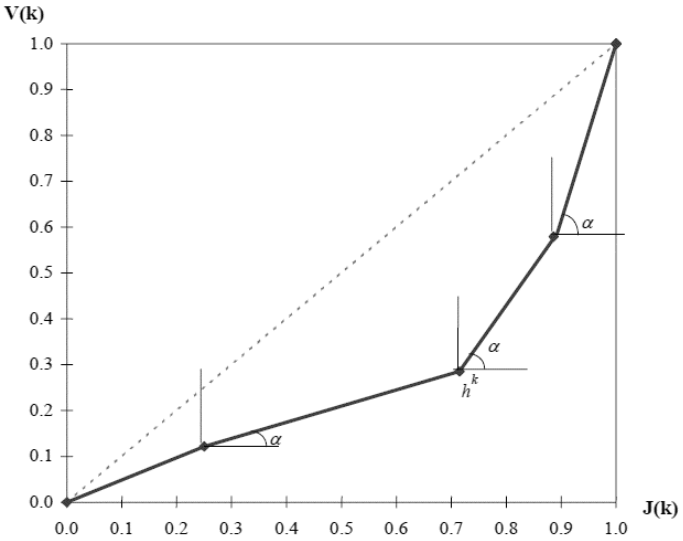
Another interesting feature of this new approach is the possibility of measuring the level of spatial autocorrelation at each distance (k) using the arctan function (AF). In fact, from a geometric point of view, these three forms of spatial autocorrelation may be assessed, as the distance k varies, by considering the tangent of the angle formed by the straight line with the *x*-axis (Mucciardi and Benassi, 2023; Benassi *et al.*, 2023a). In formula:

$$\tan^{k}(\alpha) = \frac{V_{(k)}}{J_{(k)}} \quad k = 1 \dots t. \tag{1}$$

where t coincides with the maximum territorial distance (beyond this distance the MaxMin distance method stops).

We can have 3 cases (Figure 2): *a*) $\tan^{k}(\alpha) < 1$ ($\alpha < 45°$) indicating positive spatial autocorrelation; *b*) $\tan^{k}(\alpha) = 1$ ($\alpha = 45°$) indicating no spatial autocorrelation; *c*) $\tan^{k}(\alpha) > 1$ ($\alpha > 45°$) indicating negative spatial autocorrelation. To convert the tangent function into terms of angle (degrees) we use the AF function[3].

Figure 2 – Values assumed by the angle α with varying values of distance (k) (dashed line represents case of no spatial autocorrelation - $tg^{k}(\alpha) = 1$ ($\alpha = 45°$).



Source: Mucciardi and Benassi (2023)

Furthermore, two indicators are derived from SGI for comparing the value of the spatial concentration of a variable x with respect to the value of the spatial concentration of a second variable y_{ref} usually set as a reference. We called these indicators "Delta SGI" (ΔSGI_x) and "SGI Rate" ($RSGI_x$).

We can write the following formula:

$$\Delta SGI_x = SGI_x - SGI_{ref} \text{ and } RSGI_x = \frac{SGI_x}{SGI_{ref}} \quad (2)$$

with $-1 \leq \Delta SGI_x < 1$ and $0 \leq RSGI_x < \infty$ ($SGI_{ref} \neq 0$)

[3] We recall that the AF is the inverse of the tangent function and it returns the angle whose tangent is a given number.

Considering that SGI is a dimensionless index, the ΔSGI_x can be considered as an absolute variation of the spatial concentration of a variable x with respect to the spatial concentration of the variable taken as a reference. Instead $RSGI_x$ can be considered as a relative variation of the spatial concentration of a variable x with respect to the spatial concentration of the variable taken as a reference. Clearly, the more the spatial concentration of the variable x tends to the spatial concentration of the variable set as a reference, the values of ΔSGI_x and $RSGI_x$ tend to 0 and 1 respectively. Therefore, in the following paragraph we will use this property to verify the similarity/dissimilarity of a foreign community compared to the Italian one in terms of spatial concentration.

3. Measuring the spatial concentration of the main foreign communities: principal results

3.1 Data and MaxMin algorithm output

Before analyzing the results obtained, it is necessary to examine the data used. Data is provided by the Italian National Institute of Statistics (ISTAT, 2023) and it refers to the population usually residing in the Italian municipalities broken down by country of citizenship. Territorial boundaries of 7903 municipalities have been reconstructed so they remain stable over time assuring a correct comparison. Regarding the results of the MaxMin algorithm applied to Italian municipalities, we obtain a total of 163 (k) buffer distances ranging from the minimum value of 16.3 km ($k = 1$) to the maximum value of 1223.9 km ($k \equiv t = 163$)[4].

3.2 Principal results

We calculate the SGI for the first 10 largest foreign communities in 2021 by years: 2003, 2011 and 2021 (Table 1). According to formula (2), two indicators are derived from SGI to compare the value of the spatial concentration of the single foreign community (SGI_{FOR}) with the spatial concentration of the Italian communities (SGI_{ITA}). So, for the 10 foreign communities considered we have:

$$\Delta SGI_i = SGI_{ITA} - SGI_{FOR_i} \text{ and } RSGI_i = \frac{SGI_{FOR_i}}{SGI_{ITA}} \qquad \text{with } i = 1 \dots 10 \tag{3}$$

[4] All elaborations are based on a new ad hoc library developed and implemented in Python. The library can be downloaded at the following link: https://github.com/gpirrotta/spatial-gini-index (Benassi *et al.*, 2023a).

Table 1 summarizes the three spatial concentration indicators for the 10 foreign communities. As we can see (Table 1 and Figure 3) the level of the spatial concentration for each foreign collectivity (SGI_03, SGI_11 and SGI_21) remains almost stable over time with rare exceptions (Bangladesh and Pakistan). In contrast the level of the spatial concentration for the main foreign communities is lower than the Italian population. The value of SGI for Italians varies from 0.485 to 0.483 in the three years considered, substantially showing (as expected) no spatial concentration and stability over time (population tending towards territorial homogeneousness). Instead, Pakistanis, Egyptians and Chinese communities show lower levels of SGI and consequently negative spatial autocorrelation probably due to concentrations of these populations in the North of the Italian territory (Table 1 and Figure 3).

Table 1 – *SGI, ΔSGI_i and $RSGI_i$ for Romania (ROM), Morocco (MOR), Albania (ALB), China (CHI), Ukraine (UKR), India (IND), Philippines (PHI), Bangladesh (BAN), Egypt (EGY), Pakistan (PAK) citizenship – Years 2003 (03), 2011(11) and 2021 (21). Italy (ITA) is reported for reference.*

Citizen	SGI_ 03	SGI_ 11	SGI_ 21	ΔSGI_ 03	ΔSGI_ 11	ΔSGI_ 21	RSGI_ 03	RSGI_ 11	RSGI_ 21
ROM	0.465	0.466	0.465	0.019	0.019	0.018	0.960	0.960	0.963
ALB	0.444	0.437	0.441	0.041	0.049	0.042	0.915	0.900	0.913
MOR	0.454	0.458	0.457	0.031	0.027	0.026	0.937	0.944	0.947
CHI	0.415	0.420	0.413	0.070	0.065	0.070	0.855	0.866	0.854
UKR	0.479	0.467	0.464	0.006	0.018	0.019	0.987	0.962	0.961
IND	0.451	0.451	0.466	0.034	0.035	0.017	0.930	0.928	0.965
PHI	0.434	0.437	0.440	0.051	0.048	0.043	0.895	0.900	0.910
BAN	0.486	0.469	0.467	-0.002	0.016	0.016	1.003	0.967	0.968
EGY	0.417	0.417	0.417	0.068	0.069	0.066	0.863	0.858	0.864
PAK	0.412	0.412	0.434	0.073	0.073	0.049	0.849	0.849	0.898
ITA	*0.485*	*0.485*	*0.483*	-	-	-	-	-	-

Source: Our elaboration on ISTAT data

Moreover, if these foreign communities are compared with the Italian population, we obtain the highest ΔSGI_i and lowest $RSGI_i$ values in all the years considered in this analysis (Table 1 and Figure 4). The other foreign communities seem to have values of the three indices tending towards the values of the Italian population, especially the Ukrainian and Romanian communities (see the values of ΔSGI_i and $RSGI_i$ tending to 0 and 1 respectively). Now let's examine the SLC and the AF. We recall that SGI is a global index while through the SLC and AF it is possible to inspect the spatial autocorrelation values for each distance k.

Figure 3 – *Comparison of the value of SGI by citizenship (Italians are plotted for reference).*

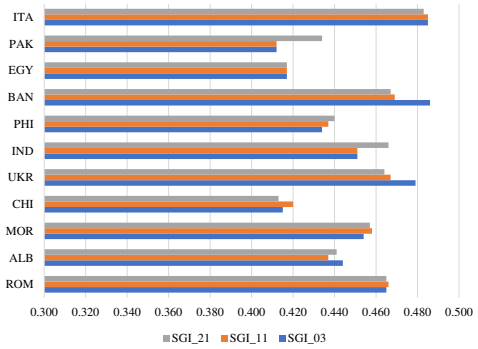
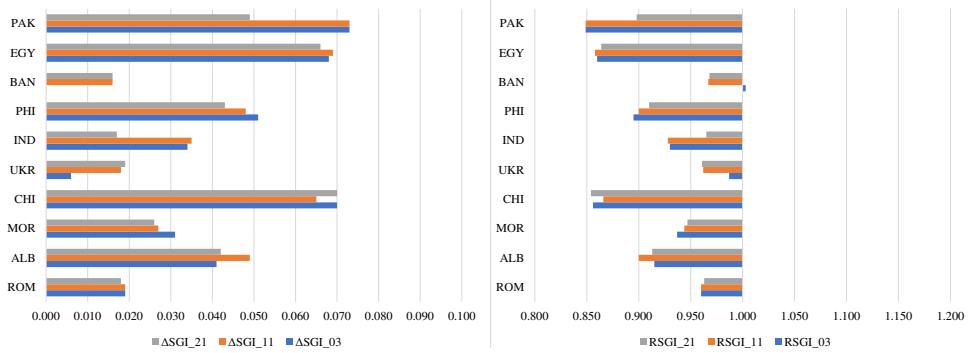


Figure 4 – *Comparison of the value of ΔSGI_i and $RSGI_i$ by citizenship (Italians are the reference).*



If we consider SLC (Figure 5) and the value of the AF (Figure 6) the differences appear greater[5]. For instance, comparing the SLC and the AF of Romanian and Egyptian (year 2021) we notice completely different trends that confirm, albeit in a different way, the settlement models known in the literature (Benassi *et al.*, 2020; Ferrara *et al.*, 2010): a dispersed model (Romanian) and a concentrated model (Egyptian). However, even in terms of SLC, the communities of Bangladesh and Pakistan do not remain stable over time (date not shown). To check the accuracy of the SGI, we compare the SGI results with the ones obtained using global concentration indices known in the literature: Delta index (DEL) (Hoover, 1941; Duncan *et al.*, 1961) and Absolute concentration index (ACO) (Massey and Denton,

[5] For reasons of space, we have reported only the main ones. All SLC can be downloaded at the following link: Supplementary files (SIEDS 2023).

1988). Although the results might appear controversial, we find a significant negative correlation between SGI and DEL (-0.58) and ACO (-0.48) applying the latter to the same database. In our opinion, the explanation for this result lies in the fact that both DEL and ACO are basically aspatial indices like the Gini index. Greater population concentration implies low territorial uniformity of the population and consequently low SGI values.

Figure 5 – *Comparison between SLC (from left to right): CHI vs ITA; EGY vs ITA; PHI vs ITA; MOR vs ITA; UKR vs ITA and ROM vs ITA (Years=2021).*

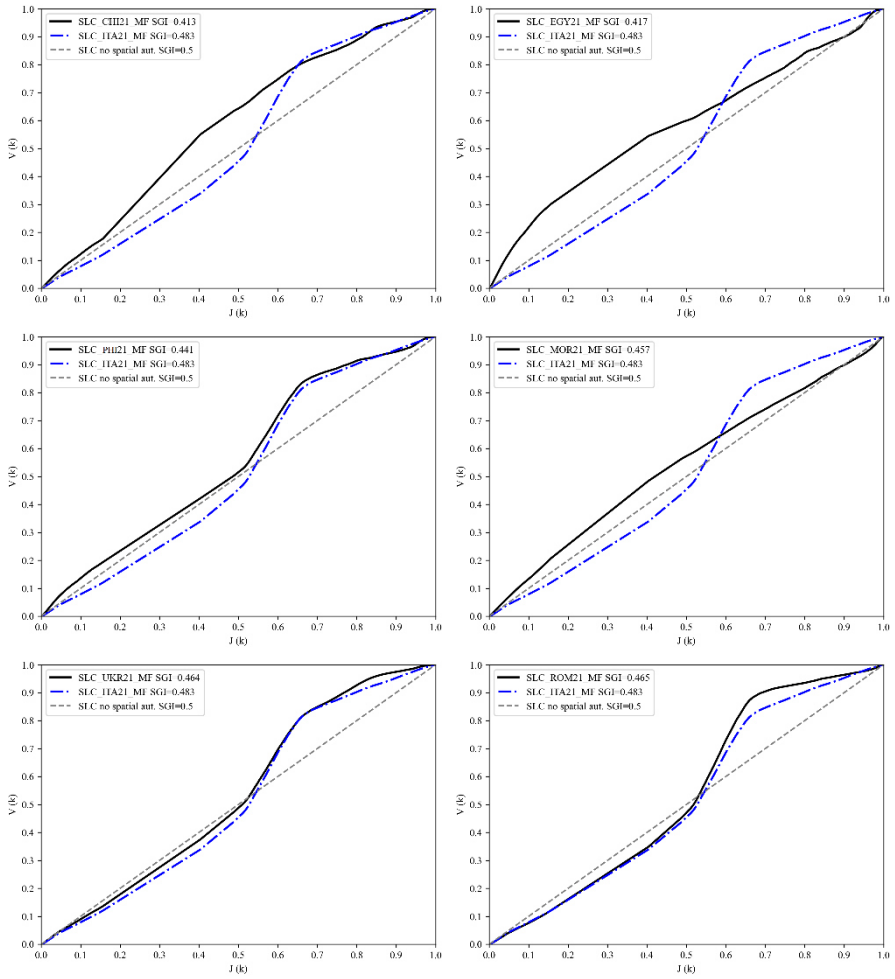
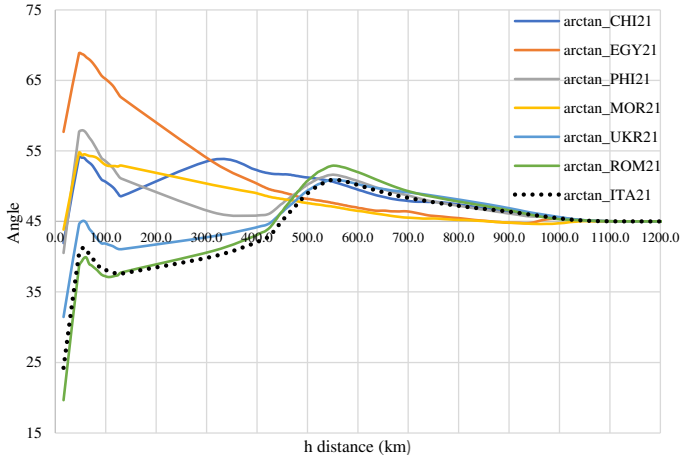


Figure 6 – *Value of the angle (degree) for the AF for the CHI, EGY, FIL, MOR, UKR and ROM citizenships (years 2021).*



4. Final remarks

In accordance with a new measure of spatial concentration proposed by Mucciardi and Benassi (2023) and Benassi *et al.* (2023a), the paper attempts to analyze the settlement models of the main foreign communities residing in Italy in the periods 2001, 2011 and 2021. The main results obtained can be summarized as follows. Compared to the traditional methodology in the field of population studies, SGI, SLC and AF represent an "*integrated tool*" for measuring spatial segregation through a global measure of spatial concentration (SGI), its graphical representation makes it possible to evaluate how a population is distributed in the territory (SLC) and the level of spatial autocorrelation of the population (AF). The application of the procedure to Italian municipal data makes it possible to differentiate the settlement models between foreign communities. Each community, even if the indicators of spatial concentration indicate stability over time (with some rare exceptions), seems to have its own settlement model: the populations of Romanians and Ukrainians demonstrate a settlement model similar to that of the Italians whereas the Egyptian and Chinese settlement models are different. To conclude, population statistics in general and especially spatial concentration measures are topics of study that can be useful in many ways. EU and national integration studies may use data obtained from spatial concentration measurement to understand the present condition of the ongoing efforts to integrate foreign populations residing in Italy. Knowing where national groups congregate also helps to understand the demand for various

occupations including construction, agriculture, domestic services and industrial labour markets. This in turn can assist in urban planning and the creation of education, employment, and social programs. It is hoped that the research methodology and results presented here will be useful to other EU member states in the effort to promote integration and social cohesion. From methodological point of view, other improvements in the model are planned by the authors. We will implement a neighbourhood system based on other buffer distances or spatial weight matrices. Furthermore, the extension of inferential methods to test the significance of SGI (i.e. Monte Carlo test) is being planned.

References

ARBIA G., PIRAS G. 2009. A new class of spatial concentration measures. *Computational Statistics and Data Analysis*, Vol. 53, No. 21, pp. 4471-4481. https://doi.org/10.1016/j.csda.2009.07.003.

BENASSI F., MUCCIARDI M., PIRROTTA G. 2023a. Looking for a new approach to measuring the spatial concentration of the human population. *Journal of Official Statistics*, forthcoming.

BENASSI F., MUCCIARDI M., CARELLA M., NACCARATO A., SALVATI L. 2023b. Spatial segregation in action? An empirical assessment of population concentration of foreigners and nationals in Italy, 2002-2018. *International Migration Review*. https://doi.org/10.1177/01979183231170808.

BENASSI F., IGLESIAS-PASCUAL R., SALVATI L. 2020. Residential segregation and social diversification: Exploring spatial settlement patterns of foreign population in Southern European cities. *Habitat International*, Vol. 101, https://doi.org/10.1016/j.habitatint.2020.102200.

BLANGIARDO G. C. 2005. I processi di immigrazione: dall'illegalità alla regolarizzazione'. In: Livi Bacci M. (Ed.), *L'incidenza economica dell'immigrazione*, Quaderni Cesifin, n. 20, Giappichelli.

DAWKINS C. J. 2004. *Measuring the spatial pattern of residential segregation*. Urban Studies, Vol. 414, pp. 833–851. https://doi.org/10.1080/0042098042000194.

FERRARA R., FORCELLATI L., STROZZA S. 2010, Modelli insediativi degli immigrati stranieri in Italia, *Bollettino della Società Geografica Italiana*, Vol. 13, No. 3, pp. 619-639.

DUNCAN O. D., CUZZORT R. P., DUNCAN B. 1961. *Statistical geography: Problems in analyzing areal data*. New York. The Free Press of Glencoe.

HOOVER E. 1941. Interstate redistribution of population, 1850-1940. *Journal of Economic History*, Vol. 1, pp. 199-205.https://doi.org/10.1017/S0022050700052980.

ISTAT 2023. Retrieved from: http://dati.istat.it/Index.aspx?QueryId=19125.

LORENZ M. O. 1905. Methods of measuring the concentration of wealth. *Publications of the American Statistical Association*, Vol. 970, pp. 209-219. https://doi.org/10.2307/2276207.

MASSEY D. S., DENTON N. A. 1988. The dimensions of residential segregation, *Social Forces*, Vol. 67, No 2, pp. 281-315.

MUCCIARDI M., BENASSI F. 2023. Measuring the spatial concentration of population: a new approach based on the graphical representation of the Gini index. *Quality & Quantity,* Vol. 57, pp.5193-5211. https://doi.org/10.1007/s11135-022-01607-2.

MUCCIARDI M. 2008. Use of a flexible weight matrix in a local spatial statistic. In: Giusti A., Ritter G. and Vichi M (Eds.), *Classification and Data Mining,* Vol. 5. Naples: Ed. Sc. Italiane, pp. 385-388.

PANZERA D., POSTIGLIONE P. 2020. Measuring the spatial dimension of regional inequality: An approach based on the Gini correlation measure. *Social Indicators Research*, Vol. 1482, pp. 379-394. https://doi.org/10.1007/s11205-019-02208-7.

REARDON S. F., O'SULLIVAN D. 2004. Measures of spatial segregation. *Sociological Methodology*, Vol. 341, pp. 121-162. https://doi.org/10.1111/j.0081-1750.2004.00150.x.

REY S. J., SMITH R. J. 2013. A spatial decomposition of the Gini coefficient. *Letters in Spatial and Resource Sciences*, Vol. 62, pp. 55-70. https://doi.org/10.1007/s12076-012-0086-z.

STROZZA S. 2004. Estimates of the Illegal Foreigners in Italy: A Review of the Literature, *International Migration Review*, Vol. 38, No. 1, pp. 309-331.

TÜRK U., ÖSTH J. 2023. Introducing a spatially explicit Gini measure for spatial segregation. Journal of Geographical Systems, Vol.25, pp.469-488. https://doi.org/10.1007/s10109-023-00412-1.

Massimo MUCCIARDI, Department of Cognitive Science, Education and Cultural Studies, University of Messina, Italy, massimo.mucciardi@unime.it.

Giovanni PIRROTTA, IT Staff, University of Messina, Italy, gpirrotta@unime.it.

Mary Ellen TOFFLE, Department of Political and Juridical Sciences, University of Messina, Italy, maryellen.toffle@unime.it.

THE STATISTICAL REGISTER FOR PUBLIC ADMINISTRATIONS: THE REFERENCE FRAMEWORK AND SOME METHODOLOGICAL ASPECTS[1]

Roberta Varriale, Nevio Albo, Cecilia Casagrande, Valeria Olivieri

Abstract. During the last decade, the Italian National Institute of Statistics has been engaged in a modernization program involving the use of statistical registers integrated into a single logical environment, the Italian Integrated System of Statistical Registers (ISSR), for supporting the consistency of statistical production processes and improving the quality of information for users. One object of the ISSR is the satellite statistical REgister for Public Administrations (REPA) that contains information on structural and economic variables on a subset of the Italian Public Administrations (PA). This subset includes specific sub-populations covered by the base business register related to the PA. Therefore, REPA extends, for each of those units, structural information coming from the base register with some economic variables obtained as the result of integration of data coming from administrative and survey sources. In this paper we describe some methodological aspects of the design and implementation of the production process, together with the structural metadata and the proposal of a structural variable for the functional classification of the statistical units.

1. Introduction

During the last decade, the Italian National Statistics Institute (Istat) has been engaged in a modernisation programme involving the use of statistical registers integrated in a single logical environment, the Italian Integrated System of Statistical Registers (ISSR) (Luzi *et al.*, 2019). ISSR comprises a series of statistical registers (basic, thematic and extended) that centralise and integrate data from administrative sources, statistical surveys carried out by the institute and new and emerging sources for the ongoing production of official statistics. The ISSR aims to ensure uniform management of the different themes (social, environmental, economic statistics, etc.) and conceptual, statistical and physical integration between the statistical units that make it up. One of the objects of the ISSR is the extended statistical REgister for Public Administrations (REPA), which includes the subset of Italian Public

Administrations in the so-called "S13 list" produced by Istat (https://www.istat.it/it/archivio/190748). In this paper we consider the units in the S13 list together with some structural information as the base Register of REPA and we will call it "Register S13" (RS13). Therefore, REPA extends the RS13 with some economic variables obtained as a result of the integration and elaboration of data coming from administrative and survey sources. The production process of the statistical Register for Public Administrations includes 3 objects: (i) the information base created by the integration of sources; (ii) the REPA itself; (iii) Frame PA.

The structure of the paper is as follows. Section 2 describes the production process of REPA and Frame PA for the whole reference population in the S13 list of public institutions, and section 3 focuses on the sub-population for Territorial Governments. Section 4 presents the metadata of REPA. Section 5 describes the new structural classification of institutions. Section 6 contains concluding remarks.

2. REPA and Frame PA

In the Section, we describe the production process of REPA and Frame PA for the whole reference population in the S13 list.

2.1. The units and the variables

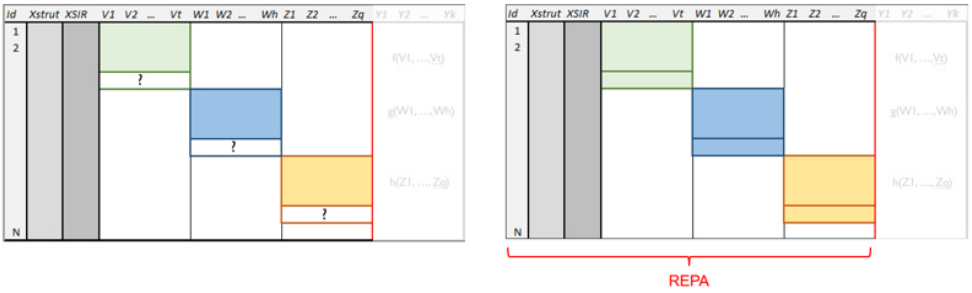
REPA is an extended Register of the base RS13, including all institutional units that are part of the general government sector and whose accounts contribute to the compilation of the consolidated profit and loss account of the General Government of Italy. RS13 is compiled according to ESA 2010, as defined by Regulation (EU) of the European Parliament and of the Council no. 549/2013, and on the basis of the interpretations of the ESA, provided in the Manual on Government Deficit and Debt published by Eurostat (Eurostat, 2019). According to the Regulation, RS13 is divided into 3 institutional sub-sectors: Central government (excluding social security funds), Local government (excluding social security funds) and Social security funds. It is possible to classify each sub-sector into different institutional typologies, some of which already populate the prototype version of REPA for Territorial Government. REPA contains a subset of the structural variables from RS13 and ISSR, a variable related to the proposed new structural classification of institutions, and a set of micro-data of an economic nature for each type of public institution. The structural variables are: identifiers and register variables, territorial variables, stratification variables, activity status, date of inclusion and possible exclusion from sector S13, transformation events. One of the ISSR variable is *Employees*. Frame PA is derived from REPA by aggregating REPA data into homogeneous variables, processable for the entire reference population and referred

to as the "Frame PA variables": *Current revenues*, *Compensation of employees*, and *Purchases of goods and services*.

2.2. REPA and Frame PA production process

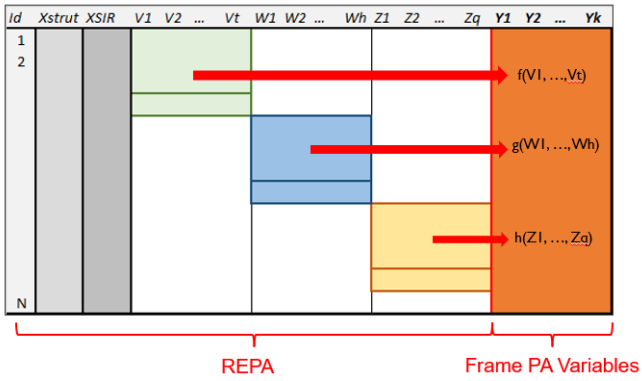
The REPA and Frame PA production process works in a differentiated way for different groups of institutions (sub-populations), defined through the classification of some structural variables (see Section 4.2). Its representation is in Figures 1 and 2, by using, as an example, three different sub-populations represented with different colours. *Id* is the identification code of each unit belonging to the RS13 and *Xstrut* are the structural variables coming from this Register together with the variable "New structural classification of institutions". *XSIR* represents the variables from the ISSR. Variables *V, W,* and *Z* are the variables of REPA for each sub-population, and *Y* are Frame PA variables that represent the final output of the process. Each sub-population has a different information structure (content of variables, sources, etc.). Figure 1.a shows how the output of data collection, harmonisation into statistical concepts and integration is differentiated by sub-population. This is followed by a phase of review, editing and imputation of total and partial non-response: the result is a complete dataset for each sub-population (Figure 1.b).

Figure 1 – *REPA production process: (a) Data collection, harmonization, and integration, (b) Data review, edit and imputation.*



As shown in Figure 2, this is followed by a process of transforming the original information into the output variables that will make up the Frame PA. This process is also differentiated according to the sub-populations.

Figure 2 – *REPA production process: Frame PA variables.*



REPA is still under development, the design and implementation of the Register is at a good stage for the sub-population of Territorial Governments (Varriale *et al.*, 2021). In the following section we describe in details its production process.

3. REPA and Frame PA for Territorial Government production process

In the Section, we describe the production process of REPA and Frame PA for the Territorial Governments (TG); we will refer to REPA TG as REPA.

3.1. The units and the variables

TG include regions and autonomous provinces and local governments i.e. municipalities, unions of municipalities, provinces, mountain communities and metropolitan cities. In 2019, the population of TG consists in 8749 units, representing the 84.3% of the total RS13 population.

The economic variables of REPA include accrual and cash values, for both revenues and expenditures. The accrual data for the revenue are the assessments (E1) while the cash data are the collections in accrual (E2) and the residual accounts (E3). For expenditures, the accrual data are the commitments (S1), the cash data are the payments on accrual (S2) and the residual accounts (S3). The information for both revenues and expenditures is organized into several hierarchical levels. By aggregating a selection of items, we obtain the economic variables of Frame PA (Guandalini *et al.*, 2022).

3.2. REPA and Frame PA production process: data sources and imputation

The primary source of information of REPA is the Public Administration Database (BDAP), providing all the information needed for REPA and Frame PA economic variables. BDAP will cover in the future also other types of institutions.

For the population of the regions and autonomous provinces, the BDAP source has no total non-response. Therefore, the variables REPA and Frame PA are obtained through a transcoding procedure. On the other hand, the population of local governments is characterised by total non-response. Before applying the transcoding procedure to obtain the REPA and Frame PA variables, a data imputation step is necessary. The imputation procedure uses an auxiliary information source, i.e. the Information System on the Operations of Public Bodies (SIOPE).

The entire REPA and Frame PA production process is repeated over time, using the same RS13 reference year and different provisions of both BDAP and SIOPE. It is important to note that BDAP and SIOPE are related to the same reference universe, the units belonging to the S13 list, but have a different periodicity in terms of availability of the data. For BDAP there are six different provisions relating to data referring to the same period, while SIOPE's provision is "continuous". In a generic year, at time *t* in the July month there is the first provisional supply of the BDAP's data referring to time *t-1*. The definitive provision usually takes place in May of the following year and the data are referred to time *t-2*. Just to give an example, let's consider the production process of REPA, reference year 2019: data from RS13 refer to 2019, the first BDAP provision for the same reference year is in July 2020, and the last BDAP provision is in May 2021.

BDAP provisions are characterised by a gradually decreasing non-response rate. Therefore, with regard to the 2019 data, the non-response rate for the total of local units decreases from 20.3% for the July 2020 BDAP provision to 7.4% and 3.6% for the following two provisions, but it remains high for mountain communities (from 58.3% to 52.3% and 44.4%) and for unions of municipalities (from 38.3% to 31.0% and 24.4%). The cases of total non-response of the BDAP source are distinguished according to the presence of information in the SIOPE source. The assumption underlying the imputation process is that the BDAP source is complete, i.e. if the BDAP source is available, no imputation is necessary.

The imputation method is described below. In the following notation, *V1-V3* denote both variables 1 to 3 of revenue (*E1,E2,E3*) and variables 1 to 3 of expenditure (*S1,S2,S3*) coming from BDAP. Variable *V4* is the result of the sum of *V2+V3* and *V4SIOPE* is variable *V4* coming from SIOPE, both for revenue and expenditure.

The first assumption underlying the imputation procedure is that for each unit *i* and for each item (148 for revenue and 22 for expenditure): $V4_i = V4SIOPE_i$, where $V4_i$ and $V4SIOPE_i$ represent the value of *V4* and *V4SIOPE* for unit *i*, respectively.

Therefore, the first step in the imputation process is to impute variable *V4* with the value of variable 4 from SIOPE: $V4_i^* = V4SIOPE_i$, where $V4_i^*$ is the imputed value of variable *V4* for each non-responding institution *i*. Subsequently, to impute the variables *V1* and *V2*, we use the median ratio between *V1* and *V4* - $r1(t)=V1(t)/V4(t)$ - and *V2* and *V4* - $r2(t)=V2(t)/V4(t)$ - calculated in each imputation stratum by using of all the information collected on the respondent units at time *t*. Therefore, the imputation of variables *V1* and *V2* is carried out by the relations:

- $V1_i^*(t)= MEDstr[r1(t)]\ V4_i^*(t)$
- $V2_i^*(t)= MEDstr[r2(t)]\ V4_i^*(t)$

where *MEDstr[r1(t)]* and *MEDstr[r2(t)]* are the median of *r1(t)* and *r2(t)*, calculated in appropriate imputation strata defined by the variables: Region, Institutional typology of entities, and aggregations of items in "Piano dei conti". Variable *V3* is then obtained through the relationship *V3=V4-V2*.

The choice of the imputation method was based on different types of evaluation. In particular, the two main working hypotheses for imputing missing values at time *t* were: (i) *cross method*: to use as auxiliary one the information collected in all responding institutions at time *t*; (ii) *longitudinal method*: to use as auxiliary information the longitudinal profile of the institution itself from the previous year, if available; and to use the information on all the responding institutions at time *t*, otherwise. First, Monte Carlo simulation studies were carried out on units with information from both BDAP and SIOPE to assess the impact of the imputation strategy on the estimates of the REPA variables at the aggregate level, i.e. by region, by title and by type of institution. Then, we analysed the longitudinal data of the institutions: the distribution of the items is subject to large variations from year to year, which invalidates the assumption underlying the longitudinal method of stability of the institutional profile. Furthermore, we evaluated the practical system management: the cross method is more responsive to any changes that occur on the respondents and is characterised by more simplicity for managing a control process. Finally, the goodness of fit of the chosen imputation strategy was confirmed by the analyses of subject matter experts of the final data.

The entire REPA production process is scheduled in this way: different provisional data are available during the year and for each reference period of the data. Frame PA is released once in a year for the final data delivery (May *t+2*). The process works in synergy with other Istat production structures, in particular those dealing with economic statistics and National Accounts.

4. Metadata

In the Section, we present the metadata activity that aims to describe how the concepts used in REPA have been structured.

4.1. REPA: metadata activity on units of analysis

The structured description of the metadata is one of the outputs of the REPA and provides information on which are the concepts that define it, framing and organizing them in a standardized way. The metadata activity therefore aims to apply a model for a structured description of concepts and it was carried out according to the principles of the standard "Generic Statistical Information Model (GSIM)" (HLG-MOS, 2020). On the basis of GSIM, it was possible to provide the REPA with the fundamental concepts for its documentation. Subsequently, we proceeded with the definition and description of the units type of the Register, the variables involved and the classifications correlated to the categorical variables.

Metadata activity, in addition to providing a set of organized and documented concepts, promotes the integration and the sharing between the REPA and the other ISSR Registers, especially with the base RS13. As introduced in Section 1, being REPA an "extended" Register, it has the specific purpose of extending the information of the base RS13, providing specific information that is not contained therein.

The elementary unit of analysis of the REPA is based on the more general concept of “economic unit”. This concept, defined in the framework of other Registers, describes an "entity that carries out economic activity of production, consumption or exchange". However, the REPA Register is just referred to "Institutional Units belonging to the institutional sector of Public Administrations" disseminated through the S13 list that is redefined every year and related to an annual reference period.

4.2. REPA: metadata activity on variables and classifications

Categorical variables are those variables that allow the reference population to be "partitioned" on the basis of specific characteristics. All those variables that have an *enumerated value domain* (according to GSIM concept), with defined modalities, belong to this type of variables. These variables are always associated with a classification that organizes their modalities in a structured way. In the case of REPA, the main categorical variables are: the Italian statistical classification of economic activities (Italian version of the Nace, *Nomenclature statistique des activités économiques dans la Communauté européenne*), the institutional typology and the legal form. These variables are inherited from the base RS13 . Some of them, together with the type of accounting of the institution and the economic operations

of the institution itself, have been used to determine the different sub-populations on which the entire REPA construction process have been built (see Section 2.2).

The variables that extend the REPA information set with respect to the base RS13 are those numerical (with *described value domain* according to GSIM), i.e. those relating to economic aggregates of income and expenditure provided, in the case of the sub-population of Territorial Governments (TG), from administrative sources BDAP and SIOPE. Up to now, the REPA prototype has been implemented only for the sub-population of TG which, as already highlighted, includes the institutional typologies of regions and autonomous provinces and local governments. The numerical variables of REPA, unlike the categorical ones, have specific sources depending on the reference sub-population and therefore the treatment relating to the non-responses is designed taking into account the source that is used for the imputation (Figure 1).

Table 1 – *The data structure components of final output FRAME PA. Territorial Governments sub-population. IU: Institutional unit.*

IU	NACE	Institut. typology	Type of Accountab.	Economic operability	CR	CE	PGS
1	…	Municip.	Financial	Current/ No current	Num.var.	Num.var.	Num.var.
2	…	Municip.	Financial	Current/ No current	Num.var.	Num.var.	Num.var.
3	…	Municip.	Financial	Current/ No current	Num.var.	Num.var.	Num.var.
…	…	…	…	…	…	…	…
i	…	Province	Financial	Current	Num.var.	Num.var.	Num.var.
…	…	…	…	…	…	…	…
j	…	Region	Financial	Current	Num.var.	Num.var.	Num.var.
…	…	…	…	…	Num.var.	Num.var.	Num.var.
n	…	…	…	…	Num.var.	Num.var.	Num.var.

Table 1 shows the data structure components of the final output FRAME PA, for the TG institutional units (IU). As introduced, the informative detail of the numerical variables within the FRAME PA output (derived from REPA) consists of three variables, one relating to income variables and two relating to expenditure variables: *Current revenues (CR)*, *Compensation of employees (CE)* and *Purchase of goods and services (PGS)*. These variables derive from the aggregation of available income and expenditures items, at a greater level of detail. The choice of these three final variables allows a comparability of the meanings of the economic variables between units classified on the basis of the variable "type of accountability" (financial or economic-patrimonial).

5. The new structural classification of a legal-functional and territorial type

The design and implementation of the REPA and Frame PA includes the definition of a new structural classification of the statistical units to support both the treatment processes and the analysis of the economic variables present in the extended register.

The need to define a new structural variable arose in the light of the descriptive limitations that distinguish the other structural stratification variables borrowed from the ISSR in the REPA, which are normally entrusted with the task of classifying all the units of the S13 list for any control activity, publication and interpretation of the related data. Those variables are: *institutional typology*, *Nace*, *territory* (municipality, province, region) of PAs. The first and most complex of these variables, the institutional typology variable, classifies the units of the S13 liston the basis of their legal, functional and territorial characteristics, but in an incomplete and incoherent manner that does not ensure that a consistent part of PAs are assigned to homogeneous sub-populations in terms of all the main characteristics just mentioned. On the other hand, Nace and the sub-variables on the location of units have the limitation of being based on a single classification criterion, namely functional or territorial, which alone is not capable of distinguishing PAs into sufficiently homogeneous sub-populations.

Therefore, another structural stratification variable was created with the aim of building an effective tool for describing the units, i.e. capable of distinguishing the largest number of PAs according to coherence and relevance, organising them into homogeneous subpopulations.

In terms of coherence, main general methodological criteria underlying the statistical classification activity have been applied for the new variable, with greater rigour than is expected for the structural variables institutional typology and Nace. These criteria are:

- completeness of the classes (few units in "other" class)
- mutually exclusive classes (same level of generality between classes)
- no underpopulated classes (< 10 units).

In terms of both relevance and coherence, it was envisaged that the new variable would classify the units on the basis of the same characteristics (mentioned above) of the individual structural variables of the institutional typology, Nace and localisation, synthesising and/or integrating them through unique and more significant dimensions. These dimensions are associated with the main characteristics of the units:

- for legal characteristics → functional autonomy level; governance model (association or institution);

- for functional characteristics → nature of activity (administrative functions, operating services, final services); sector of activity (health care, protected area management, business loans, etc.);
- for territorial characteristics → territorial level of activity (national, regional, local).

The resulting new variable produced classes of units that were completely homogeneous from a methodological and structural point of view in 92.5% of cases, compared with 37.5% of the classes belonging to the main structural stratification variable *institutional typology*.

This greater descriptive effectiveness of the new structural variable of a juridical-functional and territorial type manifests itself in a particular way with reference to the set of 894 units (8.6% of the total S13 population in 2019) whose primary source of information is the Istat statistical survey RIDDCUE (Collection of Information, Data and Documents necessary for the Classification of Economic Units in the institutional sectors established by the European System of Accounts 2010) (https://www.istat.it/it/archivio/219736), and which at the same time cannot be effectively distinguished by the structural variable of the institutional typology.

Unlike the other sources of information for the economic variables of the extended register (both from administrative archives and the statistical surveys), the source constituted by the RIDDCUE survey does not identify a specific and homogeneous sub-population of the base RS13. This is due to the specific purpose of this annual survey, which is to collect economic information on a heterogeneous set of units to determine their affiliation to one of the institutional sectors defined by ESA 2010, including sector S13. Of the units whose primary source is the RIDDCUE survey, the majority (894 units) cannot be broken down into homogeneous sub-populations from a legal and/or functional and/or territorial point of view using the institutional typology variable, because it reserves only generic, residual and underpopulated classes for this set of units. This heterogeneous set of institutions represents the second largest group of units of REPA after the sub-population of TG and the one with the highest total non-response rate.

Table 2 shows the main groups of the RS13 units according to their primary sources of information and their classifiability by institutional typology.

The importance of the new classification variable is manifold.

For a subset of PAs only the new legal-functional and territorial structural variable has the characteristics of isolating, autonomously or by crossing with other structural variables, homogeneous sub-populations. This characteristic is potentially important for the design of the imputation process of total non-response to economic data. In fact, the results of the simulations for TG have shown that imputation using a cross method based on the definition of homogeneous strata for imputation

guarantees greater stability of estimates over time and therefore less distortion of the data than a longitudinal method.

Table 2 – *Main groups of units by primary source of economic data and classifiability according to institutional typology (i.t.), year 2019.*

Groups	#	%	Primary source	Classifiability by i.t.
Territorial governments with specific source	8749	84,3	BDAP	Yes
PAs without specific source	894	8,6	RIDDCUE survey	No
Other PAs with specific source	405	3,9	General Government accounts Consolidated income statements of NHS Bodies Economic and patrimonial balance sheets of the Universities Final accounts of the Chamber of Commerce Collection of final balance sheets of Social Security Institutions (BICEP)	Yes
PAs without specific source	325	3,2	RIDDCUE survey	Yes

Furthermore, the introduction of the new structural variable in REPA and Frame PA makes it possible to deepen the description and analysis of the economic variables of these units, which would otherwise be associated with a composite and numerous subgroup of "other public administrations", thus compromising any data interpretation activity.

Finally, if the REPA and the Frame PA were to include in the future the population of public institutions of the Register Asia - Public Institutions, the new variable would ensure the same results in terms of data processing, dissemination and analysis also for other 2853 active institutional units from 2019 that at the moment are not considered in the RS13 framework. In fact, the Register Asia - Public Institutions, representing another object of the ISSR, includes more units than RS13. Those additional units would allow a better integration of all statistical registers on public units.. This macro-group of units would present similar problems concerning their breakdown into homogeneous sub-populations on the basis of the source of the economic data and of the available structural variables.

Despite the great benefits that this new classification can bring, more work is needed to validate this information.

6. Conclusion and future work

In this paper we have described some methodological aspects of the design and implementation of REPA production process, together with the structural metadata and the proposal of a new variable for the structural classification of statistical units.

Other work remains to complete the REPA, such as the design and implementation of the REPA for the other sub-populations, the completion of the metadata process and the further integration of the process with the other Istat production processes. Finally, a process for validating the quality of the register needs to be developed.

References

EUROSTAT 2019. Manual on Government Deficit and Debt IMPLEMENTATION OF ESA 2010 2019 edition, *Eurostat Manuals and guidelines, Economy and finance*. Publications Office of the European Union, available on-line: https://ec.europa.eu/eurostat/web/products-manuals-and-guidelines/-/ks-gq-19-007 (retrieved on 13/07/2023).

HLG-MOS 2020. The Generic Statistical Information model (GSIM v.1.2), available on line https://statswiki.unece.org/display/gsim (retrieved on 13/10/2023)

GUANDALINI A., PASSANTE D., VARRIALE R. 2022. The revenues of local governments in the statistical register for public administrations: inequality decomposition by sources, *RIEDS*, Vol. 76, No.3, pp. 17-28.

LUZI O., ALLEVA G., SCANNAPIECO M., FALORSI P.D. 2019. Building the Italian Integrated System of Statistical Registers: Methodological and Architectural Solutions. ESS Workshop on the use of administrative data and social statistics, Valencia,4-5June2019, available on-line:https://cros-legacy.ec.europa.eu/system/files/building-italia-integrated-system_istat_0.pdf (retrieved on 09/10/2023).

VARRIALE R., LORI M., MANTEGAZZA F. 2021. Stato di avanzamento dei lavori Focus: Frame PA enti territoriali. Nota tecnica Istat, Dicembre 2021.

Roberta VARRIALE, Sapienza University of Rome, roberta.varriale@uniroma1.it
Nevio ALBO, Istat, nalbo@istat.it
Cecilia CASAGRANDE, Istat, casagran@istat.it
Valeria OLIVIERI, Istat, vaolivie@istat.it

INSTANT MESSAGING TOOLS IN OFFICIAL STATISTICS: A USAGE MODEL IN THE 7° ITALIAN CENSUS OF AGRICULTURE

Claudia Fabi

Abstract. The data collection of the Italian 7th General Census of Agriculture took place from January to July 2021, through a synchronous multi-technique (CAPI, CATI, CAWI) design. During the survey, innovative contact tools were offered to respondents, alongside traditional communication channels. Through SMS and WhatsApp it was possible to collect the requests of over 2,000 farms, facilitating subsequent contact with the interviewers. The following work intends to show the results obtained and the propensity to use innovative contact channels. It proposes a model that allows correlating this use to some representative indicators of the territory characteristics, demographics, and the diffusion of the so-called "Digital culture".

1. The reference context

Between January and July 2021, the survey of the seventh General Census of Agriculture took place in Italy, the last one before the advent, also in this sector, of annual Permanent Censuses, on a sample basis.

For the first time, in an Italian census survey, a synchronous multi-technique design (CAWI, CAPI, CATI) was adopted, which allowed respondents a large amount of flexibility in choosing how to fill in the questionnaire (De Gaetano *et al.*, 2021 and De Gaetano *et al.*, 2022).

Important technological innovations, necessary to ensure the fluidity of data collection operations conducted by different data collection networks, have been supported by new strategies of contact with respondents, capable of responding more promptly to user requests if compared to more traditional tools such as the toll-free number or the email address of the census.

Starting from May 2021, a new contact channel for respondents was introduced, through a mobile phone number to which users could send SMS or WhatsApp messages, to request an appointment to subsequently complete the interview by phone. The main results obtained will be illustrated below, examining in detail the diffusion and propensity to use innovative contact tools on the national territory,

compared with the reference population of the Census. Finally, a correlation model will be proposed among the propensity to use innovative tools and territorial indicators, such as sectoral and socio-demographic context indicators.

2. Characteristics of the SMS and WhatsApp service in supporting respondents

The communication and assistance channels dedicated to the respondents of a survey, whether a census or a sample one, can be classified as synchronous or asynchronous[1]. The former, usually managed by operators, trained in advance for the purpose, are represented in most cases by free toll-free numbers, which users can contact to request general information on the survey, on how to participate, or ask for assistance in completing it on specific thematic issues.

However, even a synchronous tool such as the toll-free number has potential critical issues in its use, especially in some time slots of the day or in periods of intense inbound traffic, concentrated in the weeks immediately following the massive sending of information letters to the population involved in the survey. In this sense, SMS and WhatsApp take the form of a hybrid channel of support for respondents, being able to guarantee both synchronous – when managed by operators – and asynchronous support.

In the case of the Agricultural Census, the mobile telephone number, available for sending SMS and WhatsApp by the respondents, was managed by operators on the same days and times as the CATI survey and remained available as an asynchronous channel for the rest of the time, including holidays[2]. Users became aware of the existence of this innovative channel by calling the toll-free number or, for some of them, through the memorandum sent to the units still not responding in June 2021, by paper letter. The service was activated with the sole aim of offering respondents a means of requesting an appointment to complete the questionnaire with the CATI technique, and not as a complete information channel. It should be remembered that the number to send SMS and WhatsApp messages was activated just later about the start of the survey, and in particular only for the last three months of fieldwork: from May to July 2021.

[1] A communication channel is defined asynchronous when allows the contextual interaction between user and operator (e.g. Toll Free Number, Contact Center, chat), while is defined as asynchronous the communication channel through which the user forwards a request that will be processed at a later time by the operators who supervise it (e.g. email, forum, etc.).

[2] The telephone survey was carried out from Monday to Friday, from 9.00 a.m. to 9.00 p.m. and on Saturday from 10.00 a.m. to 7.00 p.m., with the exception of public holidays.

The following tables display the characteristics of service users, taking into account the two available demographic factors: age and geographic origin.

Table 1 – *Channel used by respondents (SMS or WhatsApp).*

Channel	Users	
	A.V.	%
SMS	195	9.3
WhatsApp	1,908	90.7
TOTAL	2,103	100.0

Table 1 shows that, out of the approximately 2,000 received requests, over 90% are represented by WhatsApp messages, while only 9.3% are SMS. It appears that WhatsApp has almost completely replaced the messaging services offered by mobile network operators, in a widespread and established manner.

Table 2 – *SMS and WhatsApp users by age group*[3].

Age group	SMS or WA users		Census List	
	V.A.	%	V.A.	%
Up to 30 years	15	0.7	27,816	1.7
From 31 to 40 years	63	3.1	82,882	5.2
From 41 to 50 years	210	10.3	177,750	11.1
From 51 to 60 years	381	18.8	320,700	20.1
From 61 to 70 years	581	28.6	370,022	23.2
From 71 to 80 years	438	21.6	339,186	21.2
Over 80 years	343	16.9	279,555	17.5
TOTAL	2,031	100.0	1,597,911	100.0

As Table 2 shows, the age distribution of service users reveals an interesting phenomenon. Service users are not concentrated in the younger age groups, as one would expect if instant messaging services were more widely used by age groups that are more familiar with digital tools. When comparing the age distribution of service users with that of the census list, it becomes evident that it is within the age group of 61 to 70 years that the percentage of users is higher, compared to the reference population (28.6% vs. 23.2%). Even the younger age groups, up to 40 years old, have a usage profile that is lower than the incidence of the same age groups in the census population.

[3] Missing any other source of information, the distribution by age was obtained starting from the year of birth present in the tax ID code of the units in the census list. Where this information was not available, either in the census list or for SMS and WhatsApp users, the record was excluded from the calculation of the frequency distribution.

Table 3 – *SMS and WhatsApp users by region compared to the reference population.*

Region	SMS or WA users		Census List	
	V.A.	%	V.A.	%
Piedmont	115	5.5	78,492	4.6
Valle d'Aosta	4	0.2	4,265	0.3
Lombardy	102	4.9	75,205	4.4
Bolzano/Bozen	7	0.3	27,350	1.6
Trento	9	0.4	19,374	1.1
Veneto	94	4.5	112,553	6.6
Friuli-Venezia Giulia	32	1.5	27,059	1.6
Liguria	49	2.3	23,765	1.4
Emilia-Romagna	92	4.4	78,642	4.6
Tuscany	94	4.5	81,350	4.8
Umbria	56	2.6	41,897	2.5
Marche	60	2.9	51,219	3.0
Lazio	187	8.9	117,963	7.0
Abruzzo	95	4.5	66,212	3.9
Molise	17	0.8	28,600	1.7
Campania	103	4.9	134,413	7.9
Puglia	454	21.5	266,195	15.7
Basilicata	45	2.1	49,766	2.9
Calabria	151	7.2	132,553	7.8
Sicily	277	13.2	211,179	12.4
Sardinia	60	2.9	71,890	4.2
TOTAL	2,103	100.0	1,699,942	100.0

The distribution of service users across Italian regions also highlights contrasting trends. While internet penetration among households still faces a gap between Northern and Southern Italy, favoring the northern regions, the usage of instant messaging services in the Agricultural Census is higher in the southern regions, such as Puglia and Sicily, where the agricultural sector plays a significant role in the local economy. In Puglia and Sicily, over one-third of service users are concentrated, accounting for 34.7%, a proportion higher than the presence of these regions in the census list (approximately 28%).

3. An explanatory model of the use of SMS and WhatsApp

With these premises, it seemed reasonable to delve deeper into the analysis by comparing a series of indicators selected to represent the cross-sectional dimension of "*smart farmers*" and the use of SMS and WhatsApp in the census survey.

In particular, the indicators by Italian region reported in Table 4 were considered to identify and aggregate, through a composite index, two main dimensions:

- the propensity of individuals to use digital communication channels.

This is the case of indicators Ind_062, AVQ_1, AVQ_2, and AVQ_3 which represent both the presence, in households, of connectivity devices and technology necessary for the use of digital communication tools, and their use particularly in contacts with the Public Administration (PA);

- the development and entrepreneurship in the Agricultural Sector.
  This dimension is represented by the indicators Ind_460, Ind_250, and Ind_031 which report information capable of representing the agricultural vocation of a territory and the growth trend of the agricultural sector in the local economic context. These indicators should therefore adequately capture the phenomenon highlighted in the regional distribution of Table 3, emphasizing those territories where agriculture represents a leading sector.

Table 4 – *Indicators selected for the composite index.*

Indicator	Content	Year
Ind_460	Employment rate in rural areas (15-64 years)	2021
Ind_062	Percentage of diffusion of the Internet in families	2021
Ind_250	The growth rate of agriculture	2021
Ind_031	Agricultural land productivity	2021
AVQ_1	People aged 6 and over who use the Internet	2021
AVQ_2	Possession of at least 1 mobile phone in the family	2021
AVQ_3	People who have used the Internet to request information from PA Entities in the last 12 months	2021
AVQ_4	Use of social networks in the last 3 months	2020
ISTR_CP	People with at least a high school diploma	2021

Based on these hypotheses, these dimensions could represent the two main drivers to explain the inclination towards the use of innovative communication channels, even in fulfilling obligations towards the Public Administration, such as completing the Census questionnaire.

Furthermore, in the choice, only the indicators for which was available the annual data for 2021 or, in its absence, for 2020 at the regional level, were considered to represent phenomena contemporary to the data collection period of the Census.

The elaboration of a composite index was computed following the guidelines suggested by Mazziotta-Pareto for the synthesis of non-compensatory composite indexes (Mazziotta and Pareto, 2016) using the normalization formula that follows:

$$z_{ij} = 100 \pm \frac{(x_{ij} - M_{x_j})}{S_{x_j}} 10 \qquad (1)$$

where M_{x_j} e S_{x_j} is, respectively, the mean and the standard deviation of the indicator *j,* and the polarity, in the case of the subset of indicators chosen, is the positive one.

The composite index was computed following the formula:

$$MPI_i = M_{z_i} - S_{z_i} cv_{z_i} \quad (2)$$

where $cv_{zi} = S_{zi} / M_{zi}$ is the coefficient of variation for the unit *i*.

The following Table 5 shows MPI and the correlation with the percentage of users of SMS and WhatsApp, by region, computed concerning the total number of units in the census list. The Autonomous Provinces of Trento and Bolzano values have been computed separately.

The result, somewhat surprising, is a clear absence of correlation between the propensity to use innovative channels of communication with the PA and the PMI index calculated as detailed above. The two determinants of the phenomenon (digitalization and development of the agricultural sector) do not seem to have both an impact and a role in explaining the use of SMS or WhatsApp to communicate with institutions.

However, this is not entirely true. In fact, by calculating the simple correlations between each indicator and the percentage of SMS and WhatsApp users in the census list, a slight correlation has emerged with indicators Ind_031, AVQ_4, and Ind_250 (respectively the use of social networks, the productivity of agricultural land and the growth rate of agriculture), as shown in Table 6.

The same drivers are also highlighted in the *biplot* (Figure 1) among regions and the matrix of indicators, obtained with a principal component analysis, on the set of indicators listed in Table 4.

There is therefore a slight "*smart farmers*" phenomenon, as the correlations seem to suggest, even if this phenomenon does not significantly affect the propensity to use SMS and WhatsApp among the reference population.

Table 5 – *MPI Index by Region.*

Region	MPI	% Users
Piedmont	98.8	0.15
Valle d'Aosta	102.4	0.09
Lombardy	105.2	0.14
Province of Bolzano/Bozen	106.5	0.03
Province of Trento	103.1	0.05
Veneto	99.3	0.08
Friuli-Venezia Giulia	101.4	0.12
Liguria	109.1	0.21
Emilia-Romagna	104.1	0.12
Tuscany	104.2	0.12
Umbria	98.0	0.13
Marche	98.9	0.12
Lazio	105.7	0.16
Abruzzo	98.8	0.14
Molise	91.8	0.06
Campania	96.5	0.08
Puglia	93.2	0.17
Basilicata	91.1	0.09
Calabria	88.1	0.11
Sicily	89.2	0.13
Sardinia	100.0	0.08
Coefficient of correlation		*0.098*

Table 6 – *Coefficient of correlations between SMS and WA user percentage and regional indicators.*

Indicator	Content	Coefficient of correlation
Ind_460	Employment rate in rural areas (15-64 years)	-0.115
Ind_062	Percentage of diffusion of the Internet in families	-0.069
Ind_250	The growth rate of agriculture	0.177
Ind_031	Agricultural land productivity	0.250
AVQ_1	People aged 6 and over who use the Internet	-0.063
AVQ_2	Possession of at least 1 mobile phone in the family	0.121
AVQ_3	People who have used the Internet to request information from PA Entities in the last 12 months	-0.111
AVQ_4	Use of social networks in the last 3 months	0.313
ISTR_CP	People with at least a high school diploma	0.029

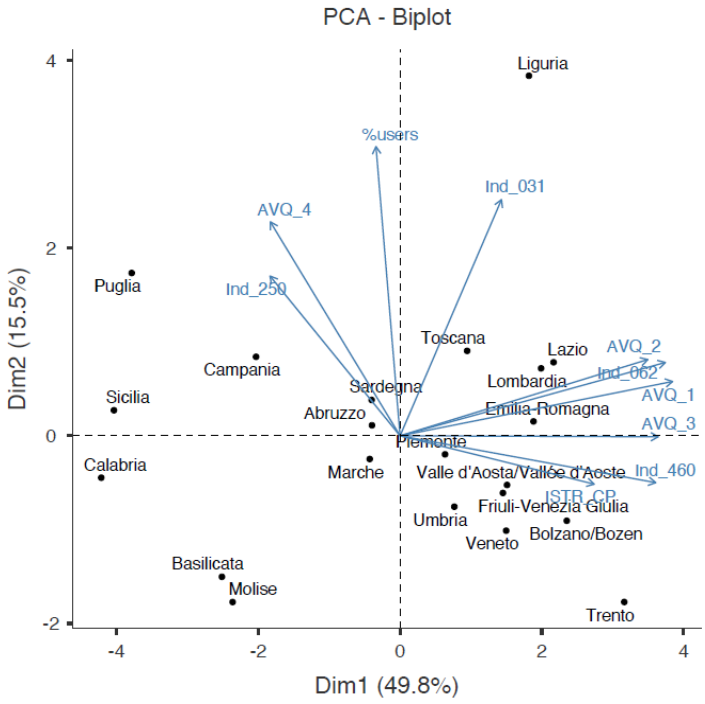
4. Conclusions

The correlation that emerges between individual/sectoral indicators and the use of SMS and WhatsApp, albeit slight, seems to suggest an expected tendency to favor these innovative channels exactly by people who possess dynamic characteristics

and a propensity to digitalization, even in the agricultural sector that is traditionally managed by small operators of high average age.

Furthermore, the absence of correlation with the PMI also suggests that investing in a digital communication channel such as WhatsApp to support respondents, is not an exclusive choice, in the literal sense of the term.

Figure 1 – *Biplot regions - indicators.*



It appears clear that there is not one (or more) specific socio-demographic profile that concentrates the majority of users of digital communication tools with the PA. This is probably due to the widespread diffusion of this tool among the population, which makes it a means of mass communication, just like we are used to considering the telephone (and consequently the Toll-Free Numbers). This makes SMS and above all WhatsApp, considering its user percentage, extremely interesting, right now, as an additional resource to complement the more usual contact strategies with respondents.

In the future, it is advisable to expand the use of these innovative contact services with the Public Administration, investing resources and conducting further analyses, not only in census surveys but also in sample surveys, particularly those concerning

households and individuals. Additionally, it will be crucial to invest in these innovative tools right from the beginning of field surveys, incorporating them clearly and prominently in the informative letters sent to sample units, to spread awareness and promote their usage.

Moreover, it is important to extend the use of these innovative channels not only for scheduling appointments for subsequent interviews but also as support tools for respondents, providing general information and assistance in completing online questionnaires. Digital assistance tools are in line with the dissemination and experimentation of Computer-Assisted Web Interviewing (CAWI) techniques in household surveys, allowing respondents to rely on quick and easily accessible assistance. Finally, it will be important to understand user satisfaction regarding the use of these instant communication services: an area of research that is still largely experimental in official statistics.

References

DE GAETANO L., FABI C., TRIOLO V. 2021. Tecniche integrate di data collection per il Censimento Generale dell'Agricoltura 2020. In *Scientific Posters of the 14° National Conference of Statistics,* https://www.istat.it/storage/14-Conferenza-nazionale-statistica/poster/022.jpg.

DE GAETANO L., FABI C., TRIOLO V. 2022. 7th Agriculture Census: strategies, and methodological and technological innovations for the survey. In *Book of Abstracts of 7° Italian Conference on Survey Methodology*. https://www.dropbox.com/s/a9uuajjvb7gbyd9/BoA%20_itacosm2022.pdf?dl=0.

ISTAT 2023. Indicatori territoriali per le politiche di sviluppo - https://www.istat.it/it/archivio/16777.

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, https://doi.org/10.1007/s11205-015-0998-2.

Claudia FABI, Istat – Istituto Nazionale di Statistica, claudia.fabi@istat.it

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

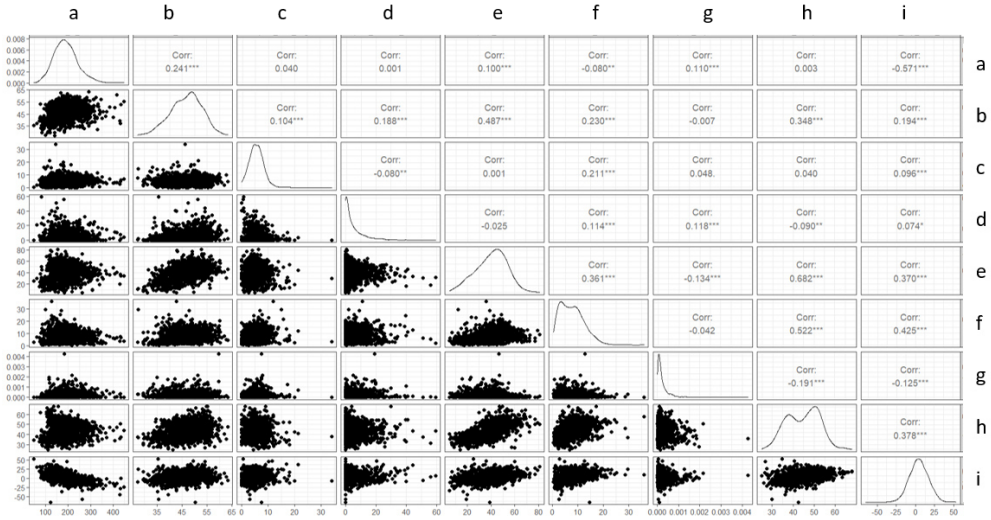
The "complexity index" was assessed using COMiC[4] (*COMposite Index Creator*), a free software designed to compute composite indices using various aggregation methods, based on the SAS programming language.

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.*

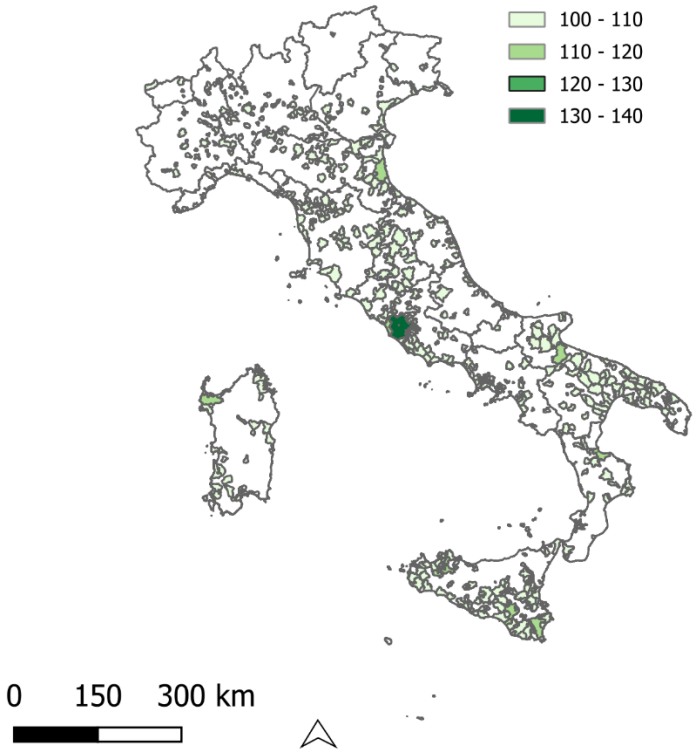
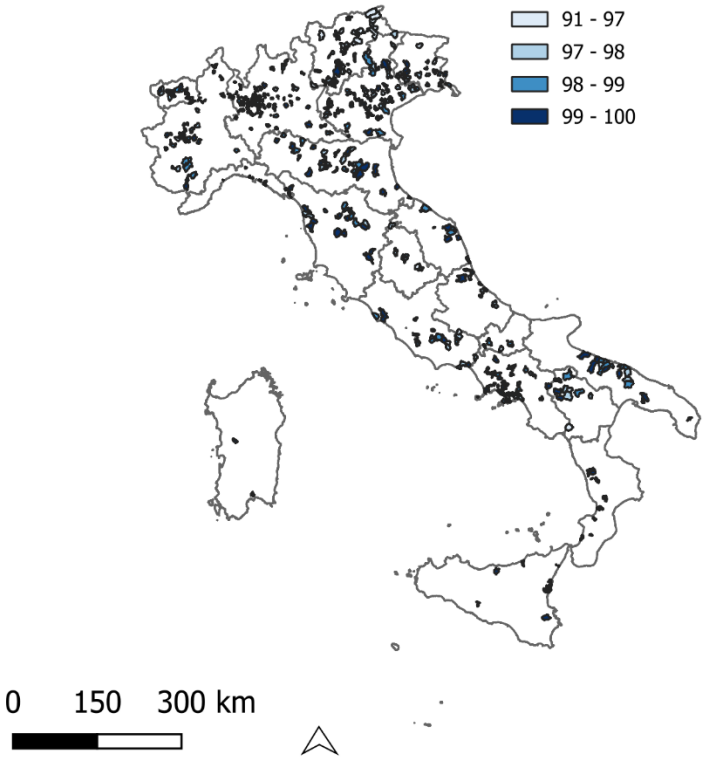


Figure 3 shows the geographical distribution of municipalities with a complexity index lower than 100. The lowest levels of complexity are represented in light blue, while the highest levels are shown in dark blue. Municipalities with a low complexity are concentrated in the North, especially in the 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

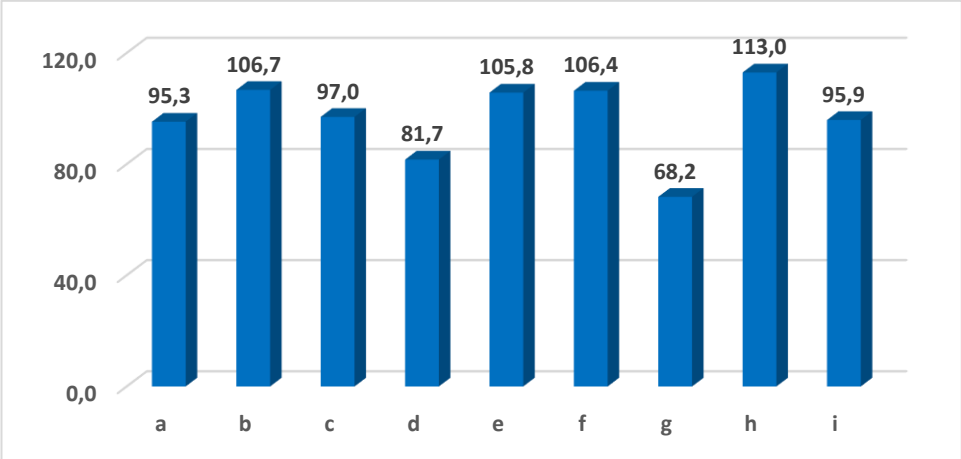
because the kurtosis index suggests that they have sharper or more peaked curves than a normal distribution.

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

THE CHALLENGE OF TRACKING EARLY CHILDHOOD DEVELOPMENT: A NEW METHODOLOGICAL APPROACH USING THE MAZZIOTTA-PARETO INDEX

Alessandra Adduci, Edoardo Latessa, Luca Muzzioli

Abstract. The last two decades have seen remarkable studies in early child development, an interdisciplinary field that involves psychology, economics, and neuroscience. This critical period in child growth exhibits high brain plasticity, and several environmental factors can shape its development by altering gene expression patterns through epigenetic mechanisms, resulting in lifelong impacts. The UN Agenda 2030 includes the Early Child Development Index (ECDI2030) to monitor the Sustainable Development Goal 4.2.1. Although the ECDI2030 has proven to be a useful tool for collecting child development data and addressing public policies, some inherent technical and methodological difficulties need to be addressed, such as, questionnaire design, response to bias in a cross-cultural country, difficulties in handling outliers, and content and duration of the training, among others. Therefore, our goal is to develop a composite index ECDI(i) that incorporates the three dimensions set by the ECDI2030 (i.e., learning, health, and psychological well-being) by using a different methodology, namely the Mazziotta-Pareto Index. Despite the preliminary nature of the results, interesting findings seem to emerge from this study: a positive strong linear correlation between the ECDI(i) and ECDI2030 and a change in ranking is observed.

1. Introduction

Early childhood is defined as the period from prenatal development to the eighth year of life. During this stage, children undergo significant cognitive, social, emotional, and physical changes that shape their future well-being. In the last decade, several studies, such as the Lancet's Series on child development (Grantham-Mcgregor et al. 2007; Walker et al, 2011; Black et al. 2017) demonstrated that the early years represent a crucial window of opportunity, establishing the foundation for lifelong learning, behavior, and health outcomes. It has been highlighted the importance of nurturing care to reach children's full potential, the burden cost of inaction for both individuals and countries, and the role of multi-sector interventions and government leadership. This paper is structured as follows: Section 2

investigates the topic; Section 3 outlines a conceptual framework with the dimensions selected and the methodology adopted and the dataset; Section 4 presents the main results; in the last section, the implications of the research findings are outlined.

1.1 The importance of the Early Child Development.

During early postnatal life, the brain exhibits high plasticity. The developing brain undergoes rapid growth, making it highly responsive to environmental signals which have a profound impact on shaping the neural circuits. The Science of Early Child Development (ECD) shows that epigenetic mechanisms, which alter the activity of genes without changing the order of their DNA sequence, play a key role in mediating the interaction between genes and the environment during early life. This causes a long-lasting change in gene expression underpinning brain functions. (Murgatroyd and Spengler, 2011).

Consequently, child's early interaction with the surrounding environment and responsive caregivers are considered essential for shaping brain architecture and promoting its development. According to Walker et al. (2011), the significant risk factors that hinder children from reaching their full potential are: inadequate cognitive stimulation, stunting, and prenatal maternal nutrition. The research also detects protective factors, such as breastfeeding and maternal education. The consequences of a poor start in life extend beyond the individual, impacting society as a whole. Investments in early childhood development are more cost-effective than remediation and produces greatest returns in human capital (Heckman, 2011). Therefore, understanding and investing in early childhood development is essential to promote cognitive, social, emotional, and physical development as well as to reduce systemic poverty and inequalities (Shonkoff and Phillips, 2000).

1.2 Tools to Measure ECD.

Measuring ECD presents several difficulties due to its multidimensional nature. Comprehensive assessments of ECD typically require highly trained professionals and significant administration time, making them unsuitable for large-scale population monitoring. To capture information about children's achievements, UNICEF, in collaboration with a technical advisory group, developed the ECDI, a 10-item index. In 2009, it was added to the Multiple Indicator Cluster Surveys (MICS) and it has been used in over 70 countries.

The ECDI aims to measure the overall developmental status of children within the physical, literacy-numeracy, social-emotional, and learning domains. It consists of specific questions for mothers/caregivers about their children's development. Subsequently, UNICEF developed a new methodology involving consultations with experts, partner agencies, and national statistical authorities. When early childhood development became part of the Sustainable Development Goals (SDGs) of the Agenda 2030, SDG indicator 4.2.1 was chosen to monitor the improvements towards this target. An updated version of the ECDI was implemented in response to the requirements of SDGs monitoring, namely ECDI2030. The ECDI2030 captures the achievement of key developmental milestones by children aged 24 to 59 months.

The index covers 12 sub-domains under three domains of ECD including health, learning, and psychosocial well-being. The index includes 20 questions for mothers or primary caregivers to assess children's behavior, skills, and knowledge in everyday situations. ECDI2030 is designed to assess a child's overall level of development across three dimensions: health, learning, and psychosocial well-being. Unlike the MICS, the ECDI2030 was specifically designed and validated to generate estimates for reporting on SDG indicator 4.2.1.

Additionally, the ECDI2030 provides broader and more comprehensive content coverage, including a larger number of developmental sub-domains that enable a more comprehensive and accurate assessment (UNICEF Technical manual, 2023)**.**

All things considered, this study aims to promote a new early childhood composite index ECDI(i) for tracking children's development at the global level, by investigating inputs that affect it. The expected index should incorporate key indicators able to capture children's capabilities in three main dimensions: health, learning, and psychological well-being.

2. Methodology

The development of a new composite indicator follows four stages:

(1) Definition of the phenomena
(2) Selection of individual indicators
(3) Standardization
(4) Aggregation

According to the relevant literature (Murgatroyd and Spengler, 2011), "*Early Child Development*" could be defined along three different dimensions: *Health*, *Psychological Well-Being*, and *Learning*. The composition of these dimensions provides a multidimensional definition of the ECD.

The reasons behind the elaboration of a new composite indicator are grounded in the necessity to construct a tool capable of measuring the inputs required for a

healthy and safe child development. The current ECDI2030 devised by UNICEF is limited to assessing whether children are on track for the three dimensions while the need to measure inputs is paramount as it makes possible an early identification of developmental challenges.

Moreover, the measure of those dimensions has other upsides such as providing a comprehensive assessment of the environment and living basic conditions of children, the possibility of being implemented remotely using available data and, lastly, it can foster targeted interventions.

2.1 Computation of the new ECDI(i)

The conceptual framework that guided the selection of dimensions included in the new index includes:

A. The capability approach (Sen, 1999), a suitable foundation for analyzing the multiple factors that influence capabilities and human well-being by considering the conversion factors.
B. The WHO guidelines provide recommendations to caregivers, health professionals, policymakers, and stakeholders for identifying areas of concern and strengthening policies to better address ECD (WHO guideline, 2020).

This composite index was developed by collecting data from the UNICEF warehouse and was tested on 28 countries selected in accordance with available data. Domains and sub-domains keep track of the three dimensions identified in the ECDI2030, indeed the final indicator is the result of the composition of three different indicators: (**H**) *Health*; (**P**) *Psychological Well-Being*; (**L**) *Learning*.

The indicators **H** and **P** were the results of the aggregation of different composite indicators. Once every individual indicator was collected, the composite indicators were built by using the **MPI** methodology. In the case of both **H** and **P** the process of standardization occurred only in the first stage of aggregation whereas in the subsequent composition steps the **Z**-matrix is already given and does not need to be computed once again.

Table 1 - *Health dimension indicators.*

Indicator	Domain	Subdomain	Individual Indicators
Health	***Maternal/Child Health***	*M&C Risk Factor*	*- Prevalence of Anaemia in pregnant Women*
			- (%) Preterm births
		Breastfeeding	*- Exclusive breastfeeding 0-5 months*
			- Continued Breastfeeding 12-23months
	Child Nutrition	*Child Food Poverty*	*- Severe child food poverty*
			- Moderate child food poverty
			- Minimum dietary diversity
			- Minimum Meal Frequency
		Child Malnutrition	*- Wasting: 0–59 months*
			- Overweight 0–59 months
			- Stunting:0–59 months
	Child Mortality	*Childhood deaths*	*- Under-5 mortality rate*
			- Infant Mortality rate
		Immunization preventable disease	*- DTP3: diphtheria, pertussis and tetanus vax*
			- Polio3: 3 doses of the polio vax
			- MCV1: 2 dose of measles vax
	Security and Safety	*Pregnant Status*	*- Antenatal care (at least one visit)*
			- Institutional delivery
		Access to Service	*- Sanitation services*
			- Drinking-water services (%)

Table 2 – *Psychological Well-Being dimension indicators.*

Indicator	Domain	Subdomain	Individual Indicators
Psychological Well-Being	**Responsive Caregiving**	*Environment and Caregiver*	*- (%) children's book*
			- (%) children's toys
			- Stimulation by parents
			- Children with inadequate supervision
			- Violent Discipline
		Care-seeking	*- Careseeking for diarrhea*
			- Careseeking for respiratory infection

Table 3 – *Learning dimension indicators.*

Indicator	Domain	Individual Indicators
Learning	**Education**	*- Adjusted net attendance rate (ANAR)*
		- Attendance in early childhood education
		- Positive Discipline 1-4

2.2 Mazziotta-Pareto Index

To aggregate individual indicators into composite indicators we used the Mazziotta-Pareto index (MPI). The MPI is widely used to calculate multidimensional phenomena. Given the original matrix $\mathbf{X=\{x_{ij}\}}$ with *n* rows and *m* columns where (Mazziotta and Pareto 2020):

$$\mathrm{M_{xj}} = \frac{\sum_{i=1}^{n} \mathrm{x_{ij}}}{\mathrm{n}} \qquad \mathrm{Sx_j} = \sqrt{\frac{\sum_{i=1}^{n} (\mathrm{X_{ij}} - \mathrm{M_j})^2}{\mathrm{n}}} \tag{1}$$

The matrix $\mathbf{Z=\{z_{ij}\}}$ is composed:

$$\mathrm{z_y} = 100 \pm \frac{(\mathrm{x_y} - \mathrm{M_{xj}})}{\mathrm{S_{xj}}} 10 \tag{2}$$

where x_{ij} is the value of the indicator *j* for the unit *i* while the values of **M** and **S** are set to a defined value of 100 and 10 respectively and the polarity of the ± sign depends on whether the phenomena is positive or negative. Given the matrix $\mathbf{Z=}\{\mathrm{z_{ij}}\}$ the vector $\mathbf{CV=}\{\mathrm{cv_i}\}$ is computed where:

$$\mathrm{cv_i} = \frac{S_{zj}}{M_{zj}} \tag{3}$$

And:

$$\mathrm{M_z} = \frac{\sum_{j=1}^{m} \mathrm{z_{ij}}}{\mathrm{m}} \qquad S_{zj} = \sqrt{\frac{\sum_{j=1}^{m} (\mathrm{z_{ij}} - \mathrm{M_{zj}})^2}{\mathrm{m}}} \tag{4}$$

The composite index is then obtained using the following formula:

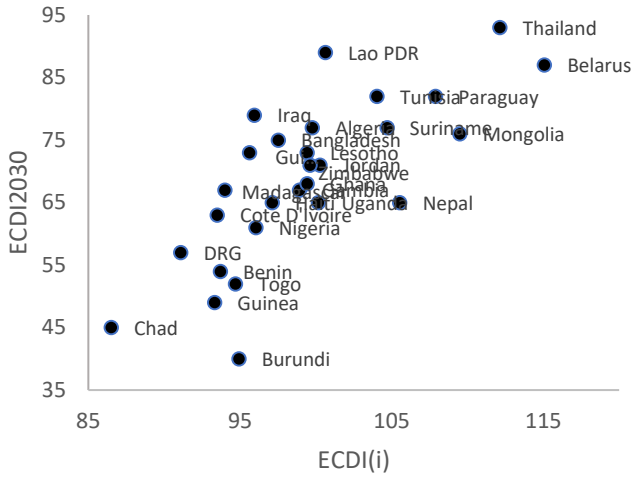
$$MPcv_i = M_{zi}(1 - cvi^2) = M_{zi} - S_{zi}\, cv_i \tag{5}$$

The arithmetic mean of the standardized indicators is corrected by subtracting a quantity (the product of $S_{zi}\, cv_i$) proportional to the standard deviation (Mazziotta and Pareto 2017). In this way units with similar standardized values are less penalized.

3. Results

To understand the effectiveness of ECDI(i) we compared it with ECDI2030, thus performing a correlation analysis. As Figure 1 shows, we found a positive correlation between the two indices.

Figure 1 – *Scatterplot between ECDI(i) and ECDI2030.*



The value of the Pearson Coefficient (r) is equal to 0.705 suggesting a substantial agreement between the two measures. Furthermore, a small p-value (< 0.001) indicates that the correlation is statistically significant. These findings reinforce the reliability of the ECDI(i) as a tool to assess early childhood development. Although the two indices are not interchangeable, this result suggests that the ECDI(i) is capable of assessing and measuring whether the inputs (such as endowments, opportunities, and planned interventions) have reached a satisfactory level for the child's development. Therefore, the ECDI(i) might be a useful complementary tool that makes it possible to draw predictions about eventual outcomes in children's development. Despite their different approaches, it can be inferred that the high correlation between ECDI(i) and ECDI2030, (one based on inputs and the other on outcomes), may be linked to the following explanation:

(a) Since ECD is a multidimensional process where inputs influence outcomes, when countries prioritize inputs, this is more likely to positively impact the overall outcomes of children's development. As suggested by the WHO, equitable access to high-impact and quality health and education services (an

input) may support child cognitive development (an outcome) (WHO guideline, 2020). In fact, countries with high MPI for each of the three dimensions are at the top of the ranking for both ECDI2030 and ECDI(i) (Table 4) (e.g., Belarus: ranking position n°1 and n°3 in ECDI(i) and ECDI203, respectively)

(b) Both ECDI(i) and ECDI2030 assess ECD based on the same framework (e.g., Health, Psychological well-being and Learning) resulting in capturing similar aspects of the ECD multi-faceted phenomenon.

Furthermore, country's ranking is shown in Table 4 by using both ECDI(i) and ECDI2030, where it can be noticed a slight but significant change in the rank positions. Ranking is a key point to compare children's living conditions between countries. Furthermore, the analysis of the index's subdimensions may provide valuable insights for policymakers to employ targeting strategies and intervention programs.

The result of this work also depends on the fact that countries with the highest scores coincide between the two indicators, as in the case of Belarus, Thailand, and Paraguay. The only exception is Mongolia which, in ECDI(i), replaces Lao PDR. In any case, the latter still maintains a score above the limit set by the MPI (100.6) for ECDI(i).

The same congruence can be found at the bottom of the ranking, where the Democratic Republic of Congo replaces Burundi, which ranks 25th in the ECDI(i) and scores well below the MPI threshold (94.3). On the other hand, when extremes are excluded, there is a significant difference in the ranking order of countries.

This behavior makes the development of a new indicator highly relevant to establish and determine distinct intervention strategies (and resource allocation) depending on the instrument used. For instance, Uganda scores above the MPI threshold in ECDI(i), while it ranks among the lowest countries in ECDI2030. The inconsistencies observed in the Ranking order may have the following explanations:

(A) Different methodologies used to compute data, result in a different assessment of the ECD phenomenon.

(B) The three dimensions of ECDI(i) do not include data exclusively related to children. The composite indices contain data regarding (i) health and well-being of the mother and (ii) the environmental conditions in which the child is developing (stimulating environment, violent education, and caregivers, among others).

In conclusion, based on these findings and considering that ECD is a complex and multidimensional phenomenon, the authors of this paper believe that a holistic approach is needed to integrate the current framework of developmental conditions by reviewing the input dimensions that causally determine the possibility of healthy child development.

Table 4 – *Rank comparison between ECDI(i) and ECDI2030.*

Country	**ECDI(i)**	Country	**ECDI2030**
Belarus	**115,08**	**Thailand**	**93,00**
Thailand	**112,14**	**Lao PDR**	**89,00**
Mongolia	**109,51**	**Belarus**	**87,00**
Paraguay	107,90	Paraguay	82,00
Nepal	105,54	Tunisia	82,00
Suriname	104,71	Iraq	79,00
Tunisia	104,04	Algeria	77,00
Lao PDR	100,63	Suriname	77,00
Jordan	100,28	Mongolia	76,00
Uganda	100,18	Bangladesh	75,00
Algeria	99,78	Guinea−Bissau	73,00
Zimbabwe	99,62	Lesotho	73,00
Lesotho	99,43	Jordan	71,00
Ghana	99,43	Zimbabwe	71,00
Gambia	98,93	Ghana	68,00
Bangladesh	97,53	Gambia	67,00
Haiti	97,13	Madagascar	67,00
Nigeria	96,05	Haiti	65,00
Iraq	95,96	Nepal	65,00
Guinea−Bissau	95,63	Uganda	65,00
Burundi	94,93	Cote D'Ivoire	63,00
Togo	94,71	Nigeria	61,00
Madagascar	94,00	DRG	57,00
Benin	93,71	Benin	54,00
Cote D'Ivoire	93,51	Togo	52,00
Guinea	**93,34**	**Guinea**	**49,00**
DRG	**91,10**	**Chad**	**45,00**
Chad	**86,52**	**Burundi**	**40,00**

4. Conclusion

A comprehensive monitoring framework for ECD should include indicators that measure both inputs and outcomes. Factors such as children's nutritional status, access to early learning opportunities, and exposure to responsive caregiving are important inputs that affect early child development. These inputs have an influence on developmental outcomes among children. The work of this research allows us to conclude that:

I. There is linear and positive correlation between ECDI(i) and Unicef's ECDI.
II. The indicator considers inputs needed for child development, although there is a need to collect more data and individual indicators.
III. A significant change in ranking is necessary to understand the degree of early child development and possibly target research, studies, and operations.

However, given the preliminary nature of these results, further development is required and necessary. Especially more data is needed for existing indicators, and possibly, the development of new indicators capable of capturing other essential aspects of child development. Indeed, along the process of data collection, a difficulty in collecting data was enconutered for many relevant indicators. In developing the ECDI(i), a recurring problem concerns the lack of data for many countries.

This problem forced us to reduce the sample of countries and the number of individual indicators. In fact, as highlighted in Table 5, some essential indicators that should be included (Shonkoff and Phillips, 2000; WHO, 2020) to build a more accurate tool have not been included due to a lack of available data.

However, the preliminary nature of the results emphasizes the need for further development and data collection, as existing indicators may require more data and new indicators to capture additional crucial aspects of child development. Many countries lack data for relevant indicators, limiting a more robust and comprehensive analysis of ECD.

Table 5– *Not included indicators due to lack of available data.*

Health	Psychological Well-Being	Learning
-Maternal consumption of Iron and folate	Maternal mental health	-Child (0-5) literacy skills
-Micronutrient deficit	-Responsive care by father	-Child (0-5) numeracy skills
-ANC and IPT3 coverage for pregnant women	-Labor force participation rate	-Mother's literacy rate

References

BLACK M.M, WALKER S.P., FERNALD L.C., ANDERSEN C.T., DIGIROLAMO A.M., LU C., MCCOY D.C., FINK G., SHAWAR Y.R., SHIFFMAN J., DEVERCELLI A.E., WODON Q.T., VARGAS-BARÓN E., GRANTHAM-MCGREGOR S. 2017. Early childhood Development Coming of Age: Science through the Life Course, *Lancet,* Vol. 389, No. 10064, pp.77-90.

GRANTHAM-MCGREGOR S., CHEUNG Y. B., CUETO S., GLEWWE P., RICHTER L., STRUPP B. 2007. Developmental potential in the first 5 years for children in developing countries, *Lancet* Vol. 369 (9555), pp.60-70.

HECKMAN J.J. 2011. The Economics of Inequality: The Value of Early Childhood Education, *American Educator*, Vol. 35, No. 1, pp.31-35.

MAZZIOTTA M., PARETO A. 2017. Synthesis of Indicators: The Composite Indicators Approach. In Maggino, F. (Ed.) *Complexity in Society: From Indicators Construction to their Synthesis*, Social Indicators Research Series, Vol 70, Springer Cham.

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

MURGATROYD C., SPENGLER D. 2011. Epigenetics of Early Child Development, *Frontiers in Psychiatry*, Vol. 2, No. 16.

SEN A. 1999. *Development as freedom*. Oxford: University Press.

SHONKOFF J. P., PHILLIPS D. A. 2000. From Neurons to Neighborhoods. *The Science of Early Childhood Development,* National Academies Press.

UNICEF. 2023. The Early Childhood Development Index 2030: A New Measure of Early Childhood Development. *UNICEF Technical manual*.

WALKER S.P., WACHS T.D, GRANTHAM-MCGREGOR S., BLACK M.M., NELSON C.A., HUFFMAN S.L., BAKER-HENNINGHAM H., CHANG S.M., HAMADANI J.D., LOZOFF B., GARDNER J.M., POWELL C., RAHMAN A., RICHTER L. 2011. Inequality In Early Childhood: Risk and Protective Factors for Early Child Development, *Lancet,* Vol. 378, No. 9799, pp. 1325-38.

WHO. 2020. Improving Early Childhood Development. *WHO Guideline*.

Alessandra ADDUCI, Rome3 University, ale.adduci@stud.uniroma3.it
Edoardo LATESSA, Rome3 University, edo.latessa@stud.uniroma3.it
Luca MUZZIOLI, Sapienza University, luca.muzzioli@uniroma1.it

SINGLE-PARENT FAMILIES AND ADOLESCENTS' WELLBEING IN EUROPE: A MULTILEVEL ANALYSIS

Andrea Ballerini, Raffaele Guetto

Abstract. This study, employing multilevel modeling, investigates adolescent subjective well-being (SWB) in single-parent families (SPFs) across Europe. It uncovers a consistent negative impact, primarily due to economic challenges, with greater penalties in regions where SPFs are more prevalent. Public family spending offers partial relief, underscoring the need for targeted interventions and comprehensive social policies to enhance adolescent SWB in single-parent families (SPFs).

1. Introduction

Household composition has evolved since the 1960s in most Western countries. The idea of family as two biological parents with children has given way to various structures due to factors as changing living arrangements and the increase in out-of-wedlock births and family instability. Accordingly, recent research has focused on how children who grow up in SPF fare compared to children raised in two-parent families (TPFs). Previous studies consistently indicate lower SWB among children living with only one parent. However, many open questions and knowledge gaps remain, as most studies focus on countries with an earlier diffusion of new family forms. Moreover, the literature provides limited and sometimes conflicting evidence regarding how the effects of family structure on children's outcomes vary based on societal characteristics (Amato, 2000; Härkönen et al., 2017). Finally, the mechanisms underlying the association between living in a SPF and adolescents' SWB are often unexplored. The aim of this paper is to fill these gaps in the literature by analysing whether and how the SWB of adolescents aged 15-19 years is associated with the structure of their family. We compare adolescents living in a SPF with those living in a TPF, by applying multilevel models with adolescents nested in countries and 3-year periods to data concerning 14 European countries. The paper tests possible cross-country differences in the association and the moderating role played by societal factors such as the diffusion of separations and family policies. Additionally, it explores possible mechanisms related to economic and relational

dimensions. Next paragraph provides a brief literature review, we will then present the data and methods used in the study, followed by empirical findings. Finally, there will be a discussion on the implications of findings.

2. Single-parent families and adolescents' subjective well-being

2.1. The prevalence of the single-parent families

From the late '60s, western countries experienced a decline of marriage and an increase in divorce, as part of the "Second Demographic Transition" (Lesthaeghe, 2020), a process which includes delayed and sub-replacement fertility, increases in cohabitation and non-marital childbearing. As a result, an increasing number of children experiences parental separation or is raised by a single parent (SP).

In the US, the percentage of SPF increased from 10% (1965) to almost 30% (early 2000s). Most European countries have seen an increase in the prevalence of SPFs as well (Maldonado and Nieuwenhuis, 2018). However, the spread has not been homogeneous across countries in terms of timing and intensity. Late-comer countries such as Italy and Spain are now catching up with early-comer countries such as Finland and Sweden, where the trend has instead stabilized. Several factors have a significant association with the risk of divorce. For instance, research has shown that in countries with greater social, economic, and legal barriers to divorce, highly educated couples are more likely to break up. Conversely, where these barriers are reduced, the gradient is reversed (Matysiak *et al.*, 2014). Findings highlight a complex interplay between socioeconomic factors and divorce, and so on the spread of SPFs. However, with exceptions of Greece, Italy and to some extent Spain, today single motherhood is more common among low-educated (Härkönen, 2018).

2.2. Children and adolescents' subjective well-being

Well-being (WB) is a strong indicator for a well performing society. It is essential for children and adolescents to thrive in all areas of their lives, including academics, social relationships, emotional resilience and physical health. Family processes play a critical role in child and adolescent WB, as families can create family environments that support child and adolescent WB (Buehler, 2020). To measure it, the best known partition is subjective and objective measures, but we can find in the literature: self-report, used to assess all five domains of child WB (physical, psychological,

cognitive, social, and economic); objective; observational, to assess social-cognitive skills; psychophysiological, for stress levels, coping skills, and emotional regulation.

The best approach to measure child WB is to use a combination of different methods. When it is not possible, self-report measures may be most appropriate for older children and adolescents, while observational measures may be more appropriate for younger children (Pollard and Lee, 2003; Tsang *et al.*, 2012).

2.3. The relationship between single-parent families and children's subjective well-being: moderating factors at the macro level

At the macro level, we consider family policies and the diffusion of SPFs as factors that may influence SPF-SWB association for children. Studies found a positive impact of policies that enhance income and employment opportunities of SPs, leading to reduced risks of poverty and fewer familial problems (Biegert *et al.*, 2022; Aerts *et al.*, 2022). However, the effects of these policies vary across contexts. Also, while policies targeting all families can improve overall income and life satisfaction, specific policies are required to bridge the gap between SPF and TPF, as between poor and non-poor families (Gornick *et al.*, 2022).

Regarding the role played by the prevalence of SPFs, sociologists and demographers have hypothesized that a process of 'normalization' and greater social acceptance of new family forms, already seen for other Second Demographic Transition-related behaviours, might occur (Härkönen et al., 2017). This normalization process could potentially reduce stress levels for parents and children and improve post-separation parent-child relationships, also due to legal reforms regarding joint custody in post-divorce arrangements. Accordingly, it has been hypothesized that the negative association between non-intact family structures and child WB decreases with the spread of SPFs. In practice, although, studies have found that children belonging to countries and cohorts with higher rates of divorce seem to suffer an even greater divorce penalty (Kalmijn and Leopold, 2021). This apparently counterintuitive result may be explained by selection effects. In situations where societal barriers to divorce are substantial, only families characterized by exceedingly high levels of parental conflict choose to pursue divorce. Under such circumstances, the separation of parents may be advantageous for the children involved. However, as divorces become more prevalent, families with relatively lower levels of conflict also decide to divorce. In these instances, the adverse consequences of parental separation are not partially offset by the advantages of reducing parental conflict (Guetto *et al.*, 2022). Furthermore, due to the reversal from positive to negative of the educational gradient of divorce, the increased prevalence of SPF is mainly caused by separations among already socially disadvantaged

couples (Kalmijn and Leopold, 2021). Young adults face unique challenges, investing in education and employment, promoting healthy lifestyles, access to healthcare, and addressing social and economic determinants of health is important.

2.4. The micro-mechanisms underlying the relationship between single-parent families and children's subjective well-being.

On average, the SWB of children living with a SP is usually found to be significantly lower than that of children living with both parents (Härkönen et. Al, 2017). This finding is consistent across studies, although the underlying causes may vary. Influential papers have identified three distinct mechanisms. First, *emotional stress and lower satisfaction with family relations* may play a crucial role, as adults and children often struggle with troubled relationships following separation and divorce. The reduced time and support provided to children by the non-custodial parent as well as the higher risk of family conflict can lead to self-esteem and social relationship problems (Amato, 2000). Children who lived their parents' divorce at a young age are particularly vulnerable (Harkonen *et al.*, 2017), and the complexity of the family may also contribute to lower SWB (Meggiolaro and Ongaro, 2014). Second, the *economic difficulties* of SPFs. A consistently higher risk of poverty for children in SPF has been found across all Europe (Maldonado and Nieuwenhuis, 2018). Some countries show an increasing risk, others (e.g. Ireland and Netherlands, which had a notable growth in SP employment) follow different trends. Lastly, lone parents face greater job challenges, often working lower paid and less stable jobs (Maldonado and Nieuwenhuis, 2018). Finally, there are *social consequences* both on parents and children. Job difficulties impact social and personal fulfilment too. Balancing family-household needs can be particularly challenging for SPs with insufficient policies. Lone parents are also more vulnerable to stress-related issues as alcohol abuse (Avison and Davies, 2005), that lead to depression and lower living standards (Amato, 2000), affecting the quality of parenting and so children's SWB.

3. Data and Methods

3.1. Data sources and variables

This study utilizes data from the European Social Survey (ESS), the European Union Labour Force Survey (EU-LFS), and the dataset provided by the Organization for Economic Cooperation and Development (OECD). The ESS data allowed us to

analyze an important part of what we wanted to test and at the same time provided us with a large sample to estimate our multilevel models. Of course, there is a price to be paid and for some variables we had to agree to use the 'closest' variable available. For the SWB in the future we would also like to implement a measure. We utilized all waves of the ESS, gathering data from 2002 to 2022 for 14 countries. The choice was based on two criteria: the first is the availability of data at the macro level, the second is the presence of (also) SPF in (almost) all groups. Although the multilevel model still allows us to include the other nations as well, we preferred a conservative choice. Lastly, our analytical sample consists of 11,045 adolescents aged 15 to 19, nested in country-period groups (Table 1).

Table 1 – *Distribution of adolescents 15-19 by type of family, country, and 3-year periods.*

Years group	02-'04		05-'07		08-'10		11-'13		14-'16		17-'19		20-'22	
N. parents	**2p**	**1p**	**2p**	**1p**	**2p**	**1p**	**2p**	**1p**	**2p**	**1p**	**2p**	**1p**	**2p**	**1p**
Belgium	174	41	115	31	149	43	149	41	185	52	87	31	ND	ND
Denmark	115	29	49	9	127	30	112	26	72	20	71	20	ND	ND
Estonia	85	29	112	32	62	24	80	36	115	41	78	14	15	3
Finland	226	46	84	24	185	52	90	20	122	36	74	13	13	9
France	47	19	79	25	122	47	53	25	103	30	62	25	45	11
Hungary	87	26	107	21	118	29	71	25	51	17	101	25	60	15
Germany	229	71	129	33	194	43	166	51	228	59	106	34	ND	ND
Ireland	75	13	116	22	44	14	156	45	73	20	75	17	ND	ND
Italy	40	7	ND	ND	ND	ND	26	5	ND	ND	253	42	ND	ND
Netherlands	122	28	40	12	84	20	48	17	99	36	93	24	4	0
Poland	319	50	137	23	220	52	109	22	147	21	79	13	ND	ND
Portugal	73	12	100	25	80	25	90	43	45	20	45	14	12	3
Spain	138	29	94	16	116	17	141	23	77	12	114	30	8	4
Sweden	163	58	90	24	137	45	107	29	114	31	41	11	ND	ND

ND: Data not (or not yet) provided by European Social Survey

The ESS provided individual-level variables, including the dependent variable, assessing life satisfaction on a scale from 0 to 10; the independent variable, a dummy indicating whether the adolescent belongs to a TPF or a SPF; control variables as age, country of birth, sex, number of persons in the hh, and parental education (up to lower secondary education (LSE), completed LSE, over LSE); intervening variables to examine the possible underlying mechanisms, as satisfaction with the household income, frequency of meetings with friends and availability of confidant for private talks (we assumed a link with the emotive issues) in order to investigate the three mechanisms highlighted in Chapter 2; average religion level in the group to account for values at macro level. EU-LFS and OECD data provided respectively percentage of SPF in the country-period group and percentage of GDP spent on family transfers.

3.2. Methods

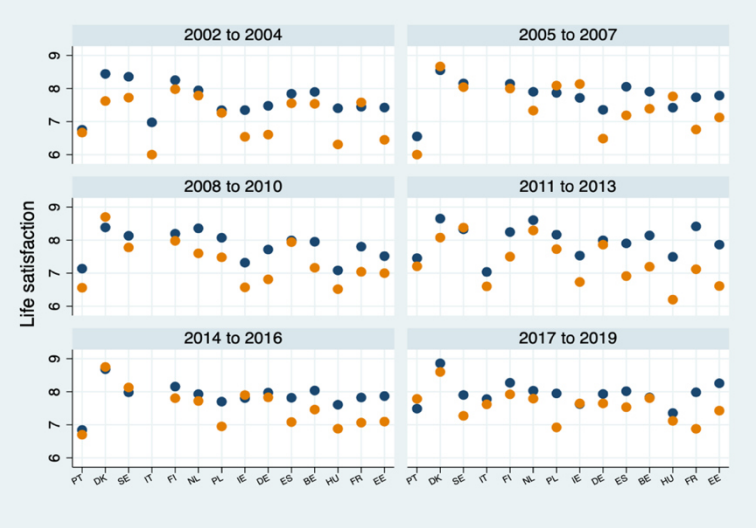
We used a multilevel modeling approach with adolescents nested within country-period groups and country fixed effects to better capture variance composition. Initially, an ANOVA was performed for the distribution of variance in dependent variable, "life satisfaction," across different levels. Then, a multilevel linear regression analysis was conducted to explore the relationship between independent variable "single-parent family" and life satisfaction, and it confirmed the necessity of utilizing a multilevel modeling. To improve the understanding of the underlying mechanisms and contextual factors, a systematic stepwise approach was employed. Control variables were introduced to account for potential confounding factors and isolate the net total effect of being in a SPF. Additionally, macro-level variables and cross-level interactions were included to examine the role of contextual factors and investigate whether the effect of living with one parent is moderated by family policies or the diffusion of SPFs. Furthermore, intervening variables were included to assess mediating mechanisms and interactions with those were lastly checked to find how the effects of SPF may vary under different conditions.

4. Results

4.1. Descriptive analysis

Descriptively, the analysis highlights a negative association between living in a single-parent family and the SWB of adolescents in Europe. This finding holds for most of the country-period groups, indicating a consistent pattern. However, the intensity of the association is highly heterogeneous (Figure 1), justifying the use of a multilevel modelling approach. The descriptive analysis also sheds light on the prevalence of SPFs in Europe. This phenomenon is increasing in almost all countries, emphasizing the need to understand the complex relationship between its prevalence and the strength of the association with adolescents' WB. Descriptive findings highlight how environments characterized by a higher prevalence of SPFs exhibit a more pronounced negative association with adolescents' SWB, contrary to the "normalization" hypothesis.

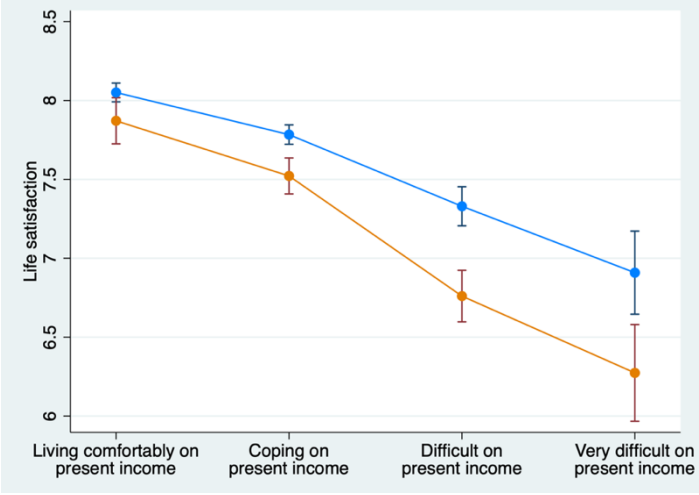
Figure 1 – *Life satisfaction of adolescents by 3-year period and country.*



Note: *Single-parent families (orange) and two-parent families (blue)*

4.2. Multivariable and multilevel results

Figure 2 – *Life satisfaction of adolescents at different level of satisfaction for hh. Income.*



Note: *Single-parent families (orange) and two-parent families (blue)*

The multilevel analysis, including control variables, provides robust evidence for the negative association between living in a SPF and the SWB of adolescents in Europe. The coefficient remains highly significant even after controlling for demographic and socio-economic factors (Table 2), as well as macro-level variables.

The observed association is primarily driven by SPFs with economic difficulties, with a gap in SWB of 0.68 for families experiencing major difficulties, compared to 0.17 for families with an income that teenagers consider satisfactory (Figure 2). The analysis thus reveals that adolescents' feeling about household income has both a mediating and a moderating role. SPFs often encounter greater economic difficulties, which can significantly impact the overall WB of their children.

Table 2 – *Variables coefficient on Life Satisfaction and SE Estimates.*

			Bivariate reg.	+ Controls	+ Macro level	+ Intervening
Single-parentfamily	No (ref.)	-	-	-	-	-
	Yes	Coef.	-0.460^{***}	-0.436^{***}	-0.438^{***}	-0.294^{***}
SPFs diffusion		Coef.	-	-	0.028^{***}	0.027^{***}
Public spending		Coef.	-	-	0.045	0.037
Average religiosity		Coef.	-	-	-0.061	-0.050
Feeling for hh income	Comfort.(ref.)	-	-	-	-	-
	Coping	Coef.	-	-	-	-0.276^{***}
	Difficult	Coef.	-	-	-	-0.842^{***}
	V. difficult	Coef.	-	-	-	-1.305^{***}
Meet usually friends	No (ref.)		-	-	-	-
	Yes	Coef.	-	-	-	0.341^{***}
Confidants	No (ref.)		-	-	-	-
	Yes	Coef.	-	-	-	0.659^{***}
Constant		Coef.	8.166^{***}	9.992^{***}	9.904^{***}	8.892^{***}
Var (Constant)			0.025	0.025	0.018	0.012
Var (SPF)			0.033	0.026	0.031	0.023
N			11,405	11,405	11,405	11,405

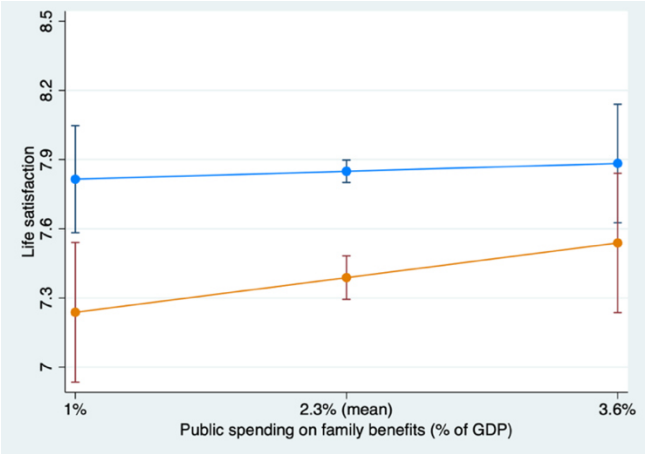
Note: ***: p<.01

4.3. How public spending on family allowances and the spread of single-parent families influence the association.

There is a slight reduction in the gap in life satisfaction between adolescents in SPF and those in two-parent ones (Figure 3). Despite the estimation uncertainty, this finding, in line with previous studies, gives important insights in favour of policies.

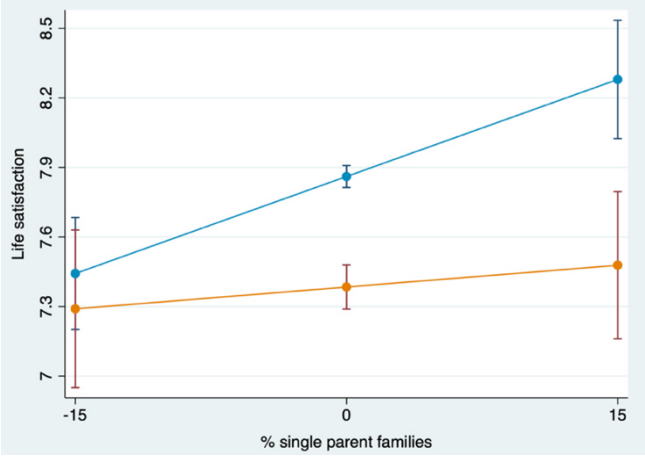
Turning to the spread of SPF, the results of the multilevel analysis align with the descriptive outcomes: countries with a higher prevalence of SPFs exhibit a significant larger disparity in life satisfaction between adolescents living in SPFs and those living in TPFs (Figure 4).

Figure 3 – *Life satisfaction of adolescents at different level of single parent diffusion.*



Note: *Single-parent families (orange) and two-parent families (blue)*

Figure 4 – *Life satisfaction of adolescents at different level of single parent diffusion.*



Note: *Single-parent families (orange) and Two-parent families (blue). % SPF centred on mean.*

5. Discussion

This study confirms the negative association between living in a single-parent family and the subjective well-being (SWB) of adolescents in Europe. This association remains robust even after controlling for various demographic and socio-economic factors. Notably, regions with a higher prevalence of single-parent families (SPFs) exhibit a stronger negative association with adolescent SWB.

Our findings make us agree with previous studies that found the primary driver of the negative association between living in a SPF and SWB in economic challenges faced by SPFs, more than the social or emotional side. The level of public spending on family benefits appears to mitigate the negative impact of economic hardship on adolescent WB. However, there is marked variation, and it is not entirely clear if this type of policy can fully bridge the gap between SPFs and TPFs. This suggests the need to incorporate complementary strategies or interventions to address the specific challenges faced by economically disadvantaged SPFs. We also attempted to incorporate spending on education to explore how investments for children and adolescents may influence this association. While our findings mirrored the strong correlation with family spending, this aspect warrants separate, in-depth analysis.

Our findings regarding the prevalence of SPFs align with arguments related to selection effects, reflecting the changing intensity of pre-separation parental conflicts and the evolving socio-economic composition of SPFs. The higher prevalence of SPF families in already disadvantaged contexts may reflect the challenges faced by individuals, including lower income levels, restricted access to resources, and limited social support networks. Consequently, adolescents growing up in these environments may encounter greater adversity and experience lower SWB compared to their counterparts in more traditional family structures. These results underscore the need for further research to unravel the underlying mechanisms and explore other potential factors that may influence the association between SPFs and adolescent SWB. The complexity of this association emphasizes the necessity of a nuanced understanding of the multiple contributing factors.

Our results align with the notion that a comprehensive approach to social policy is required to adequately support SPFs. This underscores the importance of targeted interventions that extend beyond economic support to address the specific challenges faced by these families. It calls for a re-evaluation and refinement of existing policies to ensure that they effectively cater to the specific needs of SPFs.

In conclusion, this study underscores the importance of considering family structure and economic factors in understanding the SWB of adolescents in SPFs. Addressing the challenges faced by these families can significantly contribute to enhancing the overall WB and outcomes of adolescents in Europe.

6. Implications and Future research

First, additional information is needed to differentiate between various family types, such as stepfamilies, divorced families, widowed parents, and families with a lifelong single parent. Unfortunately, due to data limitations, we could not fully explore these aspects, but recognizing the significance of these structures and their potential impact on children's outcomes would enhance future research. Second, including more country-level units is desirable. Some countries had limited units and were excluded from the analysis, but expanding the dataset to cover a wider range of countries would improve generalizability and provide a better understanding of contextual factors. It's also important to acknowledge that using frequency of meetings as a proxy for the social component and availability of confidants as a measure of emotional stress may have limitations. Unfortunately, data constraints prevented us from a full exploration of these issues. New data could allow us to investigate the influence of parents' job on children's WB too.

Future research could employ alternative measures for more accurate findings and deeper understanding of individual perceptions. Lastly, a causal framework with an expanded set of variables could offer insights for policymakers.

Acknowledgements

We acknowledge co-funding from Next Generation EU, in the context of the National Recovery and Resilience Plan, Investment PE8 – Project Age-It: "Ageing Well in an Ageing Society". This resource was co-financed by the Next Generation EU [DM 1557 11.10.2022]. The views and opinions expressed are only those of the authors and do not necessarily reflect those of the European Union or the European Commission. Neither the European Union nor the European Commission can be held responsible for them.

Special thanks to Professor Daniele Vignoli for his invaluable assistance.

References

AERTS E., MARX I., PAROLIN Z. 2022. Income Support Policies for Single Parents in Europe and the United States: What Works Best? *The ANNALS of the American Academy of Political and Social Science,* Vol. 702, No. 1, pp. 55-76.

AMATO P.R. 2000. The consequences of divorce for adults and children. *Journal of marriage and family*, Vol. 62, No. 4, pp. 1269-1287.

AVISON W.R., DAVIES L. 2005. Family structure, gender, and health in the context of the life course. *The Journals of Gerontology Series B*, Vol. 60, No. 2, pp. S113-S116.

BIEGERT T., BRADY D., HIPP, L. 2022. Cross-National Variation in the Relationship between Welfare Generosity and Single Mother Employment. *The ANNALS of the American Academy of Political and Social Science*, Vol. 702, No. 1, pp. 37-54.

BUEHLER C. 2020. Family processes and children's and adolescents' well-being. Journal of Marriage and Family, Vol. 82, No. 1, pp. 145-174.

GORNICK J.C., MALDONADO L.C., SHEELY A. 2022. Effective Policies for Single-Parent Families and Prospects for Policy Reforms in the United States. The ANNALS of the American Academy of Political and Social Science,Vol.702, No.1

GUETTO R., BERNARDI F., ZANASI F. 2022. Parental education, divorce, and children's educational attainment. *Demographic Research*, Vol. 46, pp. 65-96.

HÄRKÖNEN J. 2018. Single-mother poverty: How much do educational differences in single motherhood matter? *The triple bind of single-parent families.*

HÄRKÖNEN J., BERNARDI F., BOERTIEN D. 2017. Family dynamics and child outcomes. *European Journal of Population*, Vol.33, pp.163-184.

KALMIJN M. LEOPOLD T. 2021. A new look at the separation surge in Europe: Contrasting adult and child perspectives. *American Sociological Review*, Vol. 86, No. 1, pp. 1-34.

LESTHAEGHE R. 2020. The second demographic transition, 1986–2020: sub-replacement fertility and rising cohabitation. *Genus*, Vol. 76, No. 1, pp. 1-38.

MATYSIAK A., STYRC M., VIGNOLI D. 2014. The educational gradient in marital disruption: A meta-analysis of European research findings. *Population Studies*, Vol. 68, No. 2, pp.197-215.

MEGGIOLARO S., ONGARO F. 2014. Family contexts and adolescents' emotional status. *Journal of Youth Studies*, Vol. 17, No. 10, pp. 1306-1329.

NIEUWENHUIS R., MALDONADO L. 2018. *The triple bind of single-parent families: Resources, employment and policies to improve well-being*. Bristol: Policy Press.

POLLARD E. L., LEE P.D. 2003. Child well-being: A systematic review of the literature. *Social indicators research*, Vol. 61, pp. 59-78.

TSANG K.L.V., WONG P.Y.H., LO S.K. 2012. Assessing psychosocial well-being of adolescents, Vol. 38, No. 5, pp. 629-646.

Andrea BALLERINI, University of Florence, andrea.ballerini1@unifi.it
Raffaele GUETTO, University of Florence, raffaele.guetto@unifi.it

FERTILITY INTENTIONS IN ITALY DURING THE COVID-19 PANDEMIC. EVIDENCE FROM THE FAMILYDEMIC SURVEY

Eleonora Miaci, Raffaele Guetto, Daniele Vignoli

Abstract. The Covid-19 pandemic crisis led to sharp changes in social, work and family organisation, which may have had consequences for individuals' reproductive choices. We use data collected between July and September 2021 to examine changes in fertility intentions induced by the pandemic in Italy. Our findings do not reveal a generalized and substantial decline in respondents' fertility intentions, who, in the vast majority of cases, confirmed their pre-pandemic fertility intentions. However, parents with at least one child under 12 emerged as those more likely to experience a decline in their reproductive intentions. Also, our findings reveal how the upheaval caused by the pandemic resulted in different distributions of unpaid work within couples, with different fertility implications: those who moved towards a more unequal distribution reported a stronger reduction in fertility intentions compared to those who benefited from the opportunity to rebalance domestic responsibilities towards a more equal distribution.

1. Context and Aim

In our analysis, we examine whether the pandemic has influenced fertility intentions by considering the impact of the various transformations it generated, including the uncertainties and overturns it has caused. We take, as a starting point, previous studies that have consistently indicated a negative correlation between periods of economic crisis and a decline in fertility intentions (Matysiak *et al.,* 2021). The implications of Covid-19 regarding population health and the economic consequences that the precautionary measures adopted by governments would have caused were the first aspects to capture the interest and commitment of scholars. Subsequently, several authors turned their attention to investigating whether and how the pandemic influenced family dynamics. Significant findings from these studies encompassed various aspects, including a decline in the quality of married life, a less equal division of family roles, a decrease in marriage celebrations, deteriorating living conditions for individuals, and a decrease in fertility intentions (Arpino *et al.,* 2020; Vignoli *et al.,* 2020; Del Boca *et al.,* 2021; Guetto *et al.,* 2021, 2022). With this paper, we contribute to the literature that

focuses on this last aspect by posing the following research question: *How did fertility intentions change in Italy during the Covid-19 pandemic?*

We investigate fertility intentions because they rely upon both the characteristics of individuals (such as age, parental status, level of education, relational networks and economic situation) and are closely linked to social norms on fertility, the country's welfare system (Novelli *et al.,* 2021), and the political and economic climate, which underwent numerous changes during the Covid-19 pandemic. Fertility intentions are considerably more reliable and compromising than fertility desires. Furthermore, questioning respondents about short-term fertility (within a 3-year time frame) increases the concreteness of this prediction. This study makes an additional effort compared to the existing literature on the topic. It overcomes the limitations of only examining the very short-term impact of the pandemic, questioning respondents at a rather late stage of the pandemic and it investigates potential heterogeneity in individual responses regarding shifts in fertility intentions.

2. Gender roles and fertility

In Italy, there has been a consistent decline in the number of births since 2008, reaching a historic low of 393,000 births in 2022, which represents a decrease of 27,084 births compared to 2019 (ISTAT, 2023). Extensive research has focused on exploring the potential determinants of low fertility rates in Italy, highlighting various factors such as shifts in ideology, economic uncertainty, the availability and cost of childcare services, and the delayed financial independence of young individuals (Alderotti *et al.*, 2019; Mencarini and Vignoli, 2018). The distribution of household and caregiving responsibilities within a partnership is considered a crucial indicator of gender equality (Neyer *et al.*, 2013), and the transformation of gender roles has been proposed as a significant factor in understanding the low fertility phenomenon (Esping-Andersen and Billari, 2015). A growing body of literature has emphasized the positive relationship between an equitable division of household and childcare responsibilities among partners and fertility rates (Riederer *et al.*, 2019; Cheng, 2020), including studies specifically relates to the Italian context (Mencarini and Tanturri, 2004; Pinnelli and Fiori, 2008). However, results are not always consistent, and there is considerable variation depending on factors such as women's employment status, number of children, and the level of fathers' involvement. Despite the ongoing changes in societal dynamics, Italian society has not fully embraced the female revolution, as evidenced by lower fertility rates and lower gender equality compared to the European average. Naldini (2015) describes the challenging process of "de-traditionalization" of gender roles in the Italian context, where gender inequalities persist in both work and family spheres,

despite the convergence of life courses and the emergence of more egalitarian couple ideals (Mills *et al.*, 2008).

3. Research Hypotheses

Two possible scenarios can be envisaged regarding the potential impact of the Covid-19 pandemic: Hypothesis 1: *Worsening fertility intentions due to economic uncertainty and increased childcare duties, especially among women.* The economic crisis, and the resulting job and economic uncertainty generated by the Covid-19 pandemic, may have inhibited fertility intentions. We posit that there is a higher vulnerability among those who are already living in a precarious economic and social situation (Cazzola *et al.* 2016) and among those who already had one or more children (Modena and Sabatini, 2012). Previous studies have indicated that women interested to pursue a two-child family model are more subject to economic insecurity (Fiori *et al.*, 2013). However, it is crucial to consider that in a country like Italy, where the total fertility rate was 1.24 children per woman in 2022, it is reasonable to assume that individuals who already have one or more children have already achieved their desired fertility. Furthermore, based on literature indicating that additional care obligations may conflict with fertility plans, we hypothesise that, among respondents with (school-age) children, the closure of schools and the unavailability of care services like babysitting may have led to a worsening of fertility intentions, due to increased childcare duties. This effect is more likely among women, given their greater involvement in informal caregiving for vulnerable individuals and children (Menniti *et al.*, 2015). Overall, the pandemic crisis may have had differential impacts on men's and women's fertility intentions, with women being more burdened by household and caregiving responsibilities during the period under review (Del Boca *et al.*, 2021). Hypothesis 2: *Improvement in fertility intentions due to a more balanced gendered division of unpaid work.* A more equitable division of household and caregiving work is considered a key indicator of gender equality within the family and is also seen as an important factor in shaping fertility intentions and outcomes (Neyer *et al.*, 2013). Couples who have taken advantage of the upheaval caused by the historical moment to establish a more equal distribution of duties, along with improved relationship quality (Vignoli *et al.*, 2022), may have experienced an improvement in their fertility intentions. This may have been facilitated by the need to disrupt established routines and reorganize family dynamics, as well as the transition from conventional workplaces to home settings (Mangiavacchi *et al.*, 2021). The aforementioned scenarios may not necessarily be mutually exclusive. By incorporating insights from the gender revolution framework (Goldscheider *et al.*, 2015) and the multiple equilibrium framework (Esping-Andersen and Billari, 2015), it is plausible that these scenarios can coexist, contingent upon the specific characteristics of the couples. In particular, we assume that education and gender ideology play a decisive role.

Extensive literature has demonstrated that gender egalitarianism is more widespread among younger and more educated men and women (Mills *et al.*, 2008). Therefore, we expect to find a positive impact of higher education levels on pandemic-induced changes in fertility intentions, along with an inverse relationship between education level and respondents' inclination to delay or relinquish their fertility intentions.

4. Data and methods

The study relies on data from the *Familydemic* survey, which is a collaborative international project and a network of researchers (Kurowska *et al.* 2023). The focus of Familydemic is on the immediate and long-term consequences of policy responses to the COVID-19 outbreak for the distribution of paid and unpaid work in couples and their labour market outcomes in countries with diverse welfare regimes (Canada, Germany, Italy, Poland, Sweden and the US). For Italy, data were collected between July and September 2021. The sampling scheme imposed national quotas for age group, gender, education, macro-region of residence, presence of children and age of the youngest child (N = 7,080). Additionally, post-stratification weights were used to adjust for small deviations from the benchmark population statistics[1]. The dependent variable *change in fertility intention* is measured as the difference between respondents' fertility intentions (expressed within a 3-year time frame) at the time of the interview, and the fertility intentions before Covid-19 (January 2020), investigated retrospectively. It is based on the questions: *"**Before Covid-19** did you intend to have one (another) child in the next 3 years?"* and *"Do you **currently** intend to have one (another) child in the next 3 years?"* The survey participants could opt for a number from 1 to 10 where 1 means definitely no and 10 definitely yes. We then recoded the responses by contrasting those whose intentions had improved vs. not changed vs. worsened. The questions on fertility intentions were asked to a specific segment of the sample (N = 4,103): females aged 20-46 and male respondents who either did not have children or whose youngest child was born prior to 2020. Consistent with previous studies individuals who were expecting a child at the time of the interview were excluded from the sample (N = 360). The Familydemic survey, unlike other surveys, gathers data from both women and men. This inclusive approach allowed us to investigate men's childbearing intentions, acknowledging their important role as active participants in the reproductive process (Neyer *et al.*, 2013). The main independent variables we decided to focus on are the *presence and age of children* in the household and the changes in the *division of unpaid work* within the couple during the pandemic. About the first one, we grouped the subjects

[1] For more methodological information on the survey see (Kurowska *et al.* 2023), as well as the homepage of Familydemic at https://familydemic.wnpism.uw.edu.pl/familydemic-survey (last accessed July 8, 2023).

into the following categories: childless; respondents with a youngest child aged 0-5; respondents with a youngest child aged 6-11; respondents with a youngest child aged more than 11. The survey contains two variables that allowed us to evaluate the shifts in unpaid work: one asking respondents about the shift in the couple's division of housework chores (e.g. food purchasing, cooking, cleaning, doing the laundry), the second about the shift in the division of childcare tasks (e.g. physical care, playing/reading, helping with schoolwork, general oversight). The variables take value 0 if both partners spend the same amount of time in unpaid work (either housework or childcare) at the time of the interview compared to the pre-pandemic period; value 1 if both spend less time[2]; value 2 if both spend more time; value 3 if the respondent spends more time (and the partner spends less or equal time); finally, value 4 if the respondent spends less time (and the partner spends more or equal time). We employed two multinomial logistic regressions to investigate the impact of the independent variables on the variation of fertility intentions for both genders. In the first model, which encompassed the overall sample, we examined the effects of the presence and age of children and the change in couples' division of domestic work. Additionally, we included control variables such as the respondent's age (linear and squared functions), level of education (categorized into three levels), employment status (whether the respondent was employed before Covid-19 or not) and respondents' gender ideology. To operationalize the gender ideology of the respondent, we created a variable by combining the answers to two questions related to the roles of mothers and fathers: *"In general, fathers are as well suited to look after their children as mothers"* and *"Mothers should be as responsible for financially supporting their families as fathers."* The possible answers follow a five-point Likert scale about how much respondents agree with these statements. We built an index of gender ideology, ranging from 0 to 9, by summing up the answers to these two questions. In the second model, performed on the subsample of parents, we included two additional variables: division of childcare tasks and parity (1, 2, or 3+ children). This second model was also stratified by parity, i.e. implemented separately for respondents with one child and respondents with two or more children. To enhance the interpretability of the findings, we present the results in terms of average marginal effects (differences in predicted probabilities of changing fertility intentions).

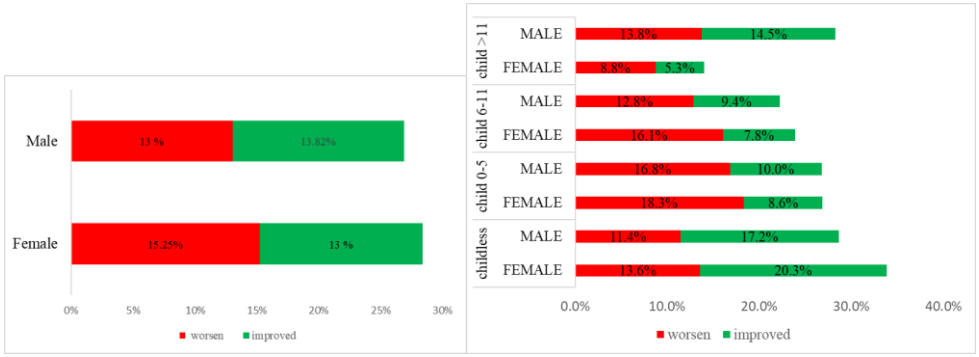
5. Descriptive findings

The majority of participants did not change their fertility intentions between the pre-Covid-19 period and the time of the survey (Figure 1). In fact, a non-negligible share of

[2] Regarding the variable on the shift in the division of childcare, this category was eliminated due to the small number of respondents (37). The variable thus only has four categories.

the sample experienced an improvement of their fertility intentions, to a similar extent among men (14%) and women (13%). Conversely, a significant share experienced a decline in their fertility intentions, with women (15%) showing a slightly higher incidence compared to men (13%). The results by parenthood status (Figure 1) confirm our hypothesis and are in line with the existing literature on the subject: parents seem to be those for whom fertility intentions worsened the most. Women and men without children, on the other hand, experienced greater improvements (around 20% for women and 17% for men). This is a reasonable outcome for a country like Italy, where the fertility rate is well below the replacement rate (1.24, ISTAT, 2019). Our results also show that, when children are present, their age plays a very relevant role in changing fertility intentions. Parents with a youngest child under 6 report the greatest deterioration in fertility intentions, and those with a youngest child between 6 and 11 also experienced noticeably worsening than respondents with a youngest child older than 11. Looking at the improvement in fertility intentions, parents with older children are also those who improved fertility intentions the most (especially for men), second only to childless respondents.

Figure 1 – *Changes in fertility intentions from the pre-COVID-19 period to the survey time, by gender (left) and by gender and parenthood status (right).*

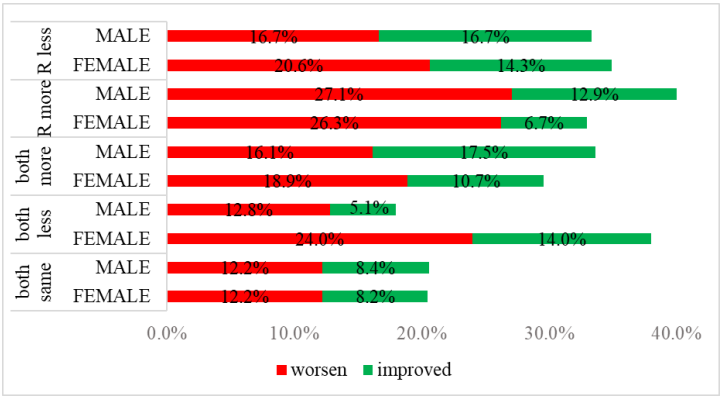


Source: Familydemic survey, 2021

One of our hypotheses was that, during the pandemic period, an equal division of unpaid work among partners could be positively associated with a shift towards higher fertility intentions. Figure 3 shows how respondents who experienced a reduction in their workload (with greater involvement of the partner) reported a more pronounced improvement in fertility intentions, while those indicating an increase in the time spent on domestic duties (with lower involvement of the partner) had a greater worsening in fertility intentions. It should be emphasized that the starting situation (before Covid-19) was highly unequal to the disadvantage of women, who were carrying over a large share of the domestic burden (Figure A1, Appendix). This clarification can account for the

finding depicted in Figure 3, specifically in correspondence with the "both more" category (which indicates an increase in domestic workload for both partners). Starting from a disadvantaged position, women who take on additional responsibilities (despite an additional effort from their partners) experience a greater decline in fertility intentions compared to men. In contrast, men, in the "both more" category, are more likely to show an improvement in fertility intentions compared to women.

Figure 3- *Change in fertility intentions between the pre-Covid-19 period and the time of the survey, by gender and partners' shift in domestic work.*

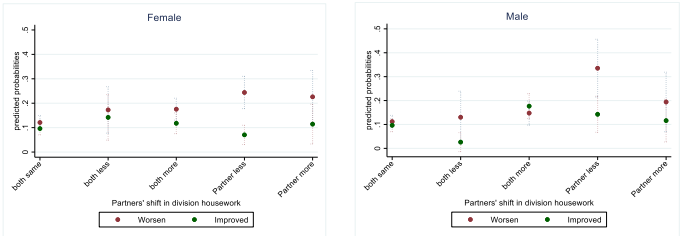


Source: Familydemic survey, 2021

6. Multivariable findings

Multivariable findings are consistent with the descriptive ones: Figure 4 shows the predicted probabilities of improving or worsening fertility intentions, depending on the shifts in the distribution of domestic workload, calculated for the overall sample (Full model results are not shown due to space constraints but are available upon request).

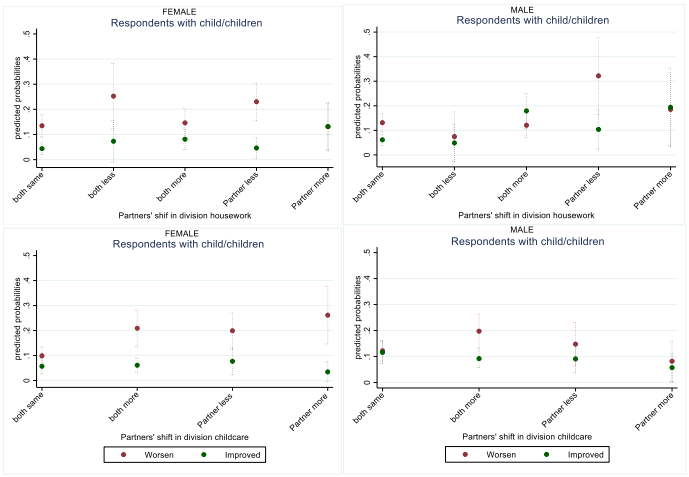
Figure 4 – *Predicted probabilities of changing fertility intentions by gender and partners' shift in housework.*



Source: Familydemic survey, 2021
Models (overall sample) control for: presence and age of children in the household, respondent's age at interview (linear and squared); level of education; respondent's pre-pandemic employment status; respondent's gender ideology.

In line with our hypotheses, the results indicate that respondents of both genders are more likely to reduce their fertility intentions and less likely to improve them when they assume a greater share of domestic work during the pandemic. However, there are significant gender differences. Women present a higher likelihood of increasing their fertility intentions from the moment they reduce some of their tasks (variable modes: "Both less" or "Partner more"). On the other hand, men are more likely to increase their fertility intentions when both partners demonstrate greater commitment ("Both more"). Figure 5 displays the predicted probabilities of changing fertility intentions depending on the shifts in the division of housework and childcare responsibilities, calculated for respondents with children.

Figure 5- *Predicted probabilities of changing fertility intentions by partners' shift in division housework and by partners' shift in division childcare.*



Source: Familydemic survey, 2021
Models (reduced sample of respondents with child/children) control for: respondent's age at interview (linear and squared); level of education; respondent's pre-pandemic employment status; respondent's gender ideology.

Regarding the shift in the division of homework, the results for parents mirror those obtained for the overall sample. With respect to the findings related to the shift in the division of childcare duties, we observe, on the contrary, that women reporting less time devoted to childcare experienced worsened fertility intentions. The variable for gender ideology is significant and positively correlated with an improvement in fertility intentions for men with children. However, when the variable on the change in the division of childcare is included in the model, the significance of 'gender ideology' is reduced (not all realised models are included in the paper but can be made available upon request). Additionally, multivariable findings indicate that, compared to childless respondents, parents (especially with children under the age of 6) have a higher

probability of reporting a worsening in their fertility intentions and a lower probability of reporting improved fertility intentions. One of our hypotheses posited that respondents with higher level of education would report an improvement in their fertility intentions: our results confirmed it for men. Furthermore, the presence of stable employment before the pandemic demonstrates a significant positive effect on the improvement of fertility intentions among women with child/children.

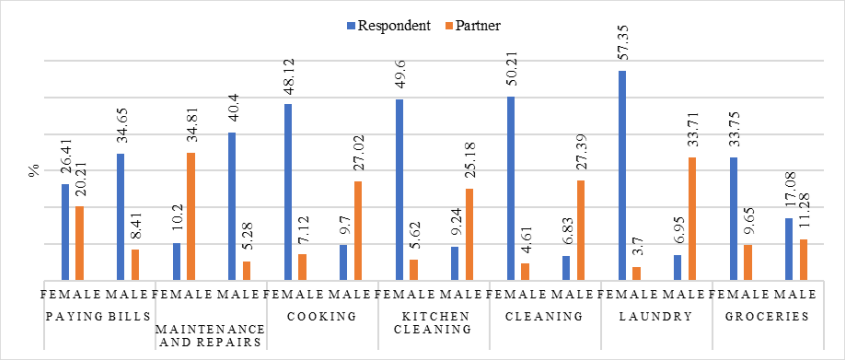
7. Conclusions

This study aimed to examine the shift in fertility intentions during the pandemic in Italy, identifying the key factors influencing these changes. The independent variables we primarily focused on were the presence and age of children and the change in the division of unpaid work between partners. We also assessed the effect of variables measuring the respondents' level of education and gender ideology. We hypothesized that a more equal division of roles could lead to improved fertility intentions, while an increase in the burden of unpaid work, especially for women, could have a detrimental impact. Our findings provided partial support for the proposed hypotheses. It was revealed that a greater domestic burden had a negative impact on changes in fertility intentions for both men and women. Among women, an enhancement in fertility intentions was observed when they experienced relief from certain tasks due to increased involvement from their partners. Conversely, for men, an increase in domestic responsibilities positively correlated with improved fertility intentions, but only when their partners also displayed an increase in involvement. More intriguing is the finding regarding the division of childcare tasks. We hypothesized that additional care obligations resulting from the pandemic might lead to a worsening of fertility intentions, particularly among women due to their greater involvement in informal caregiving. Surprisingly, our results showed that women who reduced their childcare commitments had worsened fertility intentions. It is important to underline that this finding does not imply a causal relationship or assume that a more equitable distribution of childcare responsibilities would hinder fertility intentions. Instead, this finding could suggest that respondents may have reduced the time dedicated to childcare due to various objective difficulties in balancing multiple obligations. These challenges may have been exacerbated by the contingent situation of the pandemic, which could have had an impact on fertility intentions as well. Moreover, our study revealed further evidence supporting the influence of education and gender ideology in shaping fertility intentions. Specifically, men with a more egalitarian gender ideology and higher levels of education demonstrated a higher likelihood of experiencing improvements in their fertility intentions. Our article contributes to the existing literature on Covid-19 and fertility intentions by filling the need, not addressed by previous studies, to analyse shifts in the

post-pandemic period (18 months after the onset of the pandemic) and get a longer-term perspective. The division of roles within the couple during the pandemic was identified as an important factor to explain the shift in fertility intentions and our findings emphasize the importance of rebalancing the division of roles within families through targeted gender equality policies that promote more shared practices of domestic responsibilities and ultimately favour future fertility outcomes. This study has some limitations. The data relied on an online survey, and retrospective questioning of pre-Covid-19 fertility intentions may have affected the reliability of responses. Additionally, the two-year survey period itself could have influenced the fluctuations in fertility intentions, regardless of other factors. The possibility of conducting the same analyses across countries with diverse welfare regimes presents an opportunity for future comparative studies, which can offer additional insights into this research question.

Appendix

Figure A1 – *Partners' self- reported division of domestic duties before COVID-19.*



Source: Familydemic survey, 2021

References

ALDEROTTI G., VIGNOLI D., BACCINI M., MATYSIAK A. 2019. Employment Uncertainty and Fertility: A Network Meta-Analysis of European Research Findings, *Working Paper* 2019/06, DISIA.

ARPINO B., PASQUALINI M., BORDONE V., SOLÉ-AURÓ A. 2020. Indirect consequences of COVID-19 on people's lives. Findings from an on-line survey in France, Italy and Spain. SocArXiv.

CAZZOLA A., PASQUINI L., ANGELI A. 2016. The relationship between unemployment and fertility in Italy: A time-series analysis, *Demographic Research*, Vol. 34, No.1, pp. 1-38.

DEL BOCA D., OGGERO N., PROFETA P., ROSSI M.C. 2021. Did COVID-19 affect the division of labor within the household? Evidence from two waves of the pandemic in Italy. *CESifo Working Paper Series* 9125, CESifo.

ESPING-ANDERSEN G., BILLARI, F. C. 2015. Re-theorizing family demographics, *Population and Development Review*, Vol. 41, No. 1, pp. 1-31.

FIORI F., RINESI F., PINNELLI A., PRATI S. 2013. Economic insecurity and the fertility intentions of Italian women with one child, *Population Research and Policy Review*, Vo. 32, pp. 373-413.

GOLDSCHEIDER F., BERNHARDT E. LAPPEGÅRD T. 2015. The Gender Revolution: A Framework for Understanding Changing Family and Demographic Behavior, *Population and Development Review*, Vol. 41, No. 2, pp. 207-239.

GUETTO R., VIGNOLI D., BAZZANI G. 2021. Marriage and cohabitation under uncertainty: The role of narratives of the future during the COVID-19 pandemic, *European Societies*, Vol. 23, No. sup1, pp. 674-688.

GUETTO R., BAZZANI G., VIGNOLI D. 2022. Narratives of the future and fertility decision-making in uncertain times. An application to the COVID-19 pandemic. *Vienna Yearbook of Population Research*, Vol. 20, pp. 1-38.

KUROWSKA A., BARARDEHI I.H., FULLER S. *et al.* 2023. Familydemic Cross Country and Gender Dataset on work and family outcomes during COVID-19 pandemic, *Scientific Data*, Vol. 10, No. 2, pp.1-11.

ISTAT 2019. I tempi della vita quotidiana. Lavoro, conciliazione, parità di genere e benessere soggettivo.

ISTAT 2023. Report Indicatori Demografici, Anno 2022.

MATYSIAK A., SOBOTKA T., VIGNOLI D. 2021. The Great Recession and Fertility in Europe: A Sub-national Analysis, *European Journal of Population*, Vol. 37, pp. 29–64.

MENCARINI L., TANTURRI M.L. 2004. Time use, family role-set and childbearing among Italian working women, *Genus*, Vol. 60, No. 1, pp. 111-137.

MENCARINI L., VIGNOLI D. 2018. Employed women and marital union stability: It helps when men help, *Journal of Family Issues*, Vol. 39, No. 5, pp. 1348-1373.

MENNITI A., DEMURTAS P., ARIMA S., DE ROSE A. 2015. Housework and childcare in Italy: A persistent case of gender inequality, *Genus*, Vol. 71, No.1, pp. 79-108.

MILLS M., MENCARINI L., TANTURRI M.L., BEGALL K. 2008. Gender equity and fertility intentions in Italy and the Netherlands, *Demographic research*, Vol. 18, pp. 1-26.

MODENA F., SABATINI F. 2012. I would if I could: Precarious employment and childbearing intentions in Italy, *Review of Economics of the Household*, Vol. 10, No. 1, pp. 77-97.

NALDINI M. 2015. La transizione alla genitorialità. Da coppie moderne a famiglie tradizionali. Bologna, Il Mulino.

NEYER G., LAPPEGÅRD T., VIGNOLI D. 2013. Gender Equality and Fertility: Which Equality Matters?, *European Journal of Population*, Vol. 29, No. 3, pp. 245-272.

NOVELLI M., CAZZOLA A., ANGELI A., PASQUINI L. 2021. Fertility intentions in times of rising economic uncertainty: Evidence from Italy from a gender perspective, *Social Indicators Research*, Vol. 154, No.1, pp. 257-284.

PINNELLI A., FIORI F. 2008. The Influence of Partner Involvement in Fatherhood and Domestic Tasks on Mothers Fertility Expectations in Italy, *Fathering: A Journal of Theory, Research and Practice about Men as Fathers*, Vol. 6, No.2, pp. 169-191.

RIEDERER B., BUBER-ENNSER I., BRZOZOWSKA Z. 2019. Fertility intentions and their realization in couples: How the division of household chores matters, *Journal of Family Issues*, Vol. 40, No. 13, pp. 1860-1882.

VIGNOLI D., GUETTO R., BAZZANI G., PIRANI E., MINELLO A. 2020. A reflection on economic uncertainty and fertility in Europe: The narrative framework, *Genus*, Vol. 76, No. 1, pp. 1-27.

VIGNOLI D., GUETTO R., BELLANI D. 2022. Covid-19 as an Engine of Family Reshuffling. Gender Equality and Relationship Quality during the Pandemic, *Working Paper*, DISIA.

Eleonora MIACI, Sapienza University of Rome; eleonora.miaci@uniroma1.it
Raffaele GUETTO, University of Florence, raffaele.guetto@unifi.it
Daniele VIGNOLI, University of Florence, daniele.vignoli@unifi.it

MOBILE PHONE DATA FOR POPULATION ESTIMATES AND FOR MOBILITY AND COMMUTING PATTERN ANALYSES

Fabrizio De Fausti, Roberta Radini, Tiziana Tuoto, Luca Valentino

Abstract. The use of mobile phone data (MPD) for statistical production has been widely explored in the past decade. In official statistics, MPD can be used to supplement and enrich the information available through administrative data and social surveys, taking advantage of the richness of MPD both in terms of timeliness and greater spatial availability. The National Statistical Institute (Istat) has been undertaking this investigation some years ago, particularly regarding population density estimation and small-scale mobility analysis.
In this contribution, we analyze the usability and potential of MPD, highlighting the stages at which MPDs can augment the information already available through administrative data and social surveys. First, we assess the reliability of MPDs through comparison with official estimates. In addition, to analyze and understand people's spatio-temporal behavior, it is important to better understand the location of MPDs, i.e., information about the geographic reference of cell phones during their activity. Finally, we highlight the assumptions underlying our elaborations, as well as additional potentials and limitations of the available data.

1. Objective and data description

New sources of data, including those held by private entities, are attractive for reuse for statistical purposes. Mobile phone data (MPD) are promising, since today almost everyone carries (at least) a mobile device with them during their daily activities and travels. Over the past decade, many National Statistical Institutes around the world have studied the potential and limitations of these data sources, in several fields: dynamic and present population, tourism, commuting, etc. The MPD can be used to complement and enrich the traditional surveys and the administrative data, thanks to their timeliness and greatest spatial availability in representing human behaviors. The Italian National Statistical Institute (Istat) is interested in exploring the usability and potential of Mobile Phone Data (MPD) in the production of official statistics, first assessing its reliability through comparison with official estimates.

In this contribution we describe a particular type of MPD, so-called Call Detail Records (CDRs), and the data processing steps to be undertaken to estimate the location of the MPD, i.e., information about the geographic reference of cell phones during their activity. Finally, we show some potential applications of the MPD, for estimating population density and analysing mobility patterns.

1.1. Data

ISTAT received from a mobile network operator (MNO, the data provider) Call Detail Records (CDRs) of its subscribers' calls over a 6-week period in an Italian province. This provision was managed as part of a Persons & Places research project[1]. Specifically, the data are for the weeks between January 1 and February 12, 2017, and for the province of Pisa, a medium-sized province in central Italy (Tuscany). The province is organized into 37 municipalities (as shown in Figure 1) among which the most populous are Pisa, Cascina, San Giuliano Terme, Pontedera and San Miniato.

Figure 1 – *The province of Pisa and all its municipalities.*

[1] This project complies with the privacy regulation: the data collection complies with the regulations of DL 6.9.1989 n.322; anonymity complies with DL 30.06.2003 n 196 art. 4 paragraph 1 lett. b) and n) and Opinion No. 9802796 of 09.06.2022 as reported in PSN document IST-03434. 4 paragraph 1 lett. b) and n) and Opinion No. 9802796 of 09.06.2022 as reported in PSN document IST-02834.

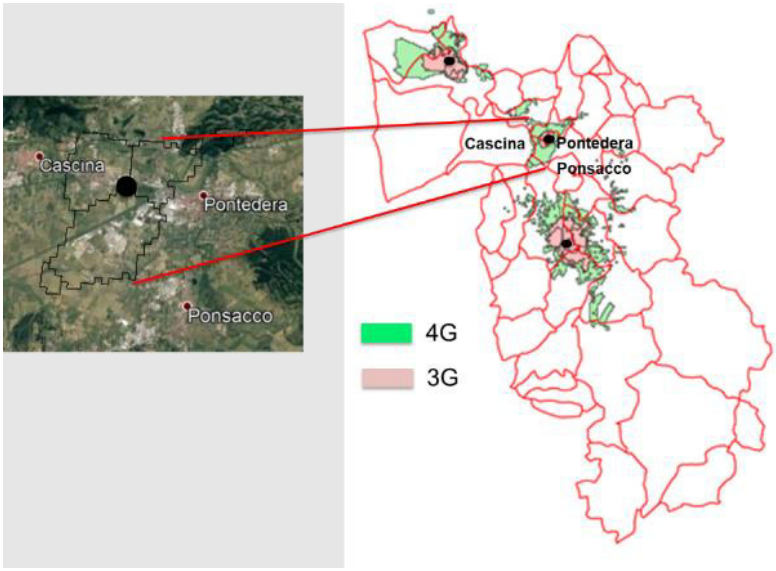
The CDR data available to ISTAT are composed as follows: Caller ID, which is a numeric code associated with each Subscriber Identity Module (SIM) by an algorithm that guarantees its anonymity; the network cell from which the call originated; the date and time the call originated; the duration of the call; and the network cell where the call ended. For text messages, the data report the date and time of the text message and the network cell from which the text message was sent.

The supply consists of about eighteen million CDRs, divided into: less than eleven million calls and seven million text messages. The total number of calling SIMs is just over four hundred thousand. Descriptive analyses are shown in Section 3. CDRs are processed to ensure anonymity.

To analyze and understand people's spatiotemporal behavior, it is important to evaluate MPD localization, i.e., information related to the geographic referencing of cell phones during their activity. In the case of CDR, we have passive localization that corresponds to the code of the antenna/sector to which the calling device has been connected. In fact, the cellular signal is picked up by an antenna and enters the network. The antennas are the cellular phone installations that receive and retransmit signals from cellular phones that are distributed throughout the territory in a capillary manner, based on population density. Each antenna is designed to serve a limited portion of territory, called a "cell". Cells are divided into different sectors. Each sector is a service characterized by a technology, a direction, and an antenna coverage area (this area is named Service Area, as shown in Figure 2).

Call location information can be obtained by different techniques. In the simplest, localization is based on the position of the antenna. This localization concentrates all calls and text messages in the municipality where the antenna is located, although antenna coverage is very wide and often covers areas that belong to more than one municipality. For example, in Figure 2 on the left, the antenna mast is in the municipality of Cascina, but the coverage covers the municipalities of Cascina, Pontedera, and Ponsacco. The methodology we have applied in this work, in collaboration with the MNO, divides the proportion of the territory according to the BSA, and assigns each call and text message as a percentage to all the municipalities served by the specific BSA.

Figure 2. – *Left: An example of best service areas (BSAs) of an antenna in 4G technology. The antenna tower is located at the point represented by the black circle, contains three sectors, and the different areas represent the corresponding three BSAs. Right: an example of BSAs for different technologies in the same antenna tower. Three antennas are shown, BSAs for 3G technology are represented in pink and those for 4G technology in green.*



Specifically, knowing the percentage of the coverage area of each sector for each municipality and considering the uniform distribution of calls for each BSA, the call rates of each BSA for each municipality were calculated based on the percentage of coverage. The following analyses are based on these calculations. Potentially, additional information can be introduced, e.g., land use of the area covered by the BSA, population counts from other sources, the presence of specific points of interest (universities, large local units and businesses, large shopping centers, large hospitals, ...), and as a result, different assumptions can be applied to distribute the calls from each BSA to each municipality. In this application, we chose the simplest hypothesis, with the intention of testing its robustness in further applications based on more sophisticated methods and hypotheses. In addition, it is worth noting that many of the additional data sources that can be exploited to enrich the current hypothesis are often not available at the level of the BSA, which is an irregular polygon not linked to any administrative territorial unit. For competitive exploitation of all additional sources, a further step of mapping the BSA on a regular grid is necessary and recommended.

2. First analyses on Pisa CDR

Figure 3 shows the number of SIM and the number of calls per day. In the period between two holidays (New Year's Day and Epiphany), the trend is irregular with respect to the following weeks. During weekdays, the trend is regular, with a sharp drop on weekends. This trend is also documented in phone traffic in other countries (de Jonge *et al.* 2012, Douglas *et al.* 2015, Furletti *et al.* 2017).

Figure 3 – *SIMs (blue) and calls (red) per day in the period between 1st January and 12th February.*

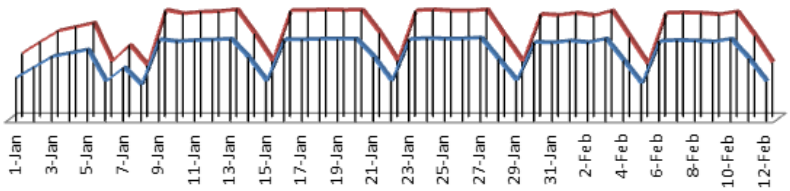


Figure 4 shows that the daily pattern of voice call events and SMS events considered together is the same as SMS events alone. For this reason, the data were used without any distinction between voice and text data.

Figure 4 - *Voice calls and text messages per day (pink), SMSs only (red).*

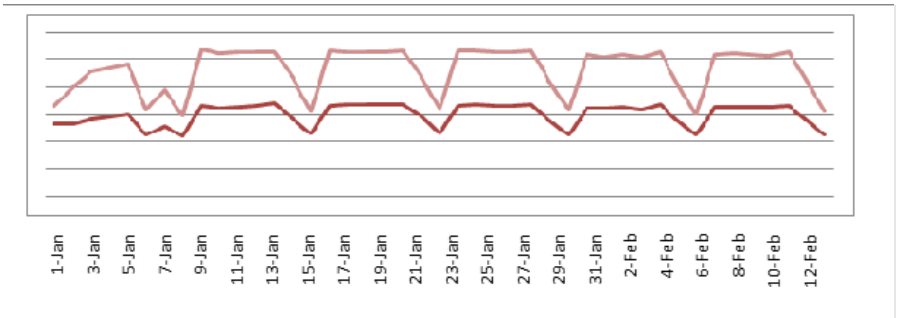
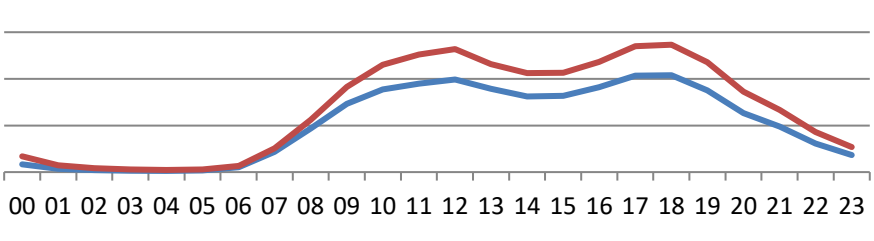


Figure 5 shows that the hourly number of voice calls is similar on weekdays and weekends. Only the volumes are different. Two peaks of voice calls are noted, around 12:00 noon and 6:00 p.m. on each day.

Figure 5 – *Voice calls during the working days and the weekend, red and blue, respectively.*



3. Results

3.1. Mobile phone data for population estimates

To properly use MPD for population estimates, we first studied the correlation between the MPD and official population counts, i.e., the count of people residing in the reference area at the reference time.

CDRs provide information on the activity of MP (mobile phone) users at a given date (with detailed time) and at a very small spatial scale. Calls and text messages can be used to produce population count estimates given certain basic assumptions:

- High level of MP penetration, which is the number of active MP users per 100 people within a specific population;
- High level of mobile network coverage in the territory, i.e., the portion of the geographical area, throughout the country, in which people can make calls and send messages from their cell phones;
- knowledge of the MNO market share, that is, the percentage of total subscribers belonging to a particular MNO.

High values for the previous indicators allow us to use the MPD to derive some estimates of population counts under reasonable assumptions; see Deville *et al.* 2014 and Douglas *et al.* 2015 for a discussion of this topic.

Italy has one of the highest MP penetration rates in developed countries; in fact, the percentage of MP per 100 citizens is about 154% in 2016[2], which means that, on average, each person owned 1.54 cell phones in 2016. Italy also shows a high coverage of MP networks in the territory, for example, 4G technology coverage is

[2] This information is published on AGCOM web site: https://www.agcom.it/servizi-di-rete-mobile (visited 22-01-2024)

97% and 3G technology coverage is 99% of the Italian territory[3]. In addition, the close cooperation with MNO ensures that the market share can be assessed on a small spatial scale, under the confidentiality constraint of business secrecy.

To investigate the correlation of MPD with official population data, we first focused the analysis on the nighttime population. The approximation of residential population with nighttime mobile phone users has been already exploited in several works (Ma and Wu 2012, Deville *et al.* 2014, Douglas *et al.* 2015). In this paper, we identify mobile phone users with SIM cards.

An initial study was conducted by considering SIM localization using antenna tower location. This work highlighted the limitations of this type of localization and suggested the need to implement a finer geolocation methodology. For this purpose, in collaboration with the MNO, we adopted the BSA-based approach mentioned in the previous section. The municipality of residence is then assigned to each SIM according to the following procedure: a SIM's home is located in the BSA most frequently recorded in the CDR records during the nighttime hours, from 8 p.m. to 7 a.m. If this BSA covers several municipalities, the SIM is counted with a percentage for each of them, and the percentage assigned to each municipality is proportional to the area covered by the BSA.

Figure 6 shows a scatter plot of the count of SIMs active at night against the January 2017 residential population estimates for the province of Pisa at the municipal level. There is a reasonably good relationship, approximately linear as indicated by the LOESS regression interpolation, in blue in the graph. In the linear regression model, the correlation coefficient is 0.94, reflecting the adequacy of the model in predicting the residential population through nighttime mobile phone users. Similar results in terms of high correlation are also obtained when the logarithmic transformation is considered, and the extreme value represented by the city of Pisa is excluded from the analysis.

The high correlation between phone users and residential population is also confirmed when the analysis focuses on SIMs active during the daytime. Specifically, we analyzed SIMs calling between 5 p.m. and 6 p.m.: the most frequent peak hour during the observed period, Monday through Friday. We used the population of phone users identified by these SIMs as a predictor of the residential population, and the SIMs are assigned to the municipality resulting from the nighttime location.

[3] Data processed from open data published by AGCOM referring to 2018, https://maps.agcom.it/

Figure 6 - *Scatter plot of nighttime mobile phone users versus residential population (municipalities in Pisa province).*

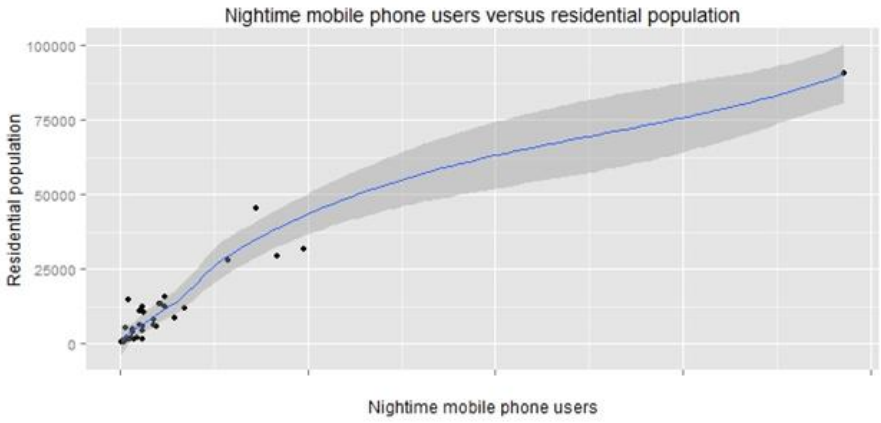
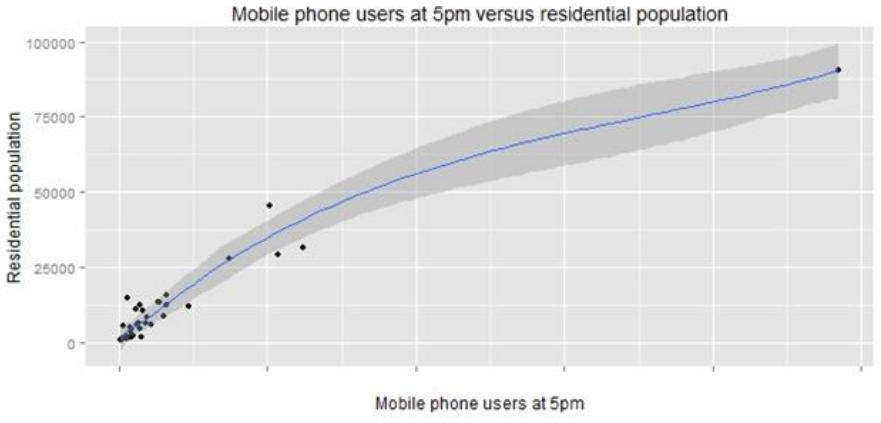


Figure 7 shows the scatter plot of the count of active SIMs from 5 a.m. to 6 p.m. against the January 2017 residential population estimates for the province of Pisa at the municipal level. Again, a reasonably good approximation of the linear relationship is observed, as indicated by the LOESS regression interpolation (in blue in the plot). In the linear regression model, the correlation coefficient improves to 0.95.

Figure 7 - *Scatter plot of mobile phone users at 5 pm versus residential population (municipalities in Pisa's province).*

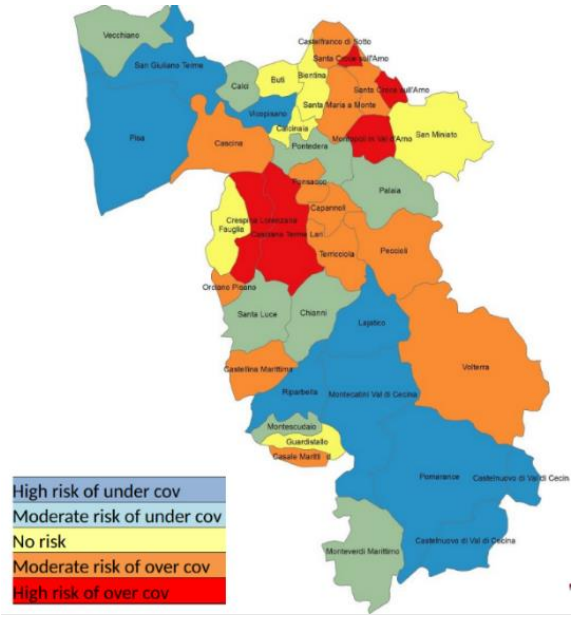


These results are considered satisfactory and encourage us to use CDRs to estimate something that cannot currently be observed with sample surveys and administrative data, but which could be a new output within official statistics.

One example of these new opportunities offered by MPDs is related to their use in combination with the coverage surveys included in the new population census framework. In the Census Transformation Program, MPDs allow us to identify areas that might be problematic for census counts; for example, areas at risk of over- or under-coverage can be identified by comparing population estimates from MPDs with counts of persons enrolled in registers.

The risk of over/under-coverage can be defined at a very small scale, and this information can be used both at the sampling stage, when designing the coverage sample survey, and at the estimation stage, when small area population estimates are to be provided. Figure 8 shows the areas at risk of under/over-coverage for the province of Pisa based on a comparison of MPD population estimates with the January 2017 residential population.

Figure 8 - *Municipalities at risk of over/under-coverage, Pisa province.*



The MPD population estimates are based on the nighttime MP users by applying a multiplier estimator (EMCDDA, 1999, Jandl 2009) that considers the market share of the MP operators in the Pisa province. The municipalities where the MP

population estimates are similar to the official estimates, with differences lower that 10% are in yellow. The municipalities with lower MP population estimates are in the light red area (up to 50% lower than the official estimates) thus highlighting a moderate risk of over-coverage. At the same time the higher ones are in the light blue area and show a moderate risk of under-coverage. The municipalities with the highest risk of over-coverage are in dark red, since the MP population estimates are lower up to 1.5 times than the counts enrolled in the registers; instead, the data referring to under coverage, as the estimates are higher, are reported in blue.

3.2. Mobility pattern analysis: the Origin-Destination Matrix

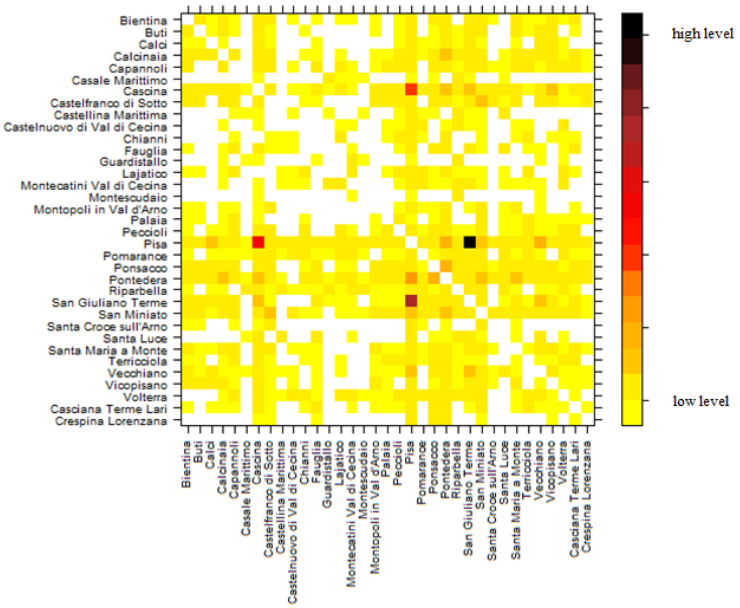
Another opportunity offered by MPD is related to understanding how and where people move, so-called mobility pattern analysis. There are at least two ways to study mobility through MPD: the first is based on the relative densities of CDRs, both across areas and over time, as shown in the previous section; the second is based on anonymized data at the individual level. Some results of the first type of analysis were shown in the previous section, while in this section we show some results using the second type as input. In this case, meaningful placement for MPD such as "home" and "work/study" is determined as follows:

- "home" is the municipality where a MP user is more frequently found during the nighttime, as in the previous section for the residential population estimates;
- "work/study" is the municipality where the MP user is repeatedly observed during the daytime hours.

By aggregating individual-level data for which home and work/study were previously derived, the origin-destination flows of home and work/study can be produced. In Figure 8 we propose an origin-destination matrix for the Pisa province at municipal level, where only movements within the province are considered. The main diagonal represents people who live ("home") and "work/study" in the same municipality. These intra-municipal movements are not of interest in this analysis, and are not shown in the matrix, although they account for 70% of the data analysed. The intensity of the movement is represented by the intensity of the colors on the matrix. Again, the results are in line with the information coming from other sources: the most used routes are those involving Pisa and its neighboring municipalities (Cascina and San Giuliano Terme), as well as the municipalities where the province's largest establishments are located (e.g., Pontedera). This result is in line with expectations since the city of Pisa is a national attraction for tourism and is home to an important university. It also has a much larger population than other centers in the province.

One disadvantage of this analysis, compared with results derived from administrative data, is that it is not possible to identify the reason for mobility; on the other hand, MPD allow us to assess the frequencies of mobility, which cannot be derived from administrative data.

Figure 9 - *This Origin-Destination (OD) Matrix describes people movement in the Pisa Province.*



4. Lesson learned and next steps

The results of the CDR analyses described in this report are definitively encouraging regarding the potential of MPD for both population estimates and the study of mobility patterns.

A key issue for effective exploitation of CDRs is small-scale localization of MP users' activities (calls and text messages). Currently, localization based on antenna position (i.e., all calls and text messages are assigned to the municipality where the antenna is located) has weaknesses that compromise the reliability of population estimates and all other statistics related to this concept. We overcame these drawbacks through collaboration with the MNO, which provided us with the percentage of territory served by the BSA. On this basis, we were able to develop a procedure that assigns each MP activity as a percentage to all municipalities served

by the specific BSA, and in this way, we greatly improved the location of CDRs and obtained reliable population estimates at the municipal level.

In the future, in agreement with the MNO, we will be able to produce statistics at a smaller scale than the individual municipality, i.e., census sections, to take full advantage of the enormous amount of information that MPDs provide us on "urban rhythms" to design and optimize mobility in dense urban centers.

Finally, the analyses proposed in this report are still a local observation, limited to one province. The availability of new data will enable us to examine these results on a larger scale, such as at the regional level and at the national level.

References

DE JONGE E., VAN PELT M., ROOS M. 2012. Time patterns, geospatial clustering and mobility statistics based on mobile phone network data. *Discussion paper 201214,* Statistics Netherlands.

DEVILLE P., LINARD C., MARTIN S., TATEM A.J. 2014. Dynamic population mapping using mobile phone data, *PNAS*, Vol. 111, No 45, pp. 15888-15893.

DOUGLASS R.W., MEYER D.A., RAM M., REDEOUT D., SOND D. 2015. High resolution population estimates from telecommunications data, *EPJ Data Science*, Vol. 4, No 1.

EUROPEAN MONITORING CENTRE FOR DRUGS AND DRUG ADDICTION (EMCDDA) 1999 Scientific Review of the Literature on Estimating the Prevalence of Drug Misuse on the Local Level. Lisbon: EMCDDA, July 1999.

FURLETTI B., TRASARTI R., CINTIA P., GABRIELLI L. 2017. Discovering and Understanding City Events with Big Data: The Case of Rome*, Information*, Vol. 8, No 3, p.74.

JANDL M. 2009. A multiplier estimate of the illegally resident third-country national population in Austria based on crime suspect data. *Working Papers 2*, Hamburg Institute of International Economics. Database on Irregular Migration.

MA X., WU L. 2012. Towards Estimating Urban Population Distributions from Mobile Call Data, *Journal of Urban Technology,* Vol. 19, No 4, pp. 3-21.

Fabrizio DE FAUSTI, ISTAT, defausti@istat.it
Roberta RADINI, ISTAT, radini@istat.it
Tiziana TUOTO, ISTAT, tuoto@istat.it
Luca VALENTINO, ISTAT, valentino@istat.it

A WEB SURVEY ON AN ELUSIVE POPULATION: A FOCUS ON INDICATORS TO MANAGE DATA COLLECTION PROCESS

Monica Perez, Linda Porciani, Federico De Cicco[1]

Abstract. In 2021, the Italian National Statistical Institute (Istat) started the study to carry out a survey about labor discrimination of LGB (lesbians, gays and bisexuals) persons without legal relationship[2]. The survey population is extremely sensitive and hard-to-count mainly for two reasons: it is (self) defined by sexual orientation and there is no list to count and identify the initial population. Moreover, labor discrimination is a topic generally underrepresented in any kind of subgroups of Italian population.

The survey fieldwork has been carried out from January to the end of May in 2022.

The project team chose the web Respondent Driven Sample (w-RDS) method as the most suitable way to sample the population. The web questionnaire has been developed following "a privacy by design" approach to preserve the privacy of respondents.

The main efforts of data collection design were devoted to the implementation of a system of ad hoc indicators able to: 1. monitor the quality of data collection process (how to define and calculate the response rate without an initial population?); 2. evaluate the goodness of the sample (how many and which respondents can guarantee the data quality criteria?); 3. allow decisions regarding the change of data collection techniques or data design *in itinere*.

The paper addresses methodological and organizational aspects of data collection design, focusing on system of monitoring survey indicators, for an experimental survey on LGB adult population resident in Italy. Lesson learned could be useful for RDS implementation in future surveys conducted by National Statistical Institute.

[1] The paper is the joint work of the authors. More in detail, the single paragraphs are as follows: par. 1 and 2 to Monica Perez; par. 2.1 to Federico De Cicco; par. 2.2 and 2.3 to Linda Porciani; par. 4 to the joint work of the authors.

[2] Italian legislation (law n. 76, 20 may 2016) recognized civil union among homosexual persons.

1. Introduction

The survey has been conducted by Istat within the project on "Access to work, working conditions, labour discrimination of LGBT+ people and diversity policies implemented at enterprises" carried out in collaboration with the National Office Against Racial Discrimination (UNAR). The objective of the study is to provide a picture of the perception and prevalence of forms of discrimination, threats and assaults that homosexual and bisexual people may have experienced according to sexual orientation in the daily life, mainly focusing on labour situation.

With reference to the part of the project aimed at investigating the experiences of the LGBT+ population, given the heterogeneity and plurality of this population which is extremely sensitive and elusive, we ideally divided the population into three subgroups that were investigated through three different surveys: (i) homosexual people who are in civil union or have been in civil union previously (selected by municipal lists as of January 1, 2020) (2020-2021); (ii) LGB people who are not in a civil union nor have been in a civil union previously (2022); (iii) trans and non-binary people (2023).

As for the survey on the LGB people without legal relationship (ii), which is the subject of this paper, the lack of a sampling frame led us to use the Respondent Driven Sampling method, which is a probabilistic sampling method based on social ties (De Rosa et al. 2020; Sheim et al. 2016; Vitalini 2012). RDS is similar to snowball sampling, a chain-referral sampling method where participants recommend other people they know belonging to the target population. The main difference between the two methods is that RDS is mathematically tweaked to add an element of randomness. RDS can be thought as a group of snowballs, each rolling down a hill in its own random direction.

This type of sampling model is usually associated with the web technique for data collection, the so called WebRDS. It has been used in surveys conducted to study the risk on gays of HIV transmission in Vietnam and Sweden (Bengtsson et al. 2012; Stromdahl et al. 2015); transgender women in San Francisco (Wesson et al. 2013); the risk on gays and bisexuals of sexually transmitted diseases in New Zeland (Ludlman et al. 2015).

Application of the RDS method as well as privacy of the respondents and the protection of data confidentiality due to the sensitiveness of the theme are crucial points that have numerous implications with repercussions on the survey design.

2. The project design

The Lgb survey, due to its experimental design, has implied innovations mainly in three phases of the Generic Statistical Business Process Model (GSBPM), that are interconnected:

a. Design phase: *Privacy by design*
b. Build and Collect phase: *Respondent Driven Sampling (RDS)*
c. Evaluate phase: *ad hoc monitoring process indicators*

Before going into the details of each of these steps, it may be helpful to specify that:

a) LGBT associations played an important role in the survey. The survey design based on 50 LGBT associations and 10 initial (potential) respondents per each (the so called "seeds") . The list of seeds was identified by the LGBT Associations among the people belonging to the associations themselves. The choice of the seeds, based on socio-demographic criteria provided by Istat, was a crucial point because the seeds must be capable of generating long chains of recruitment. Each LGBT association was provided with a different link to deliver to its own seeds, in order to trace the origin of the chains of the propagation network;
b) LGBT associations did not have to disclose the identity of selected seeds. Each association had to appoint the Data Processor according to art. 28 of GDPR;
c) each respondent was required to recruit other four individuals belonging to the target population and belonging to the circle of one's own acquaintances;
d) due to RDS method, questions concerning the respondent's acquaintance with the person who recruited him/her (reciprocity of ties) and the size of each respondent's social network need to be included. The former is necessary for calculating the balance condition in the recruitment process, the latter is an indispensable variable for estimating the probability of inclusion (par. 2.2).

2.1 Privacy by design

Given the sensitive topic of the survey privacy measures have been taken - both methodologically and organizationally - for data processing and storage, taking into account a privacy by design and privacy by default approach.

The survey design has been defined according to a specific Data Protection Impact Assessment (Art. 35 of EU Regulation 2016/679) and risk based approach, focusing on risks analysis and the measures to enhance data protection.

On the field, the privacy measures have been realized through a sequence of actions needed to access the questionnaire (Fig. 1). Each respondent receive a link to enter the "Accession module", that is an introductive web page containing preliminary questions aimed at identifying the person's eligibility to be part of the survey sample. The "Accession module" describes the key points of the survey such as the aim, data collection mode, survey period, privacy legislation and, finally, some preliminary questions to ascertain eligibility of the person in the sample, and namely: (i) being aged 18 years or more ; (ii) being resident in Italy; (iii) knowing the person sending the link.

After the compilation of this module respondent needs to indicate an email address to receive a personal link to access the "Questionnaire".

First questions of "Questionnaire" aim to complete the eligibility evaluation of the respondent, according to sexual orientation and marital status. These data are recorded in a different data server respect to the information provided compiling the "Accession module". Moreover, the email address is stored in encrypted mode to minimize the risk of individual identification of respondent.

Once the questionnaire is completed, an "invitation" link - automatically produced by the data collection system - is immediately available to the respondent on the web page to be sent to other potential respondents identified by him/her to further propagate the respondents' chain. The invitation link can be transmitted by e-mail, WhatsApp, SMS or other instantaneous messages services. Simultaneously, the system send the link to the e-mail address of the respondent, so that it remains available to recruited people even at a later time, after closing the web page.

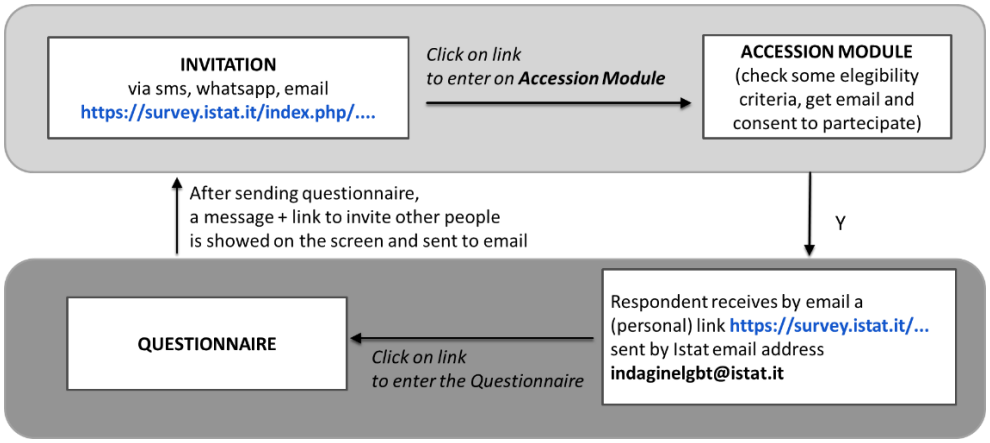
The initial respondent has two possibilities to propagate the questionnaire: i) copying the link showed on web page at the end of own questionnaire and immediately sharing it with other people (by instantaneous messages services); ii) using the link sent by email and sharing it later time.

The following steps are a part of the cycle (Figure 1) starting from again by the "Accesion module". Each respondent should be a recruiter to build a good sample.

The protection of privacy has been guaranteed by the separation of Accession module and Questionnaire: the respondent-recruiter can not access personal data of invited respondents. Moreover, a check system has been implemented on the number of the sent invitation links, which has a maximum of 10 for the associations and 4 for respondents. Each link, both for Accession module and Questionnaire, is unique and it is highly unlikely to reproduce because of the token length based on the combination of 10^{26} order.

Data protection was an element of attention in each survey step. Coded variables have been used for associations, first respondents (seeds) and further respondents in order to avoid the identification of the subjects (even indirectly) end preserve data protection not only of the investigation process but also of the survey monitoring

Figure 1 – *Data collection design scheme*



2.2 Respondent Driven Sampling (RDS)

RDS strategy is helpful to reach a population without a sampling frame to select sampling units and, consequently, without the possibility to have a probabilistic sample.

The sampling strategy based on RDS has a probabilistic approach. It combines the snowball technique - the sample is constructed by using sample units (individuals) provided by the initial recruiters (seeds) and subsequent recruits/recruiters (called nodes) - with a mathematical model that formalizes the recruitment process. Under certain conditions the recruitment process is a Markov chain (probabilistic process). (Salganik e Heckathorn, 2004; Volz e Heckathorn, 2008). The seeds need to be chosen non-randomly, based on the differentiation criterion and their ability to recruit. The recruitment process develops in waves that are generated starting from the initial recruiters until an equilibrium condition is reached, in which the probability of inclusion of the sample units stabilizes.

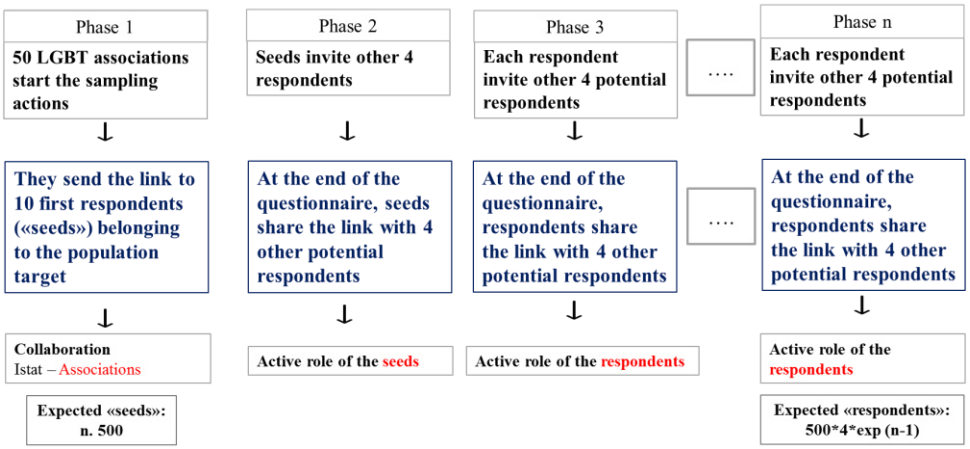
As mentioned above, in the experimental LGB survey, the seeds have been recruited by LGBT associations based on socio-demographic criteria. Afterward, the process of respondent-driven sampling started: at each wave, respondents were used to select or drive the next sampling wave by recruiting other individuals from the target population (Figure 2).

The data collected during the sampling process are used to make inferences about the structure of the social network to which they belong and to obtain unbiased estimates of the target population. Information on: (i) properties of the (responding) nodes; (ii) who recruits whom (recruitment matrix); (iii) size of the respondents'

personal network (number/strength of ties) are basic elements for generating inferences on the characteristics of the population. Under specific assumptions, RDS estimators are asymptotically unbiased (Salganik e Heckathorn, 2004).

The assumptions under the mathematical model concern both network structure and sampling. In fact, the network of the target population must be sufficiently dense and connected so that each node is reachable from the other nodes. Furthermore, the network does not have to be too segmented, to prevent the chains from becoming trapped in subgroups. Such situation would not allow equilibrium to be achieved. Respondents have to mantain symmetrical relationships and must recognize each other as members of the reference population (unoriented network). The number of ties between members must be sufficiently high to support recruitment process (recruitment chains spanning multiple waves) to ensure that each member of the population has a non-zero probability of entering the sample. As far as the sampling, the hypotheses concern: the selection with re-entry of units, the accuracy of estimate of ties, the randomness of recruitment (Gile e Handcock, 2010; Xin, 2013).

Figure 2 – *RDS scheme in LGB labor discrimination survey*



2.3. Monitoring survey process indicators

The effectiveness of the method RDS and the duration of the survey depend on the propagation capacity of the network.

If for any reason a participant decides not to "propagate" because he/she becomes discouraged, loses confidence, loses referrals, that node does not produce offspring and the network reduces its propagation effectiveness by limiting the achievement of a satisfactory sample.

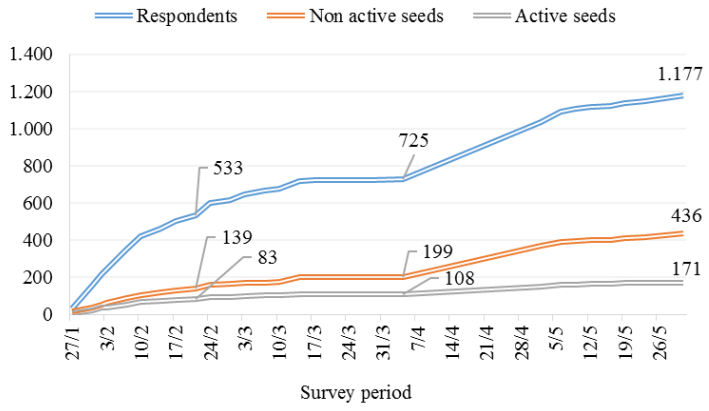
In order to monitor the survey, two sets of indicators have been elaborated: the first set refers to indicators to monitor the strength of «seeds» and network propagation (we call them "network propagation" indicators); the second set, is a set of indicators able to provide essential information about the typology of the respondents ("profile" indicators).

The "network propagation" indicators include the number of active/non active seeds; the number of active/non active respondents; the number of created chains.

The "profile" indicators monitor the number of total respondents (seeds + subsequent respondents) by sex, sexual orientation, age group and participation in LGBT association.

After a month of fieldwork, the indicators gave us some signal of criticalities of the network propagation for LGB population. On 50 involved associations, 24 percent were completely inactive (any seed has compiled the form); 76 percent were active (they have active seeds) producing 6.4 seeds on average (versus the expected 10). Nevertheless, and more seriously for the sample building, 62 percent of active seeds have compiled just its own questionnaire without any propagation activity (83 active seeds and 139 non active seeds), so they did not contribute to create respondent chain. The active seeds generated just 2.4 respondents on average versus the expected 4. Totally, the respondents were 533 after one month (Figure 3).

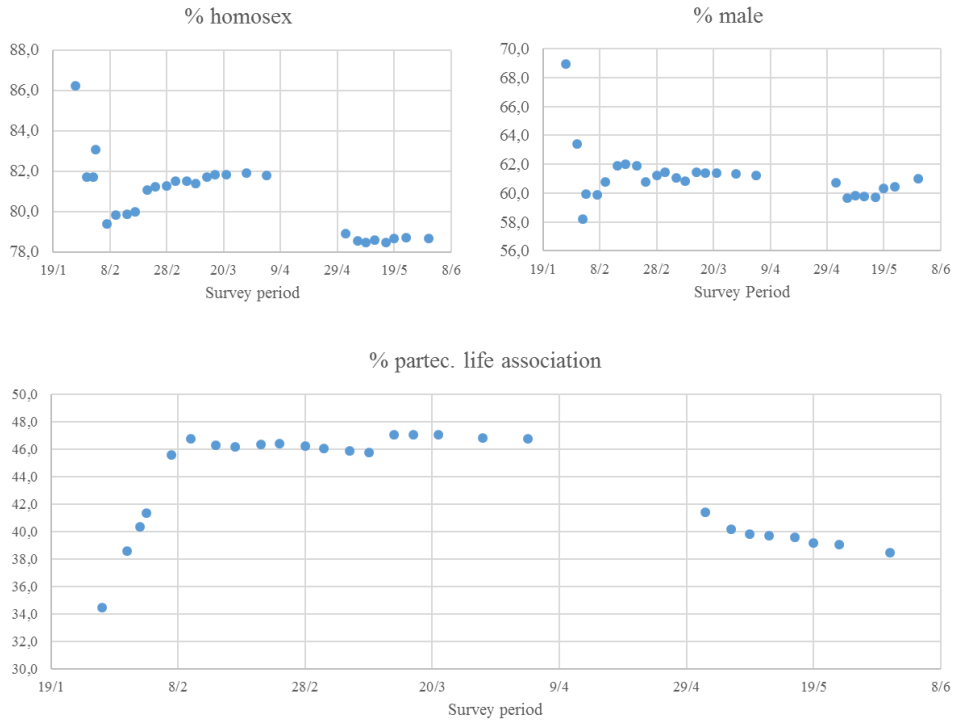
Figure 3 – *"Network propagation" indicators: respondents and non active/ active seeds (absolute values) during survey period.*



The "profile" indicators describe a homogeneous sample in terms of gender composition (more than 60 percent are men), sexual orientation (almost all units are homosexual) and participation in the life of LGBT associations (more than 40

percent of respondents have or have had experience in a thematic association): these characteristics are some of the criteria used to assess the quality of the probability sample (Figure 4).

Figure 4 – *"Profile" indicators (percentage values) during survey period.*



In order to increase the participation of all the expected seeds (10 per each association) and, above all, the generation of all the expected respondents (4 per each recruiter), reminder actions have been activated involving the LGBT associations (45 days after the start of the survey).

In spite of this, a small impact on the spread of the chains was observed. Both types of indicators remained stable (Figure 3 and Figure 4). After two months of data collection, it is only possible to observe an increase in the activity of the associations: from 76 percent to 90 percent of them had at least one active seed. However, the general scenario did not change, confirming the low participation of seeds in the survey (6.5 seeds per association on average), the low level of reproduction activity (64 percent of seeds have no offspring respondents) and the small size of the network

(2.4 respondents per seed on average). After 45 days, the total number of respondents is 725 (+36 percent). Analysis of process indicators suggested the change of data collection strategy, moving from RDS to simple snowball sampling method. This is made by the publication of the questionnaire link on the associations' web page. It implies no distinctions between a seed and a respondent, so the possibility to keep track of who referred who in the sample is loose and consequently the probabilistic approach is lost. The change of the sample strategy has been shared with LGBT associations, because it required their different participation in the survey process.

At the end of fieldwork period (lasted 4 months) the total respondents are 1,177: 54.3 percent are seeds (or first respondents in the second strategy), of which 28 percent are generative (Figure 3).

3. Final considerations

Istat has had its first experience of field application of RDS method for elusive population in LGB labor discrimination survey.

At the end of this innovative and experimental survey, the research group acquired expertise and suggestions for planning further surveys with similar characteristics.

Firstly, the need for ad hoc procedures to manage privacy issues was evident. In the LGB survey, a high privacy protection model was implemented through a two-step access to the questionnaire. The adopted data protection measures in such a survey could be more prominently promoted to enhance the involvement of distrustful people. However, challenges persist in ensuring a balance between privacy and data collection effectiveness.

Secondly, the LGB survey revealed a critical point in the activity of seeds and respondents. A low knowledge base among initial respondents (seeds) could be mitigated through an initial training activity between the research group and seeds, focusing on recruitment strategies. The low activity of respondents' propagation may also be influenced by various factors, such as the sensitive nature of the topic, the length of the questionnaire, and the intricacies of the recruitment process. These aspects warrant further detailed analysis and consideration. Moreover, exploring the potential for providing incentives to respondents could prove beneficial in enhancing participation rates in future Official Statistics surveys employing RDS.

Thirdly, there is a clear need to enhance the set of indicators with more in-depth elements on network propagation. Developing a robust evaluation benchmark for the quality of the data collection process is imperative. This includes not only understanding the breadth of the network but also its depth and the effectiveness of each step in the sampling process.

Fourthly, the representativeness of the sample is a significant challenge in RDS. Unlike traditional sampling methods, RDS does not guarantee a random sample from the target population. The method relies on the social networks of the initial participants (seeds) to recruit additional participants. This can lead to biases, especially if certain segments of the population are more connected or influential within the network. Ensuring that the seeds are diverse and well-connected within the target population can mitigate this issue to some extent. However, the degree to which the sample reflects the true population remains a challenge in RDS studies, and statistical adjustments or modeling techniques may be necessary to account for this bias.

References

BENGTSSON L., LU X., NGUYEN Q. C., CAMITZ M., LE HOANG N., NGUYEN T. A., LILJEROS F., THORSON A. 2012. Implementation of Web-Based Respondent-Driven Sampling among Men Who Have Sex with Men in Vietnam. Published: https://doi.org/10.1371/journal.pone.0049417

DE ROSA ET AL. 2020. Il Web-Respondent driven sampling per lo studio della popolazione LGBT+. Rivista Italiana di Economia Demografia e Statistica, Vol. LXXIV n.1, Gennaio-Marzo 2020.

GILE K. J., HANDCOCK M. S. 2010. ‘‘Respondent-Driven Sampling: An assessing of current methodology’’. Sociol Methodol. 40(1): 285-327.

LUDLAM A., SAXTON P., DICKSON N. P., ADAMS J. 2015. Respondent-driven sampling among gay and bisexual men: experiences from a New Zealand pilot study, BMC Res Notes, 8:549.

SALGANIK M. J., HECKATHORN D. D. 2004. Sampling and Estimation in Hidden Populations Using Respondent-Driven Sampling, Sociological Methodology, Vol. 34, pp. 193-239..

SCHEIM A. I., BAUER G. R., COLEMAN T. A. 2016. Sociodemographic Differences by Survey Mode in a Respondent-Driven Sampling Study of Transgender People in Ontario, Canada, Published Online: 1 Oct 2016 https://doi.org/10.1089/lgbt.2015.0046

STRÖMDAHL S., LU X., BENGTSSON L., LILJEROS F., THORSON A. 2015. Implementation of Web-Based Respondent Driven Sampling among Men Who Have Sex with Men in Sweden. Published: November 12, 2012 https://doi.org/10.1371/journal.pone.0049417

VITALINI A. 2012. L’uso delle reti sociali per la costruzione di campioni probabilistici. Roma: Aracne.

VOLZ E., HECKATHORN D. D. 2008. Probability-Based Estimation Theory for Respondent-Driven Sampling, Journal of Official Statistics. Vol. 24, No. 1, pp. 79–97

WESSON P., QABAZARD R. F., WILSON ERIN C., MCFARLAND W., FISHER R. H. 2013. Estimating the population size of transgender women in San Francisco using multiple methods, pp. 107-112. Published online: 28 Sep 2017 https://doi.org/10.1080/15532739.2017.1376729

XIN L. 2013. Respondent-Driven Sampling: Theory, Limitations & Improvements. Karolinska Institute. Printed by US-AB, Stockholm.

Monica PEREZ, Italian National Institute of Statistics, perez@istat.it
Linda PORCIANI, Italian National Institute of Statistics, porciani@istat.it
Federico DE CICCO, Italian National Institute of Statistics, fedecicc@istat.it

SOCIETÀ E RIVISTA ADERENTI AL SISTEMA ISDS
ISSN ASSEGNATO: 0035-6832

Direttore Responsabile: CHIARA GIGLIARANO

Iscrizione della Rivista al Tribunale di Roma del 5 dicembre 1950 N. 1864

Associazione all'Unione Stampa Periodica Italiana

TRIMESTRALE

La copertina è stata ideata e realizzata da Pardini, Apostoli, Maggi p.a.m.@tin.it – *Roma*

Stampato da CLEUP sc
"Coop. Libraria Editrice Università di Padova"
Via G. Belzoni, 118/3 – Padova (Tel. 049/650261)
www.cleup.it

ATTIVITÀ DELLA SOCIETÀ

A) RIUNIONI SCIENTIFICHE

XXXVII	La mobilità dei fattori produttivi nell'area del Mediterraneo (Palermo, 15-17 giugno 2000).
XXXVIII	Qualità dell'informazione statistica e strategie di programmazione a livello locale (Arcavacata di Rende, 10-12 maggio 2001).
XXXIX	L'Europa in trasformazione (Siena, 20-22 maggio 2002).
XL	Implicazioni demografiche, economiche e sociali dello sviluppo sostenibile (Bari, 15-17 maggio 2003).
XLI	Sviluppo economico e sociale e ulteriori ampliamenti dell'Unione Europea (Torino, 20-22 maggio 2004).
XLII	Sistemi urbani e riorganizzazione del territorio (Lucca, 19-21 maggio 2005).
XLIII	Mobilità delle risorse nel bacino del Mediterraneo e globalizzazione (Palermo, 25-27 maggio 2006).
XLIV	Impresa, lavoro e territorio nel quadro dei processi di localizzazione e trasformazione economica (Teramo 24-26 maggio 2007).
XLV	Geopolitica del Mediterraneo (Bari, 29-31 maggio 2008).
XLVI	Povertà ed esclusione sociale (Firenze 28-30 maggio 2009).
XLVII	Un mondo in movimento: approccio multidisciplinare ai fenomeni migratori (Milano 27-29 maggio 2010).
XLVIII	150 anni di Statistica per lo sviluppo del territorio: 1861-2011. (Roma 26-28 maggio 2011).
XLIX	Mobilità e sviluppo: il ruolo del turismo. (San Benedetto del Tronto, 24-26 maggio 2012).
L	Trasformazioni economiche e sociali agli inizi del terzo millennio: analisi e prospettive (Università Europea di Roma, 29-31 maggio 2013).
LI	Popolazione, sviluppo e ambiente: il caso del Mediterraneo (Università Federico II di Napoli, 29-31 maggio 2014).
LII	Le dinamiche economiche e sociali in tempo di crisi (Università Politecnica delle Marche, 28-30 maggio 2015).
LIII	Mutamento economico e tendenze socio-demografiche tra sfide e opportunità (Università degli Studi Internazionali di Roma, 26-28 maggio 2016).
LIV	Mobilità territoriale, sociale ed economica: modelli e metodi di analisi (Università degli Studi Internazionali di Catania, 25-26 maggio 2017).
LV	Coesione sociale, welfare e sviluppo equo e sostenibile (Università degli Studi dell'Insubria, Varese 24-25 maggio 2018).
LVI	Benessere e Territorio: Metodi e Strategie (Università Politecnica delle Marche, Ascoli Piceno 23-24 maggio 2019).

LVIII Tra marginalità e sviluppo. La sfida della sostenibilità in una prospettiva mediterranea (Università LUMSA, Palermo, 26-27 maggio 2022).

LIX Aspetti economici e sociali dell'invecchiamento demografico (Università degli Studi di Napoli Federico II Napoli, 25-26 maggio 2023).