rivista di statistica ufficiale

REVIEW OF OFFICIAL STATISTICS

n. 3
2022

In this issue:

Istat
Istituto Nazionale di Statistica

rivista di statistica ufficiale

REVIEW OF OFFICIAL STATISTICS

n. 3
2022

In this issue:

Istat

rivista di statistica ufficiale
n. 3/2022

Four-monthly Journal: registered at the Court of Rome, Italy (N. 339/2007 of 19th July 2007).

e-ISSN 1972-4829
p-ISSN 1828-1982

The Scientific Committee, the Editorial Board and the authors would like to thank the anonymous reviewers (at least two for each article, on a voluntary basis and free of charge, with a double-anonymised approach) for their comments and suggestions, which enhanced the quality of this issue of the Rivista di statistica ufficiale.

Editorial Preface

The last issue of the year 2022 of *Rivista di statistica ufficiale/Review of official statistics* addresses themes related to both the use of transaction and mobile phone data for increasing and enhancing the production of statistics and the application of statistical matching methods for integrating different sources, maximising results to carry out richer and more in-depth analyses.

In the first article, Ilaria Benedetti, Federico Crescenzi, and Tiziana Laureti illustrate a measure for assessing Spatial Price Indices accuracy through the adoption of a new approach based on the Jackknife replication technique.

They pursue two main objectives:

- demonstrating the feasibility of using scanner data for computing Spatial Price Indices at a detailed territorial level and at elementary aggregates;
- suggesting a framework for computing variance estimates of sub-regional Spatial Price Indices that uses Jackknife Repeated Replications when scanner data are collected following a probabilistic sample design.

For achieving these goals and explaining the potentialities of their approach, the authors provide an empirical application using scanner data managed by the Italian National Institute of Statistics – Istat in 2018, with a focus on Toscana region's ten provinces, for selected groups of products.

The second article presents the result of the fruitful collaboration of a group of researchers and experts from different institutions: Istat, Vodafone Business Italia, and Motion Analytica. The shared goals aim to evaluate potential uses of mobile phone data to strengthen current surveys and to investigate and produce broader results for official statistics.

The reference context is to develop the most appropriate methodologies to exploit mobile phone data by making them compliant with the definitions and classifications applied in official statistics, thereby meeting the requirements for robust estimate calculation. More specifically, the authors deal with topics related to the qualification and quantification of tourist flows, carrying out experimentations involving the Italian Province of Rimini and the Municipality of Roma, characterised by a strong tourist vocation. The data processed on

the activity of the mobile phone network come from Vodafone, which collects them continuously, ensuring information at very granular levels both in terms of territorial and temporal details.

The research described in this article also stands as an example of positive synergies between a private company holding certain kinds of data and a National Statistical Institute, thus inspiring further cooperation to develop and enrich this and additional fields of official statistics.

In conclusion, in the third article Gabriella Donatiello, Marcello D'Orazio, Doriana Frattarola, and Mattia Spaziani present an application of statistical matching methods for integrating the European Union Statistics on Income and Living Conditions and the Household Budget Survey.

The restructuring of the social statistics framework at European and National levels indeed enabled the use of integration techniques to exploit all information collected by existing data sources and to answer the growing demand for more detailed data.

The main objectives are measuring household income, consumption, and wealth at the micro level, and reducing both the National Statistical Institutes' production costs and the citizens' and households' response burden. For this purpose, the authors propose a change in a well-known approach to the complex sample surveys' statistical matching, with the aim of creating a synthetic dataset for in-depth multidimensional analyses of households' economic poverty in Italy.

The preliminary findings they describe are quite encouraging because they are related to an accurate ex *ante* harmonisation strategy of the reference surveys used, as well as to the collection of selected data, which proved to be particularly useful in the application of these re-designed methods.

Patrizia Cacioli
Editor

Nadia Mignolli
Coordinator of the Editorial Board

A model for measuring the accuracy in spatial price statistics using scanner data

Ilaria Benedetti[1], *Federico Crescenzi*[1], *Tiziana Laureti*[1]

Abstract

Given the crucial role of Spatial Price Indices (SPIs) for comparing standard of living across a country, there is a need to assess their accuracy on a regular basis. However, despite the importance attached to SPIs, with few exceptions reliability measures are not computed and published both at an international and national level. Focussing on SPIs, this paper aims at suggesting a measure for assessing SPIs accuracy through the adoption of a new approach based on the Jackknife replication technique. To illustrate the potentialities of the suggested approach an empirical application is provided using the 2018 Italian National Institute of Statistics - Istat scanner data on the ten provinces of Toscana for selected groups of products. Our results demonstrate a large spatial heterogeneity between the SPIs of the Toscana provinces and their standard errors.

Keywords: Scanner data, uncertainty, spatial price indices, price statistics, Jackknife replication method.

DOI: 10.1481/ISTATRIVISTASTATISTICAUFFICIALE_3.2022.01

1 Ilaria Benedetti (i.benedetti@unitus.it); Federico Crescenzi (federico.crescenzi@unitus.it); Tiziana Laureti (laureti@unitus.it), Università degli Studi della Tuscia/University of Tuscia - Viterbo, Lazio, Italy.

The authors wish to thank Antonella Simone (Istat) for her support and assistance.

The views and opinions expressed are those of the authors and do not necessarily reflect the official policy or position of the Italian National Institute of Statistics - Istat.

The authors would like to thank the anonymous reviewers for their comments and suggestions, which enhanced the quality of this article.

1. Introduction

The popularity and availability of transaction or scanner data for the compilation of Consumer Price Indices (CPIs) have increased over the past twenty years.

Consensus has emerged on the fact that, besides reducing the administrative burden and cost for both National Statistical Institutes (NSIs) and retailers, scanner data may allow the reduction of both sampling and non-sampling errors thanks to the detailed information available for individual products (product characteristics, quantity sold, etc.), the wide coverage both in terms of product groups and territorial areas, and the opportunity to implement superlative index based on product weights within elementary aggregates.

In addition, the availability of new data sources, such as scanner data and web-scraped data, have stimulated research for adopting more developed statistical techniques for constructing CPIs and for assessing their accuracy (Fenwick and Ball, 2001; Smith, 2021). The number of European NSIs using scanner data for CPI computations is steadily increasing and different choices are made regarding the index formula, which should be able to reduce chain drift bias and substitution bias. In the time domain context, in order to deal with the chain drift problem Ivancic *et al.* (2011) proposed the use of multilateral indices based on scanner data following the work of Balk (1981) who first noted that multilateral index methods, originally developed for price comparisons across countries, can be easily adapted to price comparisons across time. Since these pivotal studies, various multilateral methods have been suggested in literature following different approaches with the common characteristic that price indices are constructed simultaneously for the entire sample period as in the case of spatial price comparisons (Diewert and Fox, 2022).

However, in spite of their potential, to the authors' knowledge, only a few NSIs explored the use of scanner data to measure price level differences across region by computing sub-national spatial price indices (SPIs) which are crucial for assessing regional disparities in the distribution of real incomes and supporting regional policy-making (Rokicki and Hewings, 2019). These indicators, also known as sub-national Purchasing Power Parities (PPPs), are particularly important, especially in the case of EU member states where

regional economic analyses have become essential due to the implementation of EU Cohesion Policy promoting more balanced and sustainable territorial development.

Several European NSIs have been using scanner data for replacing on-field collected prices needed for international PPP computations in the framework of the OECD-Eurostat Programme. NSIs use scanner data to identify products and collect prices thus increasing the number of products priced (and assessing their representativity using turnover as weights) and expanding the number of cities where prices are collected (not only national capitals). In this way, it is also easier to compute Spatial Adjustment Factors as required by Eurostat, thus obtaining average national prices that are more representative of the whole country.

Although sub-national price comparisons are not as widespread as international price comparisons through the International Comparison Programme (ICP), several NSIs and individual researchers have conducted interesting research studies on the compilation of sub-national SPIs in various countries (Laureti and Rao, 2018). However, systematic attempts to compile sub-national SPIs on a regular basis have been hindered by the labour-intensive analyses required for processing traditional price data, *e.g.* data used for compiling CPIs, and by the costs involved for carrying out ad-hoc surveys for collecting price data. In this context, the use of scanner data is both a challenging yet feasible solution for solving the difficulties NSIs face when making spatial price comparisons worldwide. The Italian case study by Laureti and Polidoro in the recent Guide to the Compilation of Subnational Purchasing Power Parities (Biggeri and Rao*,* 2021) illustrates how scanner data can be blended with data from other sources (traditional collected prices, administrative and internet data) in the compilation of subnational SPIs covering the total household consumption expenditure.

The use of scanner data, which are often obtained through a probability sample design, may also allow to provide accuracy measures of point estimates of price differences across space. Uncertainty in the SPIs comes not only from the choice of aggregation procedure but also from the dispersion of relative prices. For countries across which relative prices are very different, the sampling of goods matters, and the PPPs and SPIs are much more uncertain (Deaton, 2012). This source of variation induces substantial uncertainty in

the spatial price indices. Measures of statistical errors for CPI and SPI have many uses: to inform on the quality for users, to guide CPI and SPI compilers in allocating resources for compilation in the most efficient way, and to detect possible serious errors in the data when output editing.

Yet, despite their importance, no standard errors or reliability measures are computed and published both at an international and national level (Deaton and Aten, 2017; Rao and Hajargasht, 2016; Smith, 2021). In view of the complexity of price index structures and the common use of non-probability sampling in compiling CPIs and SPIs, an integrated approach to variance estimation appears to be problematic. A single formula for measuring the variance of CPIs and SPIs, which captures all sources of sampling errors, may be impossible to find.

However, it is often possible to develop partial measures, in which only the effect of a specific single source of error is quantified. In recent years, there has been a growing concern for the more explicit use of the concepts and tools of statistical inference to produce estimates of official CPIs/SPIs and, especially, to define the targets of the estimates following a framework typical of statistical survey methods also known as sampling approach in index theory (Balk, 2005).

This paper contributes to the advancement of this literature by exploring the issue of evaluating the uncertainty associated to point estimates of sub-national SPIs, using the Italian scanner data for the year 2018 as the reference source. To the authors' knowledge, the evaluation of uncertainty among SPIs has not been explored yet.

The aim of this paper is twofold: firstly, we demonstrate the feasibility of using scanner data for computing SPIs at a detailed territorial level (NUTS-3) and at elementary aggregates by referring to the recent study carried out by Laureti and Polidoro (2022). Secondly, we suggest a framework for computing variance estimates of sub-regional SPIs that uses Jackknife Repeated Replications (JRR) when scanner data are collected following a probabilistic sample design. We also provide an application using a subset of the 2018 Italian National Institute of Statistics (Istat) scanner data where the sample of large-scale retail outlets is selected according to a probabilistic stratified random design from a universe of more than 9,000 outlets stratified by three variables: province (NUTS-3), distribution chain and outlet type

(supermarket or hypermarket). In particular, we consider the outlets of the ten provinces of Toscana (one of the Italian regions) and three group of homogeneous products, called basic headings (BHs), namely, Mineral water, Coffee and Pasta.

The remainder of this paper is structured as follows. Since the aim of this paper is not to provide a detailed literature review on research on spatial price comparisons, Section 2 reports the main methods that have been suggested and case studies in which scanner data have been used. In addition, a brief overview of the issue of measuring the accuracy of a consumer price index is provided. A description of Istat scanner data and our dataset is reported in Section 3, while the methodological approach to point and standard error estimates of SPIs are described in Section 4. Section 5 reports our results for Toscana and for each of the BHs considered. Section 6 draws some concluding remarks.

2. Literature Review

2.1 Scanner data and Spatial Consumer Price Indices

In countries characterised by large territorial differences in prices and quality of products and household characteristics, such as Italy, it is essential to calculate sub-national SPIs in order to assess inequality in the distribution of real incomes and consumption expenditures.

In all spatial price comparisons, the concept of PPP[2] is used to measure the price level in one location compared to that in another location; therefore, PPPs are essentially SPIs[3]. At international level, PPPs facilitate cross-country comparisons of Gross Domestic Product (GDP) and its major aggregates as they can be used in converting aggregates into a common currency. Likewise, sub-national PPPs allow for intra-country spatial comparisons and can serve as inputs and/or improve other inputs for estimating key economic indicators produced by countries, such as real regional price comparisons, real income dimensions and poverty estimates.

The process of compiling PPPs (or SPIs) is quite complex and is carried out in two stages. First, elementary spatial price indices are computed by aggregating, generally without using weights, prices of items belonging to a group of similar well-defined product goods or services (called Basic Headings, BHs). In the second stage, the elementary PPPs (or SPIs) are aggregated using expenditure weights to obtain PPPs (or SPIs) for higher-level aggregates such as consumption, investment and GDP.

Some NSIs and many researchers have conducted computations of subnational spatial price indices for household consumption. Indeed, given the limited resources, the goal of producing sub-national PPPs or spatial CPIs at a subnational level is most feasibly achieved using information from the national CPIs.

2 PPP for a given country represents the number of currency units required to buy a similar basket of goods and services in the given country in relation to a reference or base country). At international level, PPPs for countries are compiled by the ICP, administered by the World Bank with the collaboration of the OECD, EUROSTAT and other organisations (World Bank, 2013; Biggeri and Rao, 2021). A description of the framework adopted by the International Comparison Programme (ICP) is presented in Rao (2013) and the full set of ICP procedures are discussed in various chapters of World Bank (2013), "Measuring the Real Size of the World Economy: The Framework, Methodology, and Results of the International Comparison Programme (ICP)".

3 In this paper, we will use the term sub-national SPIs instead of PPPs as we refer to spatial comparison using scanner data for grocery products.

Early research on sub-national price comparisons was mostly conducted in the United States in the early 1990s (Kokoski *et al.*, 1999). These pioneering efforts at the Bureau of Labour Statistics and at the Bureau of Economic Analysis, were later continued by Aten (2008) finally leading to the regular compilation of spatial price differences in the United States through the computation of Regional Price Parities (RPPs). The Australian Bureau of Statistics started a research project for compiling SPIs. Similarly, Statistics New Zealand has been evaluating the possibility of carrying out spatial price comparisons of prices since 2005 and two experts have been assigned to develop a methodology for constructing spatial cost of living indices (Melser and Hill, 2007). However, to the authors' knowledge, estimates of sub-national SPIs have never been disseminated (Tam and Clarke 2015).

The General Statistical Office (GSO) of Vietnam started a pilot research project in 2010 to compute subnational PPPs in terms of the SCOLI (Spatial COst of LIving) index, supported by the World Bank, based on CPI data available. The GSO produced and published SCOLI indices for the period 2010 to 2017 (GSO, 2019).

At the beginning of the 2000s, the Italian National Institute of Statistics (Istat) conducted experiments on the use of CPI data to calculate regional consumer price level indices and disseminated results on two occasions: in 2008 with reference to price data from 2006; and in 2010 with reference to 2009 data (Istat, 2008 and 2010).

In France, the National Institute of Statistics and Economic Studies (INSEE) conducted *ad hoc* price surveys in 1985, 1992, 2010, and 2015 and published analyses based on these surveys (INSEE, 2019).

As far as the estimation of sub-national consumer spatial price index numbers using CPI data and/or COICOP classification, numerous attempts have been made by researchers in various countries (Janský and Kolcunová, 2017; Biggeri *et al*, 2017*a*; Chen *et al*, 2019; Rokicki and Hewings, 2019; Weinand and von Auer, 2020; Biggeri *et al*, 2017*b*; Montero *et al.* 2019; Fenwick and O'Donoghue, 2003; Roos, 2006, Kocourek *et al.*, 2016).

Recently, several NSIs around the world have been integrating price data collected from traditional sources in the process of compiling official CPIs with other new sources of data. Based on the experienced gained, for example in

Italy, scanner source of data makes it possible to identify representative and comparable products across subnational areas, resolving the important issue of balancing the two competing requirements of representativity and comparability. Indeed, when compiling SPIs, the first step is the definition of a product list consisting of goods and services that are to be priced which should adequately cover goods and services purchased in different regions of the country.

To this respect, scanner data allow for the construction of fine level SPIs as the volume of information contained in such data makes it more likely for spatial price comparisons to be made on the same product across geographical areas and it allows using information on sales for obtaining weights at individual product level. Indeed, GTIN codes uniquely distinguish products, and they are generally the same for each item at national level. Thanks to the high territorial coverage which characterises scanner data, it is possible to compare price levels at different territorial levels within a country (NUTS-3, NUTS-2 and NUTS-1). In this way, the issue of comparability can be solved.

Moreover, scanner data may allow to provide information on quality characteristics that may influence the price of a product, such as the chain or the type of outlet in which the product is sold. It is also possible to add a time dimension to multilateral spatial price comparisons since detailed data are usually available at the point of sale and usually on a weekly basis. In addition, using transaction data it is possible to account for the economic importance of each item in its market by using data on turnover, thus providing a reliable indicator of the importance of individual products. Finally, as already mentioned, using scanner data to carry out spatial comparisons will result in cost efficiencies since price data collection can then be limited to traditional outlets thus lowering data collection costs for the NSIs.

When it comes to the usage of scanner data in computing SPIs however we notice that scanner data have been mainly used for making international price comparisons (Heravi *et al.*, 2003; Feenstra *et al.*, 2017) leaving its usage for sub-national SPIs unexplored.

A first attempt in this context is the study by Laureti and Polidoro (2018) who estimated Italian SPIs for 2017 using a scanner dataset constructed for experimental CPI computation. Later, Laureti and Polidoro (2022) demonstrated the feasibility of using scanner data for comparing consumer prices at different territorial levels, which are representative of local

consumption patterns and comparable based on a set of prices determining characteristics. The encouraging results obtained stimulated further research by Istat to achieve the aim of producing sub-national SPIs for Italy on a regular basis through the adoption of a multi-source approach to cover all the retail trade channels and product baskets.

2.2 Addressing the issue of accuracy in Spatial Consumer Price Indices

The issue of measuring the accuracy of temporal and spatial consumer price indices has a long history and it is still controversial. Indeed, it has not been clear if it is feasible to make such assessments. The underlying population price index, for even the smallest of countries, involves so many transactions on so many individual products in so many places as to be inaccessible. In addition, usually, the universe of individual products on the market is dynamic, thus introducing additional difficulty in the definition of an appropriate sampling design. As a result, most official price index numbers are published without an explicit statement about their accuracy. Therefore, the situation has not changes since Morgenstern's remark on this issue "In spite of the widespread use of government price indices, there has been little done in attempting to determine the error inherent in these indices" (Morgenstern, 1963).

However, during the last years, there has been a growing interest in a more explicit evaluation of the accuracy of CPIs thanks to the availability of new data sources (De Haan *et al.*, 1999; Fenwick *et al*, 2001). However, since CPI and SPI are not in general obtained from a single survey, the sampling and non-sampling errors, being related to all the surveys used for the construction of the index, cannot be easily specified by a single complex model.

The pioneering works on this subject date back to Banerjee (1956) and Adelman (1958), while systematic research on the sampling variance of CPI indices has been developed since the mid-eighties (Balk and Kersten, 1986; Balk, 1989). In the same framework, Biggeri and Giommi (1987) suggested a structure for the classification of errors within a price index based on a breakdown of the mean squared error.

Several approaches have been used to estimate the variance of CPIs (Smith, 2021). Model-based and design-based approaches have both been applied for CPI variance calculation.

More specifically, design-based approaches with Taylor linearisation have been adopted by several authors in the literature. Andersson *et al.* (1987a) used this approach to assess the variance due to sampling outlets in the Swedish CPI, Leaver and Valliant (1995) used this approach in the US context in order to give an approximation of the variance of the Laspeyres index between two time periods. Despite several applications, there is a suggestion that Taylor linearisation may underestimate the empirical variances for smaller sample sizes (see, *e.g.* Andersson *et al.*, 1987b). Moreover, the computation of Taylor linearisation estimator may be extremely complicated, and may require several layers of approximation, whose aggregate effect is not clear without detailed investigation (Smith, 2021).

A replication-based approach is the longest-standing approach in the computation of standard errors for the US CPI (Leaver *et al.*, 1991). The drawback of this method is related to the resampling procedure. Indeed, if the sampling procedure generates small samples, this approach can produce large variances

The Jackknife and the Bootstrap methods have become increasingly popular for CPI variance estimation since they can be used in complex designs. The Jackknife is already in use in the US for special item categories. Leaver and Cage (1997) discussed the implementation of the Jackknife in the US CPI by comparing variance estimates for a series of alternatively aggregated prices. Klick and Shoemaker (2019) used the Jackknife to evaluate significant differences between urban population and other populations (wage earner and clerical worker population, and elderly population) in some cities in the US.

Some authors have used a model-based approach to estimate the price index formula variance. For example, Kott (1984) modelled the variance of the Laspeyres type index where the prices are assumed to be "nearly homogeneous". The key advantage of this approach is that sampling weights are not required to obtain the estimates, on the contrary, it is more difficult to justify a particular model and to claim objectivity for the variance estimates.

During the last years, various research studies have been carried out for adopting probability sampling design and assessing CPI accuracy by using simulation data and new sources of data (see, for example, De Gregorio, 2012; De Vitiis *et al.*, 2017). De Gregorio (2012) analysed the sample sizes needed to estimate Laspeyres consumer price sub-indices under a combination of

alternative sample designs, aggregation methods, and temporal targets using simulated data. The author found that the optimal sample size depends crucially on the degree of relative variability and skewness of items and underlined the crucial role of stratification in saving sample size. De Vitiis *et al.* (2017) studied the properties of alternative aggregation formulas of the elementary price index in different sampling schemes implemented on scanner data. Bias and efficiency of the estimated indices are evaluated through a Monte Carlo simulation.

At official level, only the Swedish NSI provides official variance estimates for CPIs. Official statistics published by Statistics Sweden are disseminated with a quality declaration per survey year and the sampling uncertainty measures are assessed annually for monthly change, annual change (inflation rate), and monthly change in the inflation rate (Norberg and Tongur, 2022).

In this context, scanner data may play a crucial role in stimulating research on sampling errors by improving sampling methods, checking the representativity of the achieved sample, and controlling initial sample selection. Recently, Tongur (2019) focussed on the case of scanner data for daily consumer products and their inclusion in the Sweden CPI, particularly regarding the issue of the trade-off between item related variance and the bias from disregarding explicit quality adjustments. Results show that the contribution to the variance from a randomly sampled item in the daily products survey is rather small and would tend to decrease with appropriate sampling, given that the samples are based on size-proportional sampling strategies. The sample size related variance is estimated through a Jackknife method.

It is worth noting that scanner data may also reduce non-sampling errors. Among these types of errors, it is possible to distinguish between measurement errors, representativeness of items and coverage errors. The firsts are mitigated thanks to the increased number of products priced and to the improved territorial and population coverage. Indeed, prices may be collected in each city across the province and not only in the provincial capital. In contrast, the traditional basket is a relatively small subset of the complete universe of goods while quantities sold are not available. Moreover, the use of unit value (calculated as the total expenditure for that item code divided by the total quantities sold) instead of price, represents a more accurate measure of the actual price of an individual product than an isolated price quotation (Balk,1995).

In addition, by using scanner data, it is possible to consider a wide range of methods for calculating spatial price indices due to the availability of quantities and expenditure information (Heravi *et al.*, 2003) which is usually not collected in traditional surveys. In fact, information on expenditure and quantities allows to calculate indices based on a variety of "superlative" index number formulae, including the Fisher ideal index (see Imai *et al.*, 2015; Laureti and Polidoro, 2018). This may reduce the so-called formula error (Dorfman *et al.*, 2006).

According to the previous literature review, this paper offers an advancement to the literature on spatial price index variance estimation by applying the Jackknife as a variance estimation technique to scanner data. To the authors' knowledge, the use of the Jackknife technique for evaluating uncertainty among spatial price indices computed for geographical areas within a country has not yet been explored.

Although in the context of CPI both design-based and model-based estimators have been used, the Jackknife method has been selected since it can be easily applied to the scanner data sampling design by setting BHs as strata and market classification as Primary Sampling Units (PSU). As a result, Jackknife estimates have an interpretation in terms of the error arising from the sampling processes for prices, and the proposed stratification variables isolate possibly homogeneous product groups and clusters of pricing policies.

3. Italian scanner data for computing Spatial Price Index

In this paper, we use a scanner dataset provided by Istat for the year 2018[4]. Since January 2018, the Italian NSI has introduced in its consumer price production process the use of scanner data from large-scale retail trade/ modern distribution chains (hypermarkets and supermarkets) for grocery products (packaged food, household, and personal care goods).

Istat acquires data, through a market research company, for individual outlets of 16 large-scale retail groups in Italy for all 107 provinces of the national territory by type of outlets (hypermarkets and supermarkets).

The sample of large-scale retail trade outlets is representative of the entire universe of large-scale retail trade and includes 1,781 outlets, of which 510 hypermarkets and 1,271 supermarkets distributed throughout the country.

The 16 chains collaborating with Istat represent, at national level, more than 90% of the total turnover of hypermarkets and supermarkets, with a high coverage also at regional level (the highest coverage value is recorded in Toscana with 99.9% and the lowest in Basilicata with 67.9%).

Italian scanner data cover all grocery products for a total of 79 product aggregates, belonging to five ECOICOP[5] divisions (01, 02, 05, 09, 12) and substitute on field price collection. The individual products included in the index calculation are identified by codes (GTINs), which uniquely identify products (items) throughout the country.

The sample of large-scale retail outlets is selected according to a probabilistic stratified random design. The universe, made up of over 9,000 outlets, is stratified by considering three variables: the province (all 107 provinces), the chain or group to which it belongs (16 large-scale retail trade distribution) and the outlet type (supermarket or hypermarket). The sampled outlets are extracted within each of the 888 strata of the universe with probability proportional to the sales turnover of the previous year. Since scanner data do not provide the "shelf price" of the product but quantity sold the purchasing price should be defined as unit value of average weekly price.

4 The scanner data set was provided by Istat and treated in order to respect the statistical confidentiality. The authors worked at Istat to perform the various analyses.

5 European Classification of Individual Consumption according to Purpose (ECOICOP).

In our analysis, the unit value price for each individual product/GTIN is calculated using turnover and quantities sold (unit value = turnover/quantity) of each province obtained as a weighted sum of outlet sampling weights. All the products sold in each sampled outlet are selected within homogeneous groupings of products corresponding to the markets. Probabilities of selection were assigned to each outlet based on the corresponding turnover value. The classification of homogeneous products within markets represents an objective and detectable identification of commodity products shared by industrial and distribution companies. This classification is provided by the Efficient Consumer Response (ECR) community, used for category management by both industrial and distribution companies. In the Italian case, the market classification is made available by the Nielsen company, but in general they could be retrieved through statistical registers of enterprises and local units.

The ERC classification is based on a hierarchical structure. Each of the defined categories is linked to a reference sheet that contains the definition of the category and the criteria for exclusion and inclusion of products. As an example, for the Water BH one of the possible markets is "sparkling water". Classification includes the following description: "Natural water from mineral water sources, with various chemical and therapeutic properties. They have carbon dioxide added and are labelled 'carbonated' or 'sparkling'". Only mineral waters with added carbon dioxide called carbonated or sparkling are included. On the contrary non-carbonated mineral waters (natural or still), natural sparkling waters and lightly or slightly carbonated waters are excluded. The classification reports also some examples of inclusion, *e.g. "Boario gassata", "San Pellegrino frizzante"*, and example of exclusion: *e.g. "Ferrarelle effervescente naturale"*.

Provincial prices are calculated as weighted arithmetic mean according to the probability proportional to size (PPS) sampling design within each stratum.

In our paper, we refer to a portion of this big dataset since we used 2018 data for all the outlets sampled for Toscana region, assembled after the stratification process. To illustrate the potential of the suggested methodology, in this analysis we considered three BHs namely: Mineral water, Coffee, and Pasta. The dataset consists of 8,856 annual price quotes from the ten Toscana provinces, which make up Toscana: Arezzo (AR), Firenze (FI), Grosseto (GR), Livorno (LI), Lucca (LU), Massa-Carrara (MS), Pisa (PI), Prato (PO), Pistoia (PS) and Siena (SI).

4. Methods

4.1 Computation of Spatial Price Index among provinces

We estimate provincial SPIs at BH level for the 10 Toscana provinces. As a first step, we compute within-province SPIs using the Fisher and the Törnqvist formulas, as they are known to be superlative (Diewert, 1976). In particular, the Fisher type formula is used for computing international SPIs also by the Eurostat-OECD (2012).

The Fisher index is an "almost ideal" index since it satisfies all the standard properties and tests except the transitivity test and approximates a cost-of-living index (Diewert, 1976). Instead, the Törnqvist index satisfies the country reversal test, and it is superlative and exact. The two indices are very close to each another due to the similarity in their definitions (Pilat and Rao, 1996). In the second step, we compute a multilateral index by combining bilateral comparisons between pairs of provinces.

When constructing multilateral SPIs is important to satisfy two basic properties: transitivity and base invariance (province invariance). Transitivity requires that the SPI computed between two provinces should be the same whether it is computed directly or indirectly through a third region. Base invariance means that all provinces be treated equally in deriving the matrix of SPIs that satisfy transitivity.

Let us assume that we are attempting to make a spatial comparison of prices between M provinces. At BH level, p_{ij} and q_{ij} represent the price and quantity of the *i-th* item in the *j-th* province with $i=1,2,\ldots,N$ and $j=1,2,\ldots, k,\ldots, M$. In most cases, the price indices for elementary aggregates are calculated without the use of explicit weights. Contrastingly, our scanner data set contains detailed quantity and sales information within an elementary aggregate so that there are no constraints on the type of index number that may be used.

We calculate matched-model indices, *i.e* Laspeyres and Paasche indices, between areas j and k. Since not all items may be priced in all areas included in the comparison, N_{jk} represents the number of products that are priced in both areas j and k. Therefore, N_{jk} is usually smaller than the number of commodities N in the basic heading. If a commodity is not priced in one of the two provinces,

that item cannot be included in the SPI computation. Because SPI_{jk}^{P} and SPI_{jk}^{L} use only information on areas j and k from the price tableau, the resulting indices are not transitive.

In a spatial framework, the Laspeyres index measures the change in the fixed basket cost, taking a base province as a reference, where substitution is not considered when there is a change in relative prices (Paredes Araya and Iturra Rivera, 2013).

As known, the Laspeyres and Paasche indices can be calculated in two ways: either as the ratio of two value aggregates or as an arithmetic weighted average of the price ratios for the individual products using the hybrid expenditure shares as weight:

$$SPI_{jk}^{L} = \frac{\sum_{i \in N_{jk}} p_{ik} \cdot q_{ij}}{\sum_{i \in N_{jk}} p_{ij} \cdot q_{ij}} \equiv \sum_{i \in N_{jk}} \left(p_{ik}/p_{ij}\right) s_{ij} \tag{1}$$

$$SPI_{jk}^{P} = \frac{\sum_{i \in N_{jk}} p_{ik} \cdot q_{ik}}{\sum_{i \in N_{jk}} p_{ij} \cdot q_{ik}} \equiv \sum_{i \in N_{jk}} \left[\left(p_{ik}/p_{ij}\right) s_{ik}{}^{-1}\right]^{-1} \tag{2}$$

For each pair of provinces, we calculate the two bilateral SPIs, that is SPI_{jk}^{P} and SPI_{jk}^{L}, using expenditure shares equal to s_{ij} and s_{ik} respectively. These expenditure shares are computed using only the common goods in each pair of provinces. Therefore s_{ij} represents the expenditure share of *i-th* item in the province *j-th*, while s_{ik} represents the expenditure share of *i-th* item in the province *k-th*.

By following this procedure each BH is provided with a matrix of Fisher SPIs. The Fisher price index, which has good axiomatic and economic properties (Balk, 1995), is a geometric average of the Laspeyres and Paasche indices given by:

$$SPI_{jk}^{F} = \sqrt{SPI_{jk}^{L} \cdot SPI_{jk}^{P}} \tag{3}$$

From an economic approach, the Fisher index is preferred as it uses quantities at different times and allows for substitution effects. Since the Fisher SPIs, SPI_{jk}^{F}, are not transitive, the Gini-Éltető-Köves-Szulc (GEKS) methodology is used to obtain a transitive index that deviates the least from

a given matrix of binary comparisons. The GEKS SPIs can be obtained as an unweighted geometric average of the linked (or chained) comparison between provinces j and *k* using each province *l* in the comparisons as a link:

$$SPI_{jk}^{GEKS-F} = \prod_{j=1}^{M}\left[SPI_{jl}^{F} \cdot SPI_{lk}^{F}\right]^{1/M} \quad (4)$$

The second method that has been used for computing sub-national SPIs is the Törnqvist spatial price index, that is defined as a geometric average of the price relatives weighted by the average expenditure shares in the two provinces *j* and *k*:

$$SPI_{jk}^{T} = \prod_{j=1}^{M}\left(\frac{p_{ik}}{p_{ij}}\right)^{\frac{s_{ij}+s_{ik}}{2}} \quad (5)$$

Where $\frac{s_{ij}+s_{ik}}{2}$ is the arithmetic average of the share of expenditure on item i in two provinces. Fisher and Törnqvist are superlative indices and show up as being "best" in all the approaches to the index number theory as they satisfied all the axiomatic properties expected of a price index number formula with the exception of the circularity test (Hill, 2004; ILO, 2020). More specifically, they satisfy the base invariance test and commensurability test; they are symmetric indices given that they make equal use of prices and quantities in both the areas compared and treat them in a symmetric manner. Within the economic approach, Diewert (1976) obtained a characterisation of the Törnqvist price index, as being the economic price index that corresponds to a linearly homogeneous translog unit cost or revenue function.

Unfortunately, similarly to Fisher, the Törnqvist SPIs express in equation (5) are not transitive. To obtain a set of transitive SPIs, the bilateral SPIs must be transitivised using the GEKS procedure:

$$SPI_{jk}^{GEKS-T} = \prod_{j=1}^{M}\left[SPI_{jl}^{T} \cdot SPI_{lk}^{T}\right]^{1/M} \quad (6)$$

The Törnqvist version of GEKS is often referred to as the CCD method (Caves *et al.*, 1982). One attractive feature of the CCD method is that it can also be represented as a star method with an artificial country at the centre of the star (Hill and Timmer, 2006).

4.2 Addressing the issue of accuracy in Spatial Price Indices

Originally introduced as a technique of bias reduction, the Jackknife Repeated Replication (JRR) method has by now been widely tested and used for variance estimation (Durbin, 1959). Like other resampling procedures, the JRR method estimates the sampling error from comparisons among sample replications which are generated through repeated resampling of the same parent sample. Each replication needs to be a self-representative sample and to reflect the full complexity of the parent sample. The JRR variance estimates consider the effect on variance of aspects of the estimation process which are allowed to vary from one replication to another. In principle, these can include complex effects such as those of imputation and weighting. The basic JRR model which shall be adopted in this work can be summarised as follows. Consider a design in which two or more primary units have been selected independently from each stratum in the population. As in the case of the linearisation approach, sub-sampling of any complexity may be involved within each PSU, this does not affect the variance computation formulae. In the standard delete one-PSU at a time Jackknife version (Leaver and Cage, 1997), each replication is formed by eliminating one sample PSU from a particular stratum at a time and increasing the weight of the remaining sample PSUs in that stratum appropriately to obtain an alternative but equally valid estimate to that obtained from the full sample. This procedure involves creating as many replications as the number of primary units in the sample.

Let r be a subscript to indicate a sample PSU and let h indicate its stratum; moreover, let $a_h \geq 2$ be the number of PSUs in stratum h, assumed to be selected independently. Let λ be a full sample estimate of any complexity, and $\lambda_{(hr)}$ the estimate obtained after eliminating primary unit r in stratum h and increasing the weight of the remaining $a_h > 1$ units in that stratum. Also, let $\lambda_{(h)}$ be the simple average of the $\lambda_{(hr)}$ over the a_h values of r in h. The variance of λ is then estimated as (Betti *et al.*, 2018):

$$var(\lambda) = \sum_h [(1 - f_h)\left(\frac{a_h - 1}{a_h}\right)\sum_r \left(\lambda_{(hr)} - \lambda_{(h)}\right)^2] \quad (7)$$

Where $1 - f_h$ is the finite population correction which in typical social surveys is approximately equal to 1.

Under quite general conditions for the application of the procedure, the

same and relatively simple variance estimation in Formula (7) holds for λ of any complexity. This in fact is the major attraction of the JRR method for practical application.

De Gregorio (2012) showed that the choice of the strata is of paramount importance, since it involves theoretical and microeconomic issues, including for example market criteria, to isolate possibly homogeneous product groups and clusters of pricing policies. Our dataset considers 8,856 observations, subdivided in 3 Strata (BHs) and 46 PSUs (the markets).

Therefore, by conditioning on each BH, we use the JRR to estimate the uncertainty in SPIs due to the allocation of products in Markets. This is done by substituting the λ-s in Formula (7) with their counterparts illustrated in Formulas (4) and (6) respectively.

5. Results

In our analysis, we consider a total of 8,856 price observations referring to a total of 1,726 unique products. Table 5.1 reports the number of observations, number of unique products and number of markets in each BH.

In Figure 5.1, we compare price variability observed across Toscana provinces in the three BHs included in our analysis. The highest price variability value is reported for Coffee products where the coefficient of variation (CV) is equal to 96%. Contrastingly, Pasta products show similar prices as the value of the CV is equal to 59%.

Considering the variability within each product group, it is worth noting that, among the various provinces, for the Mineral water product group the highest heterogeneity in price levels is observed for Pistoia and Lucca provinces (CV equal to 98% and 97% respectively). The highest values of price variability for Pasta products are observed in Lucca and Firenze (CV equal to 62% and 59% respectively), while for the Coffee BH, Firenze and Pisa show the highest levels of price variability (CV equal to 64% and 63% respectively).

Table 5.1 - Number of observations, GTIN and markets in each stratum

BHs	N. of Observations	N. of GTIN	N. of Markets
Mineral water	2,010	298	13
Coffee	1,184	337	12
Pasta	5,662	1,091	21

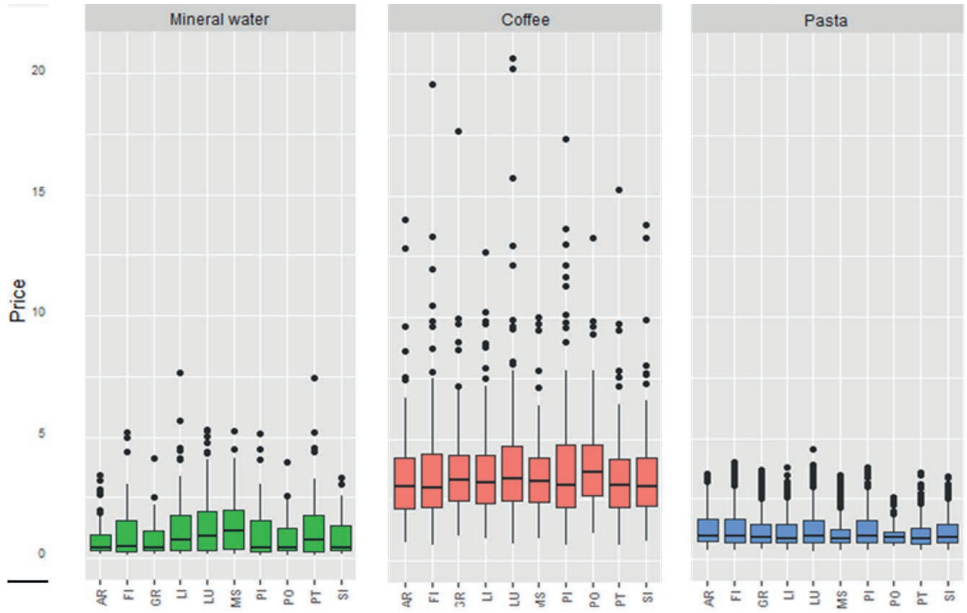
Source: Authors' elaboration from Italian Scanner Data

To ensure comparability among products sold in two provinces, individual products are matched by using GTIN codes, which contain elementary information and identify unique products.

For price comparisons at the product level, we identified 1,091 different individual products for Pasta, 337 different products for Coffee and 298 different products for Mineral water that are priced in at least two different provinces. Table 5.2 reports the number of different individual products (GTINs) sold in each province of Toscana by group of products (BH) and outlet type. This Table shows that there is a large variety of products over provinces that can be explained by the fact that the large-scale retail trade

distribution is not uniformly distributed across the Toscana territory in terms of types of outlets, retail chains and market share. Unfortunately, our data do not allow us to analyse the distribution of retail chains among the provinces in question due to confidentiality constraints. In addition it is worth noting that consumers may purchase different individual products due to different consumption behaviours strictly related to the city in which they live. While well-known brands are sold in all the Toscana provinces (*e.g.* "De Cecco", "Barilla", etc.) other local-produced pasta are sold in only few provinces (*e.g.* "Antichi Poderi Toscani").

Figure 5.1 - Price distributions in Toscana provinces by groups of products: Mineral water (left), Coffee (centre), Pasta (right)



Source: Authors' elaboration from Italian Scanner Data

Table 5.2 - Number of GTINs priced in each province for each product groups

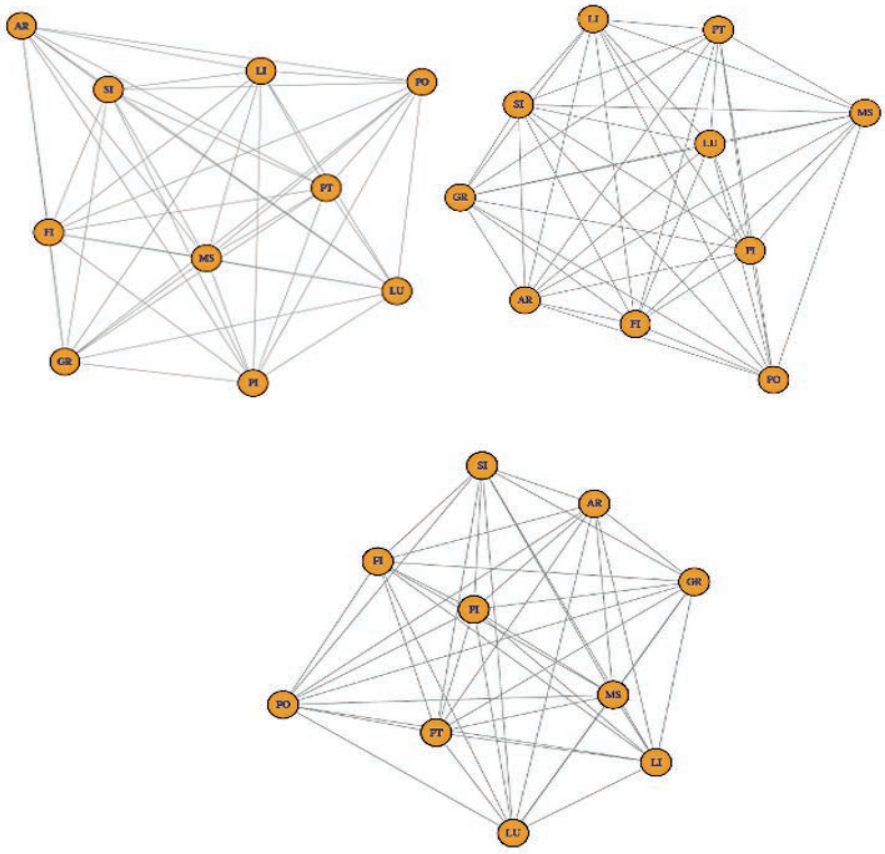
Province	Basic Heading		
	Mineral water	Coffee	Pasta
Arezzo	103	131	500
Firenze	159	176	470
Grosseto	112	110	354
Livorno	210	148	538
Lucca	193	194	611
Massa-Carrara	168	103	327
Pisa	142	172	437
Prato	78	76	178
Pistoia	177	141	485
Siena	87	108	352
Total	298	337	1091

Source: Authors' elaboration from Istat scanner data 2018

Given that not all the GTINs are priced in all the provinces, our data lead to an incomplete tableau of prices. In this case, it is important that the collected price data are connected to allow price comparison between all the areas involved (World Bank, 2013). For binary SPIs, little overlap in the products priced by the two provinces implies that the two geographical areas are very different and, by implication, inherently difficult to compare (Hill and Timmer, 2006).

Figure 5.2 demonstrates that reliable price comparisons can be carried out by illustrating the existence of the links among all the provinces. Indeed, we can state that our price data for the three product groups in question are connected among provinces and therefore it is not possible to place the provinces in two groups in which no GTINs sold by any province in one group is sold by any other province in the second group. We note multiple links between the same two nodes, even if not all the provinces are directly linked. Left panel of Figure 5.2 reports an interesting case in which the Pistoia (PT) province is not directly linked with Grosseto (GR) province. However, four indirect links exist. For example, one of the links compares prices across Pistoia and Firenze (FI) and then across Firenze and Grosseto. Therefore, in this case, through indirect links it is possible to make reliable price comparisons across provinces (World Bank, 2013).

Figure 5.2 - Figure 5.2 Product links among provinces. Mineral water (left), Coffee (right), Pasta (bottom)



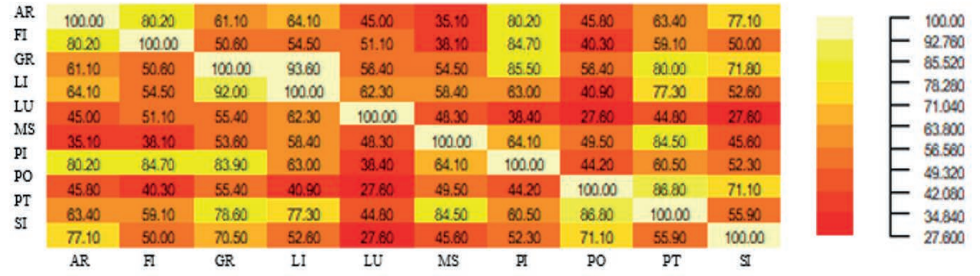
Source: Authors' elaboration from Istat scanner data 2018

In order to evaluate the strength of the overlaps among provinces, Figures 5.3, 5.4 and 5.5 show the product-matched matrices calculated for the 10 provinces for the aggregates included in our analysis. Each value in the matrix reports the matching products sold in two provinces. As an example, Figure 5.3, which reports the matrix for Mineral water, illustrates that 71.40% of products sold in Firenze are also purchased in the province of Arezzo. Considering Pasta products, the highest overlap is observed between Prato and Pistoia (97.75%), while the minimum overlap is observed between Pistoia and Lucca (25.90%).

Figure 5.3 - Product-matched matrix for Mineral water BH across provinces

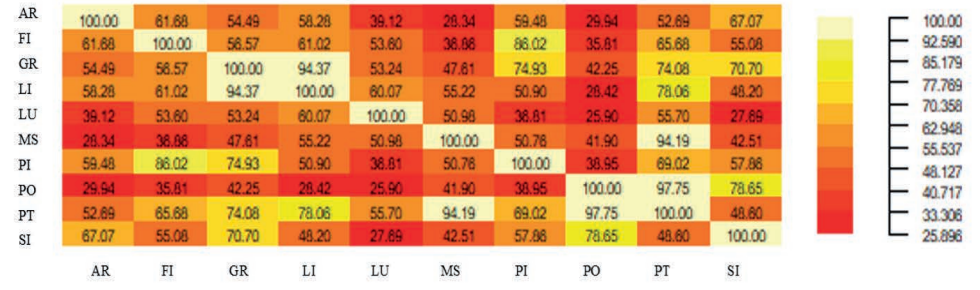
Source: Authors' elaboration from Istat scanner data 2018

Figure 5.4 - Product-matched matrix for Coffee BH across provinces



Source: Authors' elaboration from Istat scanner data 2018

Figure 5.5 - Product-matched matrix for Pasta BH across provinces



Source: Authors' elaboration from Istat scanner data 2018

SPIs point estimates obtained with GEKS-Fisher and GEKS-Törnqvist methods are reported in Table 5.3, SPIs are computed by considering Firenze province as reference (Firenze=100). As already mentioned, Fisher and Törnqvist price indices are symmetric and allow for substitution effect. Fisher price index measures the change in the expenditure function described by the Cobb-Douglas function, while Törnqvist price index measures the change in the expenditure function in a Translog utility function. Results reported in Table 5.3 reveal that Fisher and Törnqvist indices are very close to each other as shown by Diewert (1978). In addition, the Törnqvist index is approximately bounded by the Laspeyres and Paasche indices (Reinsdorf *et al.*, 2002).

The territorial heterogeneity among Toscana SPIs provinces is highlighted in Figure 5.6. From our results, it is possible to observe that Firenze is the less expensive province for Pasta and Coffee BHs, while for the Mineral water product group the less expensive area is Siena, followed by Arezzo and Prato.

It is interesting to note that concerning the Mineral water product group, Massa-Carrara is found to be the most expensive province (SPIs equal to 104.09 estimated with GEKS-Fisher and 104.07 estimated with Törnqvist-GEKS), followed by Grosseto (SPIs equal to 101.95 estimated with GEKS-Fisher and 101.92 estimated with Törnqvist-GEKS). For Coffee product group, the provinces of Grosseto and Siena proved to be the most expensive provinces according to the two procedures. The SPIs for Grosseto are equal to 105.35 with GEKS-Fisher and 105.20 with Törnqvist-GEKS, while the SPIs for Siena are equal to 105.13 with GEKS- Fisher and 105.10 with Törnqvist-GEKS. Interestingly, for the Pasta product group, the most expensive province is Siena (SPIs equal to 104.09 estimated with GEKS-Fisher and 104.07 estimated with Törnqvist-GEKS).

Table 5.3 reports the standard errors (in italics) obtained using JRR for the GEKS-Fisher and the Törnqvist-GEKS SPIs for each BH[6]. These are similar to each other although the GEKS-Fisher procedure has standard errors slightly greater than the Törnqvist-GEKS.

The highest standard errors are observed for Mineral water product group estimates obtained with the Fisher-GEKS procedure: the standard errors range from 0.003 in Pisa province to 0.049 in Grosseto province. By focussing on

6 These correspond to the inner summation in Formula (3).

Coffee product group, the highest standard error computed with Fisher-GEKS procedure, is observed for Grosseto (0.029) while the lowest standard error is observed for Pistoia (0.003). The Pasta product group has the lowest standard errors among the groups considered either for the Fisher-GEKS and the Törnqvist-GEKS. This is reasonably related to the high number of products and markets included in the analysed BHs. High standard errors may be due to the different territorial distribution of types of outlets, retail chains and market share in the Toscana territory.

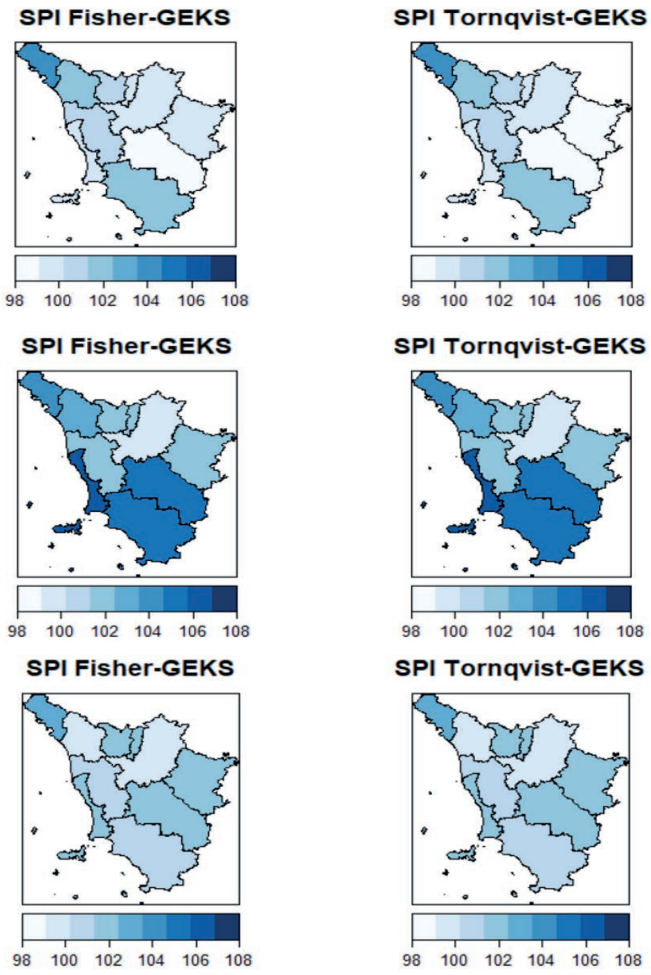
On the basis of the JRR replications, we derived 95% confidence intervals for each Provincial SPI. Confidence intervals overlap for the provinces of Livorno and Lucca for the three product groups analysed in this paper, thus suggesting that the SPIs in these two provinces are not significantly different from each other. As an example, for Mineral water product group, the confidence interval for Livorno province at 95% becomes [101.52; 101.68] and for Lucca province, the confidence interval becomes [101.59; 101.62].

Table 5.3 - Fisher-GEKS and Törnqvist-GEKS SPIs estimates. Standard errors in italics (Firenze=100)

Province	Fisher-GEKS			Törnqvist-GEKS		
	Basic Heading			Basic Heading		
	Mineral water	Coffee	Pasta	Mineral water	Coffee	Pasta
Arezzo	99.12	101.58	102.25	99.09	101.53	102.23
	0.010	*0.014*	*0.002*	*0.009*	*0.013*	*0.002*
Grosseto	101.95	105.35	100.95	101.92	105.20	100.94
	0.049	*0.029*	*0.020*	*0.048*	*0.028*	*0.020*
Livorno	101.61	102.65	100.18	101.55	102.66	100.21
	0.045	*0.022*	*0.018*	*0.044*	*0.021*	*0.018*
Lucca	101.61	102.65	100.18	101.55	102.66	100.21
	0.008	*0.014*	*0.007*	*0.008*	*0.014*	*0.007*
Massa-Carrara	104.09	104.50	102.80	104.07	104.40	102.82
	0.038	*0.018*	*0.014*	*0.038*	*0.018*	*0.014*
Pisa	100.45	102.02	100.70	100.46	101.99	100.70
	0.003	*0.008*	*0.005*	*0.003*	*0.008*	*0.005*
Prato	99.17	101.81	101.37	99.16	101.76	101.35
	0.010	*0.007*	*0.008*	*0.009*	*0.007*	*0.008*
Pistoia	100.63	102.41	102.21	100.61	102.34	102.19
	0.004	*0.003*	*0.006*	*0.004*	*0.003*	*0.006*
Siena	98.72	105.13	105.31	98.67	105.01	105.24
	0.028	*0.017*	*0.007*	*0.028*	*0.017*	*0.007*

Source: Authors' elaboration from Istat scanner data 2018.

Figure 5.6 - Fisher-GEKS and Törnqvist-GEKS SPIs estimates Mineral water, Coffee, Pasta. Firenze=100



Source: Authors' elaboration from Istat scanner data 2018

6 Conclusions

Scanner data has received considerable attention from statistical agencies during recent years since they proved to have significantly improved the efficiency of traditional price collection techniques in that they contain transactions on the goods sold, the prices paid by consumers, and the quantities sold for each item code or GTIN (Laureti and Polidoro, 2018). Scanner data may help to overcome the issue of price data availability in the various areas involved in spatial price comparisons thus fulfilling the requirements of representativeness and comparability that emerge when compiling sub-national SPIs. Due to the high territorial coverage, which characterises scanner data, it is possible to compare price levels at different territorial levels within a country (NUTS-3, NUTS-2, and NUTS-1). In this paper, we provided point estimates of SPIs of three basic headings for the provinces of Toscana in Italy using scanner data provided by Istat. Our results reveal that Fisher and Törnqvist-based GEKS numerically approximate each other and the two index methods tend to coincide, even if in some cases the SPIs estimated with Fisher-GEKS procedure are slightly greater than those obtained with Törnqvist-GEKS. This means that the unit elasticity of substitution implied by the geometric formula seems to overestimate the extent to which consumer responds to process changes relative to the market level in each product group based on ECR classification.

Along with point estimates, we used Jackknife Repeated Replications to provide an estimate of the associated standard errors with. In the introduction of the paper, we outlined that the measurement of price level differences across regions within a country is essential for assessing inequalities among populations residing in different parts of a country, for example in the distribution of real incomes (Laureti and Rao, 2018). Consequently, assessing the uncertainty of point estimates of price differences is essential for a better understanding of the phenomenon that we are investigating. In this paper we estimated the variance using the Jackknife estimator. However, the Jackknife depends on the sampling design and therefore to have more accurate estimates of SPIs it may be of crucial importance reconsidering the sampling design. On the basis of the JRR replications, we derived 95% confidence intervals for each Provincial SPI. Confidence intervals overlap for the provinces of Livorno and Lucca for the three product groups analysed in this paper, thus

suggesting that the SPIs in these two provinces are not significantly different from each other. Moreover, our results showed that sometimes the uncertainty due to the reference selection obtained stratifying the GTINs by market (ECR group) is such that for some area we are not able to say whether the SPI obtained is significantly higher or lower than that of the base area. Of course, the Jackknife is not the only estimator of variance that has been proposed in the literature (*e.g.* bootstrap, linearisation) but no comparison has been addressed in the literature so far (at least, to the authors best knowledge). This paper is a stepping-stone to the development of variability associated to SPIs. This is an important topic that will constitute a further line of research focussed on estimating uncertainty and comparing the results from different models.

References

Adelman, I. 1958. "A New Approach to the Construction of Index Numbers". *The Review of Economics and Statistics*, Volume 40, N. 3: 240-249.

Andersson, C., G. Forsman, and J. Wretman. 1987a. "On the Measurement of Errors in the Swedish Consumer Price Index". *Bulletin of the International Statistical Institute*, N. 52, Book 3: 155–171.

Andersson, C., G. Forsman, and J. Wretman. 1987b. "Estimating the Variance of a Complex Statistic: A Monte Carlo Study of Some Approximate Techniques". *Journal of Official Statistics - JOS*, Volume 3, N. 3: 251–265.

Aten, B.H. 2017. "Regional Price Parities and Real Regional Income for the United States" "Regional price parities and real regional income for the United States". *Social Indicators Research*, Volume 131, Issue 1: 123-143. https://doi.org/10.1007/s11205-015-1216-y.

Aten, B.H. 2008. "Estimates of State and Metropolitan Price Parities for Consumption Goods and Services in the United States, 2005". *Papers*, N. 05. Suitland, MD, U.S.: Bureau of Economic Analysis – BEA, U.S. Department of Commerce.

Balk, B.M. 2005. "Price Indexes for Elementary Aggregates: The Sampling Approach". *Journal of Official Statistics* - JOS, Volume 21, N. 4: 675–699.

Balk, B.M. 1995. "Axiomatic Price Index Theory: A Survey". *International Statistical Review*, Volume 63, N. 1: 69–93.

Balk, B.M. 1989. "On calculating the precision of consumer price indices". In Proceedings of the *47th Session of ISI World Statistics Congress* – WSC, Paris: contributed papers, Book 1.

Balk, B.M. 1981. "A simple method for constructing price indices for seasonal commodities". *Statistische Hefte*, Volume 22, Issue 1: 72-78.

Balk, B.M., and H.M.P. Kersten. 1986. "On the Precision of Consumer Price Indices Caused by the Sampling Variability of Budget Surveys". *Journal of Economic and Social Measurement*, Volume 14, N. 1: 19-35.

Banerjee, K.S. 1956. "A Note on the Optimal Allocation of Consumption Items in the construction of a Cost-of-Living Index". Econometrica, Volume 24, N. 3: 294-295.

Betti, G., F. Gagliardi, and V. Verma. 2018. "Simplified Jackknife Variance Estimates for Fuzzy Measures of Multidimensional Poverty". *International Statistical Review*, Volume 86, Issue 1: 68-86.

Biggeri, L., G. Ferrari, and Y. Zhao. 2017*a*. "Estimating Cross Province and Municipal City Price Level Differences in China: Some Experiments and Results". *Social Indicators Research*, Volume 131, Issue 1: 169-187.

Biggeri, L., and A. Giommi. 1987. "On the accuracy and precision of the consumer price indices. Methods and applications to evaluate the influence of the sampling of households". *Bulletin of the International Statistical Institute*, N. 52, Book 3: 155–171.

Biggeri, L., T. Laureti, and F. Polidoro. 2017*b*. "Computing Sub-national PPPs with CPI Data: An Empirical Analysis on Italian Data Using Country Product Dummy Models". *Social Indicators Research*, Volume 131, Issue 1: 93-121. https://doi.org/10.1007/s11205-015-1217-x.

Biggeri, L., and D.S. Prasada Rao (*Eds.*). 2021. *A Guide to the Compilation of Subnational Purchasing Power Parities (PPPs)*. Washington, D.C., U.S.: World Bank, International Comparison Program, Inter-Agency Coordination Group.

Caves, D.W., L.R. Christensen, and W.E. Diewert. 1982. "The Economic Theory of Index Numbers and the Measurement of Input, Output, and Productivity". *Econometrica*, Volume 50, N. 6: 1393-1414.

Chen, M., Y. Wang, and D.S. Prasada Rao. 2019. "Measuring the Spatial Price Differences in China with Regional Price Parity Methods" *The World Economy*, Volume 43, Issue 4: 1103-1146. https://doi.org/10.1111/twec.12899.

De Gregorio, C. 2012. "Sample size for the estimate of consumer price sub-indices with alternative statistical designs". *Rivista di statistica ufficiale/ Review of official statistics*, N. 1/2012: 19-47. Roma, Italy: Istat. https://www.istat.it/it/archivio/72183.

De Haan, J., E. Operdoes, and C. Schut. 1999. "Item selection in the Consumer Price Index: Cut-off versus probability sampling". *Survey Methodology*, Volume 25, N. 1: 31-41.

De Vitiis, C., A. Guandalini, F. Inglese, and M.D. Terribili. 2017. "Sampling Schemes Using Scanner Data for the Consumer Price Index". In Petrucci, A., F. Racioppi, and R. Verde (Eds.). *New Statistical Developments in Data Science. Atti Convegno della Società Italiana di Statistica - SIS 2017*. Springer Proceedings in Mathematics & Statistics, Volume 288: 203-2017. Cham, Switzerland: Springer.

Deaton, A. 2012. "Calibrating measurement uncertainty in purchasing power parity exchange rates". International Comparison Program – ICP, *7th Technical Advisory Group Meeting*, Washington, D.C., U.S., 17-18 September 2012.

Deaton, A., and B.H. Aten. 2017. "Trying to Understand the PPPs in ICP 2011: Why Are the Results so Different?". *American Economic Journal: Macroeconomics*, Volume 9, N. 1: 243-264.

Diewert, W.E., and K.J. Fox. 2022. "Substitution Bias in Multilateral Methods for CPI Construction". *Journal of Business & Economic Statistics*, Volume 40, Issue 1: 355-369.

Diewert, W.E. 1978. "Superlative Index Numbers and Consistency in Aggregation". *Econometrica*, Volume 46, N. 4: 883-900.

Diewert, W.E. 1976. "Exact and superlative index numbers". *Journal of Econometrics*, Volume 4, Issue 2: 115-145.

Dorfman, A.H., J. Lent, S.G. Leaver, and E. Wegman. 2006. "On Sample Survey Designs for Consumer Price Indexes". *Survey Methodology*, Volume 32, N. 2: 197-216.

Durbin, J. 1959. "A Note on the Application of Quenouille's Method of Bias Reduction to the Estimation of Ratios". *Biometrika*, Volume 46, N. 3/4: 477-480.

Eurostat, and Organisation for Economic Co-operation and Development - OECD. 2012. "Eurostat-OECD Methodological Manual on Purchasing Power Parities. 2012 edition". *Methodologies and Working papers*. Luxembourg: Publications Office of the European Union.

Feenstra, R.C., M. Xu, and A. Antoniades. 2017. "What is the Price of Tea in China? Towards the Relative Cost of Living in Chinese and US Cities". NBER, *Working Paper*, N. 23161. Cambridge, MA, U.S.: National Bureau of Economic Research - NBER.

Fenwick, D., A. Ball, P. Morgan, and M. Silver. 2001. "Sampling in Consumer Price Indices: what role for scanner data?". Paper presented at the *Sixth Meeting of the International Working Group on Price Indices*, Canberra, Australia, 2-6 April 2001.

Fenwick, D., and J. O'Donoghue. 2003. "Developing estimates of relative regional consumer price levels". *Economic Trends*, N. 599: 72-83, Office for National Statistics – ONS. London, UK: Palgrave Macmillan.

General Statistical Office of Vietnam - GSO. 2019. "Subnational PPPs in Vietnam". Paper presented at the *4th Meeting of the ICP Task Force on the Country Operational Guidelines and Procedure*, Paris, France, 2-3 May 2019.

Heravi, S., A. Heston, and M. Silver. 2003. "Using scanner data to estimate country price parities: A hedonic regression approach". *Review of Income and Wealth*, Volume 49, Issue 1: 1-21. https://doi.org/10.1111/1475-4991.00071.

Hill, R.J. 2004. "Constructing Price Indexes across Space and Time: The Case of the European Union". *American Economic Review*, Volume 94, N. 5: 1379-1410.

Hill, R.J., and M.P. Timmer. 2006. "Standard Errors as Weights in Multilateral Price Indexes". *Journal of Business & Economic Statistics*, Volume 24, N. 3: 366-377.

Imai, S., E. Diewert, and C. Shimizu. 2015. "Consumer Price Index Biases. Elementary Index Biases vs. Sampling Biases". Paper presented at the *Fourteenth Meeting of the Ottawa Group – the International Working Group on Price Indices*, Tokyo, Japan, 20-22 May 2015.

Institut National de la Statistique et des Études Économiques – INSEE, Direction des statistiques démographiques et sociales. 2019. "Enquete de comparaison spatiale des niveaux de prix à la consommation entre territoires francais". *Sources statistiques et indicateurs*. Paris, France: INSEE.

International Monetary Fund – IMF, International Labour Organization – ILO, Eurostat, United Nations Economic Commission for Europe – UNECE,

Organisation for Economic Co-operation and Development – OECD, and The World Bank. 2020. *Consumer Price Index Manual. Concepts and Methods* - 2020. Washington, D.C., U.S.: IMF.

Istituto Nazionale di Statistica – Istat, Unioncamere e Istituto Guglielmo Tagliacarne. 2010. "Le differenze nel livello dei prezzi al consumo tra i capoluoghi delle regioni italiane. Anno 2009". *Comunicato Stampa*. Roma, Italy: Istat. https://www.istat.it/it/archivio/6279.

Ivancic, L., W.E. Diewert, and K.J. Fox. 2011. "Scanner data, time aggregation and the construction of price indexes". *Journal of Econometrics*, Volume 161, Issue 1: 24-35.

Janský, P., and D. Kolcunová. 2017. "Regional differences in price levels across the European Union and their implications for its regional policy". *The Annals of Regional Science*, Volume 58, Issue 3: 641-660.

Klick, J., and O. Shoemaker. 2019. "Measures of Variance Across CPI Populations". *Research papers*. Washington, D.C., U.S.: U.S. Bureau of Labor Statistics, Office of Survey Methods Research.

Kocourek, A., J. Šimanová, and L. Šmída. 2016. "Modelling of Regional Price Levels in the Districts of the Czech Republic". Technical University of Liberec, 2018-09-25T11:50:33Z.

Kokoski, M.R., B.R. Moulton, and K.D. Zieschang. 1999. "Interarea Price Comparisons for Heterogenous Goods and Several Levels of Commodity Aggregation". In Heston, A., and and R.E. Lipsey (Eds.). "International and Interarea Comparisons of Income, Output and Prices": 123-166. NBER, *Studies in Income and Wealth*, Volume 61. Chicago, IL, U.S.: The University of Chicago Press.

Kott, P.S. 1984. "A Superpopulation Theory Approach to the Design of Price Index Estimators with Small Sampling Biases". *Journal of Business & Economic Statistics*, Volume 2, N. 1: 83-90.

Laureti, T., and F. Polidoro. 2022. "Using Scanner Data for Computing Consumer Spatial Price Indexes at Regional Level: An Empirical Application for Grocery Products in Italy". *Journal of Official Statistics* - JOS, Volume 38, N. 1: 23-56.

Laureti, T., and F. Polidoro. 2018. "Big data and spatial price comparisons of consumer prices". Presented at the *49th Scientific meeting of the Italian Statistical Society*, 20-22 June 2018, Palermo, Italy.

Laureti, T., and D.S. Prasada Rao. 2018. "Measuring spatial price level differences within a country: Current status and future developments". Estudios de Economía Aplicada, Volume 36, N. 1: 119–148.

Leaver, S.G., and R.A. Cage. 1997. *Estimating the Sampling Variance for Alternative Estimators of the U.S. Consumer Price Index*. Washington, D.C., U.S.: U.S. Bureau of Labor Statistics, Office of Survey Methods Research.

Leaver, S.G., J.E. Johnstone, and K.P. Archer. 1991. "Estimating unconditional variances for the U.S. consumer price index for 1978–1986". In American Statistical Association - ASA. *Proceedings of the Survey Research Methods Section:* 614-619. Alexandria, VA, U.S.: ASA.

Leaver, S.G., and R. Valliant. 1995. "Statistical Problems in Estimating the U.S. Consumer Price Index". In Cox, B.G., D.A. Binder, B. Nanjamma Chinnappa, A. Christianson, M.J. Colledge, and P.S Kott (Eds.). *Business Survey Methods*, Chapter 28: 543–566. Hoboken, NJ, U.S.: John Wiley & Sons, *Wiley Series in Probability and Statistics*.

Melser, D., and R.J. Hill. 2007. "Methods for Constructing Spatial Cost of Living Indexes". *Statisphere - Official Statistics Research*, Series 1, Article 3: 1-113.

Montero, J.-M., T. Laureti, R. Mínguez, and G. Fernández-Avilés. 2019. "A Stochastic Model with Penalized Coefficients for Spatial Price Comparisons: An Application to Regional Price Indexes in Italy". *Review of Income and Wealth*, Volume 66, Issue 3: 512-533. https://doi.org/10.1111/roiw.12422.

Morgenstern, O. 1963. *On the Accuracy of Economic Observations*. Princeton, NJ, U.S.: Princeton University Press.

Norberg, A., and C. Tongur. 2022. "Balancing the Swedish CPI". Paper presented at the *Seventeenth Meeting of the Ottawa Group – the International Working Group on Price Indices*, Roma, Italy 7-10 June 2022.

Paredes Araya, D., and V. Iturra Rivera. 2013. "Substitution bias and the construction of a spatial cost of living index". *Regional Science*, Volume 92, Issue 1: 103-117.

Pilat, D., and D.S. Prasada Rao. 1996. "Multilateral Comparisons Of Output, Productivity, And Purchasing Power Parities In Manufacturing". *Review of Income and Wealth*, Volume 42, Issue 2: 113-130.

Rao, D.S. Prasada. 2013. "The framework for the international comparison program (ICP)". In World Bank (Ed.). *Measuring the Real Size of the World Economy*, Chapter 1: 13-45. Washington, D.C., U.S.: World Bank.

Rao, D.S. Prasada, and G. Hajargasht. 2016. "Stochastic approach to computation of purchasing power parities in the International Comparison Program (ICP)". *Journal of Econometrics*, Volume 191, Issue 2: 414-425.

Reinsdorf, M.B., W.E. Diewert, and C. Ehemann. 2002. "Additive Decompositions for Fisher, Törnqvist and Geometric Mean Indexes". J*ournal of Economic and Social Measurement*, Volume 28, N. 1-2: 51-61.

Rokicki, B., and G.J.D. Hewings. 2019. "Regional price deflators in Poland: Evidence from NUTS-2 and NUTS-3 regions". *Spatial Economic Analysis*, Volume 14, Issue 1: 88-105. https://doi.org/10.1080/17421772.2018.1503705.

Roos, M. 2006. "Regional price levels in Germany". *Applied Economics*, Volume 38, N. 13: 1553-1566. https://doi.org/10.1080/00036840500407207.

Smith, P.A. 2021. "Estimating Sampling Errors in Consumer Price Indices". *International Statistical Review,* Volume 89, Issue 3: 481-504. https://doi.org/10.1111/insr.12438.

Tam, S.-M., and F. Clarke. 2015. "Big Data, Official Statistics and Some Initiatives by the Australian Bureau of Statistics". *International Statistical Review*, Volume 83, Issue 3: 436-448. https://doi.org/10.1111/insr.12105.

Tongur, C. 2019. "Inflation Measurement with Scanner Data and an Ever-Changing Fixed Basket". *Economie et Statistique*, N. 509: 31-47.

Weinand, S., and L. von Auer. 2020. "Anatomy of regional price differentials: evidence from micro-price data". *Spatial Economic Analysis*, Volume 15, Issue 4: 1-28. https://doi.org/10.1080/17421772.2020.1729998.

World Bank. 2013. *Measuring the Real Size of the World Economy: The Framework, Methodology, and Results of the International Comparison Program - ICP*. Washington, DC, U.S.: World Bank.

Exploring mobile network data for tourism statistics: the collaboration between Istat and Vodafone Business Italia

Lorenzo Cavallo[1], Erika Cerasti[1], Mascia Di Torrice[1], Alessandra Righi[1], Maria Teresa Santoro[1], Tiziana Tuoto[1], Luca Valentino[1], Dario Di Sorte[2], Mauro Rossi[2], Andrea Zaramella[2], Dario Bertocchi[3], Glauco Mantegari[3], Bruno Zamengo[3]

Abstract

The paper describes the collaboration between Istat and Vodafone Business Italia to innovate and enhance tourism statistics. The common goal is to evaluate the potential uses of mobile phone data in current surveys and to investigate new outputs for official statistics, such as visiting routes and means of transport. The analysis concerned inbound tourism (foreigners in Italy), domestic tourism (Italians in Italy), and outbound tourism (Italians abroad). The work presents analyses and results for the Province of Rimini and the Municipality of Roma, referred to August 2019/2020 and April 2020, and a trial of the use of the "Welcome SMS" for the estimate of the residents in Italy who travel to foreign countries. Phone data required specific treatments to meet the definitions of official statistics. Some aspects related to the location and definition of overnight stays will require further investigation.

Keywords: Data for good, privately-held data, mobile phone data, data analytics, tourism statistics, inbound tourism, domestic tourism, outbound tourism.

DOI: 10.1481/ISTATRIVISTASTATISTICAUFFICIALE_3.2022.02

1 Lorenzo Cavallo (cavallo@istat.it); Erika Cerasti (cerasti@istat.it); Mascia Di Torrice (maditorr@istat.it); Alessandra Righi (righi@istat.it); Maria Teresa Santoro (masantor@istat.it); Tiziana Tuoto (tuoto@istat.it); Luca Valentino (luvalent@istat.it), Italian National Institute of Statistics – Istat.

2 Dario Di Sorte (Dario.Disorte@Vodafone.com); Mauro Rossi (Mauro.Rossi2@Vodafone.com); Andrea Zaramella (Andrea.Zaramella@Vodafone.com), Vodafone Business Italia.

3 Dario Bertocchi (dario.bertocchi@motionanalytica.com); Glauco Mantegari (glauco.mantegari@motionanalytica.com); Bruno Zamengo (bruno.zamengo@motionanalytica.com), Motion Analytica.

The views and opinions expressed are those of the authors and do not necessarily reflect the official policy or position of the Italian National Institute of Statistics - Istat.

The authors would like to thank the anonymous reviewers for their comments and suggestions, which enhanced the quality of this article.

1. Introduction

In the framework of a collaboration protocol for the identification of innovative statistical methodologies, the Italian National Institute of Statistics - Istat and Vodafone Business Italia carried out experimentation on themes of domestic, inbound, and outbound tourism. The aim was to evaluate the potential of using mobile phone data in the production of high-quality official statistics. The trial allowed us to assess the value of mobile phone data in complementing and enriching current tourism surveys.

In this work, we describe the statistical information useful for the qualification and quantification of touristic flows that Istat could extract from mobile phone data while highlighting some aspects that need further investigation.

The trial involved the Province of Rimini and the Municipality of Roma, two areas with a strong tourist vocation. We had the opportunity to compare the pre-pandemic tourism flows (August 2019) with flows during the lockdown period (April 2020) and after the tourism restart (August 2020). We developed three case studies:

1. Arrivals and nights spent estimates for domestic tourism and inbound tourism;
2. Same-day visits and visit routes for domestic tourism and inbound tourism;
3. Outbound tourism estimates.

Data come from Vodafone Analytics technology, which continuously collects data on the activity of the mobile phone network, thus ensuring a very granular territorial and temporal detail of information.

We preprocessed mobile phone data with the aim of fitting the international definitions of official tourism statistics in an intense preliminary design phase. Also the developed algorithms aspire to meet the requirements of official statistics for estimate calculation.

Colleagues from Vodafone and Motion Analytica implemented the co-designed algorithms and provided the corresponding results in an aggregated form, suitable to be compared with official Istat estimates.

We believe this work is an interesting example of cooperation between a private company and a National Statistical Institute on the investigation of possible usages of mobile phone data for official statistics on tourism statistics and can inspire further cooperation in this field, facilitated by the multinational dimension of Vodafone company and the active role of Istat in the official statistics community. In addition, parts of this work on definitions and data processing are not restricted to tourism statistics and may be beneficial also for the usage of mobile phone data in other fields of official statistics.

This work is organised as follows: in Section 2 the international context of mobile phone network data usage in official statistics is outlined, with a focus on their applications for tourism statistics; Section 3 describes the main feature of the collaboration between Istat and Vodafone Business Italia; in Section 4 the data from Vodafone mobile network are illustrated. Section 5 introduces methods, concepts, and definitions of tourism applied to the mobile phone network data for the identified case studies. Sections 6, 7, and 8 present results on arrivals and nights spent estimates for domestic tourism and inbound tourism; same-day visits and visit routes for domestic tourism and inbound tourism; outbound tourism estimates, carried out using the "Welcome SMS" received by Italian residents arriving in foreign countries. Concluding remarks and future perspectives are provided in Section 9.

2. The international and national context of mobile phone network data for tourism statistics

Mobile phone network data have been more and more attractive for the production of statistics in the last decades. They have been used for several statistical goals, among others for estimating the present population, for disaster management, already before the COVID-19 pandemic outbreak, for human mobility, and for designing smart cities and special events. Several countries in Europe and all around the world have been experimenting with these data.

The usage of data coming from telecommunication companies for tourism statistics production is part of the agenda of principal international official statistical producers (Ahas *et al.*, 2014; De Meersman *et al*., 2016). As an example, the Department of Economic and Social Affairs - Statistics Division of the United Nations Organisation established an expert task team on this subject (UN-CEBD, 2021). Moreover, since 2016 the European Commission funds collaboration projects among National Statistical Institutes of Member States to use data coming from new sources, including data from telecommunication companies (Eurostat, 2021). In addition, Eurostat established a dedicated task force on mobile network operator data used for official statistics, in order to support member states (European Commission, 2021).

In many countries, the usage of mobile phone data still occurs in an experimental stage, and, as far as concerns tourism statistics, they are used for focussing on particular areas by exploiting their spatial granularity (Saluveer *et al.*, 2020). When mobile phone data are used for the official production of tourism statistics, Statistical Institutes are often supported by companies' experts in data analytics (Kuusika *et al.*, 2014). Results of several analyses confirm the high potential of those sources: they show a high correlation with statistics based on traditional data, especially for some target variables (like occupancy in accommodation establishments) and they allow to complete information on tourism categories not included in collective accommodation establishments (*e.g*. relatives and friends' houses, second houses, *etc*.). They can give also real-time information on the attendance to big events with a strong tourist attraction, and can be further analysed in terms of places of origin, length of stay, and other places visited nearby (Guidotti *et al.*, 2017; Nurmi and Piela, 2019).

When phone data are used for official production on tourism statistics, they have to be "mapped" to the internationally harmonised concepts, definitions, and standards of the official statistics on tourism (UNWTO, 2010). In Europe, NSIs are additionally subject to formal legislation (as will be described in detail in Sections 6 and 7). Sometimes, the difficulties in adapting mobile phone data to derive these specific tourism definitions may create limitations to the usage of mobile phone data in the current production of official tourism statistics. Currently, only a few countries (*e.g.* Indonesia and Estonia) use phone data to complete official statistical production.

The COVID-19 pandemic and the consequent lack of traditional statistics, particularly direct face-to-face surveys, sharply increased the use of mobile phone data for public and official statistics. For instance, the Bank of Italy used mostly phone data in March 2020 to get information on touristic flow for the compilation of the balance of payments at the requested monthly planning, due to the abrupt interruption of the survey because of the COVID-19 pandemic diffusion. The promptness in the activation of this source came as a result of experimental projects realised in previous years (Carboni *et al.*, 2020).

In addition, to enrich the understanding of people's movements in Italy during the outbreak of Coronavirus in 2020, Vodafone proposed a solution to anonymously monitor the daily movements of Vodafone SIMs in Italy (at an aggregated level and different spatial and temporal granularities) to provide insights about the movements of Italians aimed at supporting the decisions taken by local authorities in Italy (Calabrese *et al.*, 2021).

The collaborations among national statistical offices and telecommunication companies raise new challenges for the official statistics, relative to the usage of data collected by private parties (Ricciato *et al.*, 2018).

The indication provided by Eurostat on the outputs of the so-called *Trusted Smart Statistics*[4] produced with privately-held data explicitly recalls the need to leave that the data are processed by the holders and to apply new technologies and methodologies to ensure privacy, as well as transparency, verifiability and public control on the whole elaboration process (Ricciato *et al.*, 2021).

National statistical offices guarantee the quality assessment and transparency commitment of the processes and data produced, consistently with the

4 Statistical outputs derived from services provided by smart systems, embedding auditable and transparent data life-cycles, ensuring the validity and accuracy of the outputs, respecting data subjects' privacy and protecting confidentiality (see https://ec.europa.eu/eurostat/cros/content/trusted-smart-statistics-nutshell_en)

European quality framework, *e.g.* the European Statistics Code of Practice (Eurostat, 2017) and the Fundamental Principles of Official Statistics of the General Assembly of the United Nations (UNSD, 2014). The applications of these principles are well-established, at least in European countries, for statistics derived from surveys and administrative data. When statistics are produced from data collected by private parties, there is still the need for further insights on many quality aspects, and the requirement for unified standardised approaches and solutions to grant quality and transparency still animates the debate at the national and international levels.

Privacy guarantees are critical points. To this extent, it is important to remind that mobile phone data need to be used in full compliance with the privacy legislation, and the General Data Protection Regulation (EU GDPR, European Parliament, 2016), for instance following a Privacy by Design methodology. These requirements are added to the crucial attention to data privacy and confidentiality protection by the official statistics community strongly focussed on the output privacy preserving methodologies.

3. Istat – Vodafone Business Italia cooperation mode

In 2018, with the aim of developing studies on the opportunities for the use of Mobile Network Operators (MNO) data, Istat decided to set up a collaboration with Vodafone Italia, which was going to develop the sector of analytics. After a long phase of internal evaluation and bilateral negotiation, Istat decided to sign a first biennial Collaboration agreement in 2019. Istat has opted for an agreement, free of charge, for the exchange of methodologies and not of data, not yet feeling ready to carry out a tender on the market for the provision of MNO analytics services, since the real information scope of MNO data was not yet sufficiently known by Istat at that time.

The agreement provided that the data was elaborated by the Operator, for evaluating the possible use of mobile phone data in the official statistics production in sectors such as sub-populations, mobility and tourism flows, transport and traffic flow, smart cities, and lifestyles. From the Vodafone Business Italia point of view, an additional goal of the agreement was the comparison of its analytics with the official statistical standards.

The agreement stated that all data, documents, contents, and information exchanged were subject to non-disclosure rules (*i.e.* types of data available on their own databases; information on the services developed, algorithms and techniques to process the information; information derived from market analysis; data descriptors used; results of analyses conducted on aggregated data; information on the platforms, methodologies, and solutions adopted). Only a subsequent agreement, in force since 2021, set a relaxation of these rules, and any disclosure of results, if previously communicated by Istat, would be subject to authorisation from Vodafone.

However, the sole methodological exchange soon proved insufficient to investigate some key aspects of the resulting estimates. It was decided to carry out two projects which would have allowed Istat, against the compensation of the costs of using the machine time: 1) having “almost” direct access to the databases of network traffic information; 2) verifying the operating mode with which Vodafone realises the Analytics; 3) making prototypes, also verifying their accuracy.

This work is the product of one of these projects, which was carried out by means of an innovative kind of collaboration, called “Sprint methodology”.

A group of Istat methodologists and thematic experts worked online together with the experts of Vodafone Business Italia and Motion Analytica by means of a dashboard (Mirò, on software license provided by Vodafone) providing their contributions for building the output prototype, in terms of definitions of aggregates to estimate, algorithms to implement, *etc*.

Vodafone colleagues ran the programmes on their machines and returned the results for validation at each round to arrive at a shared output.

In the Sprint, we adopted careful but limited data exploitation, given the budget available for the machine time. It was necessary to arrive at the common Planning phase with clear objectives regarding the extension (territorial vs temporal) of the elaborations, to find a trade-off between the exploitation of spatial data at a given date and the opportunity to make temporal comparisons (*e.g.* before, during and after COVID-19 pandemic period) for a smaller territorial set.

We discussed concepts, definitions, and methodologies applied to adhere as much as possible to the international standard and definitions. In the presentation of the main results, we stress the points in which the adherence to the international definitions may be problematic and where there is a need for further investigation and elaboration on the conceptual mapping between the official statistics concepts and the mobile phone-derived concepts. We have fruitful insights into the data and methods used by Vodafone analytics, all the specific methods have been discussed and approved by the group, although the non-disclosure agreement limits the description of some methodological choices in this paper.

4. Big Data Analytics in Vodafone's mobile network

The continuous growth of mobile phone services and usage, together with the pervasive deployment of network coverage, have made wide-area mobile networks a valuable source of an extreme amount of data. The analytics of this time-space big data can indeed be an innovative way to explore several insights into subscribers' behaviour, presence, and movements.

Vodafone Analytics is the service offered by Vodafone that allows performing statistical analysis on non-personal data generated by the Vodafone 4G and 4.5G Networks. The essential component at the very base of Vodafone Analytics services is the ability to collect data from a network monitor analyser (probe data) which allows collecting space-time information on which cellphone tower and on which cell of that tower a SIM is connected to at a specific point in time (Calabrese *et al.*, 2014). The maximum spatial resolution is then at the single cell level, which usually covers an area with a radius from a few hundred meters (in urban areas) to a few kilometers (in rural areas). The frequency of observation over time depends on how often the mobile device connects to the network infrastructure, and on the network technology used. For instance, for 4G connections, we can identify events in the order of hundred times during a day (Pinelli *et al.*, 2015). The Vodafone Network consists of more than 200,000 telephone cells located throughout the country, covers almost 99% of the population, and collects up to tens of billions of positions daily, coming from interactions with approximately 20 million human SIMs (see AGCOM 02/22). Furthermore, the availability of a time series up to 18 months allows us to compare the phenomenon considering a complete cycle of seasonality and offers the possibility to measure trends and differences year over year.

The data are anonymised and irreversibly aggregated, in compliance with the privacy legislation, and the provisions of the EU GDPR, according to the Privacy by Design methodology. Customers are informed by Vodafone through the privacy policy that the data generated by the mobile network are used for these purposes in an anonymous and aggregate form, pursuant to and for the purposes of the legislation on data protection and as established in this regard by the Opinion 05/2014 on the anonymisation techniques of the Working Group pursuant to Article 29 for data protection. Vodafone, in compliance with this legislation, guarantees that it is technically impossible

for anyone to trace the personal data that would allow the identification of the interested party from the anonymised data stored in the cloud.

To calculate the stay location of a SIM (SIM position) we need to aggregate network data at the cell network towers level. A dwelling time algorithm to estimate stops in a specific area is used on the probe data. The algorithm receives in input a series of parameters, one of these is the minimum time threshold to be spent in a certain location (*i.e.* under the coverage of a cell if we are interested in a small area, or under the coverage of a group of cells if we are studying the presence, for instance, in a municipality). As a minimum temporal threshold, we used 30 minutes (minimum stop duration); this allows us to filter out noise from the data and also to catch only significant events. This ability to collect big data about SIM positions from a cellular mobile network is a function of the parameters described here below.

Space granularity – The density of mobile radio cells is paramount in guaranteeing the supply of reliable data. Vodafone can rely on more than 200,000 phone cells located across the national territory, which guarantees an advanced granularity of service because the higher the density of cells and the smaller their coverage and thus more precise the information about the SIM positions from the network big data. Within densely populated areas, the radius of a cell can be reduced to a very few hundred meters; on the other side, within rural areas where cellphone towers are definitely less dense, the dimension of a cell coverage may raise up to a few kilometers. An Additional increase of the space granularity can be achieved with the new coverage cells, and the DAS (Distributed Antenna System) technology, which Vodafone regularly installs indoors.

Additional information to improve the precision – The precision with which spatial data is acquired is decisive in obtaining high-quality geo-referenced information. A further improvement to spatial precision measurements, in respect to phone cell coverage, is obtained by accessing other data sources and complex interrelated methods. An example of sources are existing device applications; roughly 10 million devices provide geo-localisation with an accuracy of just a few meters (AGPS). This allows us to better associate the radio coverage of a cell with a geographical area since the mapping algorithm considers not only the output from simulators but also real information from the field (*i.e.* from the devices). It is possible, therefore, to build probabilistic

maps that allow for statistically distribute the users, resulting in an effective increase of spatial detail with respect to analysis based entirely on phone cells.

Time granularity – The frequency of the SIM positions sampling is of the utmost importance to enable the profiling process, along with accurate analysis, especially when it is necessary to know the actual presence or passage number in a limited geographical area (*i.e.* train stations, motorway tolls, border crossings, *etc*.). A sample every 30 minutes or more would not guarantee the effective presence of customers in limited geographic zones and, therefore, it would drastically diminish the reliability of the sample and most of the insights. Vodafone, however, can count on a high-frequency sample that guarantees presence notifications multiple times per minute, thanks to both the Circuit Switched (CS) network and the Packet Switched (PS) network. The ability to access PS data across the entire national territory enables access to data having a frequency in the order of minutes (roughly up to 1,000 records per day per SIM /device), in contrast to the Circuit Switched data that, on average, produces roughly only tens of records per day per SIM/device). CS networks typically generate Call Detail Record (CDR) information, which is collected based on user-generated events (voice calls and SMS), whereas PS networks also register IP-based service (data connections and app usage) and signalling.

Network coverage extension – In the case of the absence of mobile coverage an operator will not be able to collect data or supply data to the contracting authority. Therefore, it is essential for the network coverage extension to be optimal, as the time and costs of dedicated coverage constructions are not compatible with the expectations of the data users. The percentage of the population covered by 2G is close to 100%, whilst the percentage of the population covered by 4G is close to 99% (Vodafone, 2022).

Privacy by design – The Analytics services fully respect the current privacy laws and are compliant with GDPR rules in that all analysed data and information are always strictly anonymised and aggregated by design methodology. Customers are informed via privacy information that the data generated from the mobile network is used anonymously and in an aggregated form. Vodafone guarantees that it is technically impossible to trace back to any individual via data anonymously given and memorised in the cloud. For privacy reasons, data supplied by Vodafone is anonymous and aggregated, showing only cluster groups of more than 15/30 people.

5. Methods, concepts, and definitions

One of the main challenges we met is how to relate phone users and the resident population, which is in general the reference population for official statistics. This challenge requires solving several issues: how the subscribers of a single mobile network operator are representative of the resident population, how to consider people who do not have or use phones, how to deal with people subscribing to more than one contract, and so on. Additional issues are related to the identification of "usual" places (*e.g.* place of residence or place of tourism/leisure, in our case) for people on the territory on the basis of the phone locations and the time aggregation. All those aspects and many others have been discussed and illustrated during the working sessions of the cooperation. As already mentioned, the cooperation agreement leaves to the parties the ownership of the applied algorithms and they cannot be fully revealed at the time this paper is written. Nevertheless, in this Section, we provide some general references on the methods that have been applied.

Concerning the representativeness of the total resident population, Vodafone Analytics is based on a sample of about 1/3 of the Italian population, and geographic information is collected several times per minute. All the provided information is expanded to the entire reference universe. The process to expand the Vodafone sample to the total reference population (all people with a phone that correspond approximately to those who are at least 12 years old) is based on a machine learning algorithm. Indeed, an owner-calibrated algorithm allows to represent the entire universe of users and not only the Vodafone SIM owners or foreigners connected to the Italian network. The main input of this model are:

- Local market share of the operator, obtained from internal market analysis of the Italian SIMs. This is available by province and age groups bases.
- Market share on a national basis of foreign SIMs.
- Market share by type of SIM (business/consumer) obtained from market studies and official reports, such as the telecommunications observatory or AGCOM.
- Socio-demographic characteristics of users from proprietary data. Two levels of data refinement are applied to these indicators:

- "real user" information: Vodafone, through its customer interface points (contact centre, shop, *etc.*) updates the personal information of the real user of the SIM at each contact with its customers. The real user is the person who actually uses the SIM regardless of who signed the contract.
- Correction of any distortions related to the different market penetration between the different age groups of the population using official statistics such as Istat census data.

The combinations of calibration factors have led to the identification of different classes of inferential algorithms according to the types of analysis, creating very specific areas, such as, for example, (non-exhaustive list): indoor presence, at high speed, territorial presence, the passage through limited places. Continuous effort is put into the verification and calibration of the inferential model. When a reliable benchmark for comparison is available (such as, for example, official data of sports and musical events and of rail and air transport), these situations are exploited for testing the inferential algorithms.

The validation of the algorithms used to produce the total resident population and its quality assessment are quite far from the scope of this cooperation. Nevertheless, the discussion on these aspects have been really active during the common working sessions.

Hence, we decided to move forward with the traditional definition of usual residents and adopt a new definition of 'phone residence', to identify the location of the prevalent cell where a user (SIM) spent a night. A SIM is associated with a night phone cell if it is connected for at least 6 hours within the time window from 8 pm to 8 am. The concept of prevalent cell recalls the place where the SIM records the highest number of activity. We built three different user profiles (residents in the area, daily visitors, and tourists) by ascertaining their daily presence during the reference period, by quantifying the time spent in the area, and by identifying places where they spent nights.

We deeply analysed how to better operationalise concepts and definitions used in tourism official statistics using mobile phone data.

In Schema 5.1 the concepts and definitions of the tourism official statistics are connected with the solutions adopted in using mobile phone data.

Schema 5.1 - Concepts and definitions of the tourism official statistics and the corresponding definition adopted using Vodafone data

Concept	Official tourism statistics definition	Definition adopted
Usual place of residence	Municipality of usual residence of the tourist	*Night Cell:* the most frequent cell the user is attached to between 8 pm and 8 am. *Phone residence:* it corresponds to the municipality of the most frequent night cell, in the last 12 months before the reference period.
Tourist	Traveler taking a trip with an overnight stay to the main destination outside his/her municipality of usual residence, for any main purpose (business, leisure, or other personal purposes) other than to be employed by a resident entity in the country or place visited. *Domestic tourist* is a resident in Italy who make trips in Italy, inbound tourist is a resident abroad who makes trips to Italy; outbound tourist is a resident in Italy who makes trips abroad.	A user with a night cell referring to a municipality that differs from his/her *phone residence*.
Arrival	A person (tourist) who arrives at a tourist accommodation establishment and checks in	A tourist is registered as an "*arrival*" on the first day of his/her trip.
Night spent	Each night a guest/tourist actually spends (nights of sleep or stays) in a tourist accommodation establishment. The sum of the *Nights* in the municipality of the destination spent by all the *Tourists* is the *Total nights spent*.	Each night spent by a user with a night cell referring to a municipality that differs from his/her *phone residence*
Same-day visitor	Traveler who makes a visit outside his/her municipality of usual residence for at least three hours without an overnight stay. Visits made during a trip are excluded.	A user who visits a municipality for at least 3 hours being neither a tourist nor a phone resident.
Resident tourist	Tourist resident in Italy	Tourist with an Italian SIM or Foreigner tourists who spent at least 3 months in Italy.
Non-resident tourist	Tourist non-resident in Italy	*Foreign users*: users owning a foreign SIM.

We also used information on the frequency of the presence of a SIM in a municipality in order to distinguish *Frequent (or regular) visitors (i.e.* SIMs who visit the same municipality at least 4 times within a month) from *Infrequent (or non-habitual) visitors*: sum of *Italian* and *Foreign users* net of *Frequent (or regular) Visitors*.

Based on all these definitions, we estimated inbound flows by approximating the foreign nationality of residence with that of the foreign operator who issued the SIM, net of *foreign users resident in Italy*. For the estimation of the domestic tourism flows we used the determination of the phone residence of SIMs observed in the destination to differentiate tourists from users residing in the observed municipality.

6. Domestic and inbound tourism – Arrivals and nights spent

Istat currently measures arrivals and nights spent relative to domestic and inbound tourism through the "Occupancy in collective accommodation establishments" survey. The survey is a census, carried out on a monthly basis and aimed to estimate the tourist flows on the national territory, in particular flows recorded in official accommodation establishments in the country. It responds to the framework of the EU Regulation no. 692/2011 of the European Parliament and the Council concerning the European statistics on tourism (EU Regulation, 2011), as amended by the EU Delegated Regulation no. 2019/1681 (EU Commission, 2019). The scope of this Regulation concerns all tourist accommodation establishments providing as a paid service (although the price might be partially or fully subsidised) short-term or short-stay accommodation services[5] (Eurostat, 2015).

The survey quantifies, for each month and each municipality, arrivals and nights spent by Italian residents (by region of residence) and non-residents in Italy (by country of residence) at tourist accommodation establishments, disaggregated by category of hotels and similar accommodation and by type of other collective accommodation establishments. The official accommodation establishments - hotels and others - involved in the survey are currently over 218 thousand, resulting from the local registers, held and managed by the regional intermediate bodies[6]. The establishments' owners transmit daily data on occupancy to their local public administration in charge of the survey; data are then summarised monthly, at the municipal level, and submitted to Istat by the intermediate bodies. The quality of traditional tourism statistics is certainly high and it has improved over time. Since the survey is intermediated by Regions, quality may /vary slightly from Region to Region, although, it is definitely kept under control by continuous monitoring.

For the trial with MNO data, we identified two areas with a strong tourist vocation but with quite different characteristics. The province of Rimini is indeed a seasonal and marina destination, characterised by the presence of

5 These services are classified and described in the three following groups of the NACE rev.2 classification: 55.1 (hotels and similar accommodation), 55.2 (holiday and other short-stay accommodation) and 55.3 (camping grounds, recreational vehicle parks and trailer parks). Non-rented accommodation are out of scope

6 The "Occupancy in collective accommodation establishments" survey is intermediate by local public administrations in charge of the survey (Statistics Offices in the Regions or Provinces, local tourism bodies) and data collection is entrusted to them.

many second homes. Rome, on the other hand, is extremely attractive, not only for tourists but also for commuters, and less seasonal. For the Province of Rimini, the Istat tourist flows include also of the values registered in private (rented) accommodations that the Emilia-Romagna Region collects and supplies to Istat. The same information is not available for the Municipality of Rome, despite this phenomenon being relevant also for this destination, according to Airdna's elaboration on Airbnb data (AirDNA, 2022). Hence, tourist flows using private rented accommodations might be currently underreported in official figures for the Municipality of Rome.

The analysis concerned *Arrivals* and *Nights spent*, but focussed more in-depth on *Nights spent,* a variable with a clear defining perimeter, for the province of Rimini and the Municipality of Rome, during August 2019 and August 2020, and April 2020. The latter period can be considered as a benchmark of the presence of the population in their usual place of residence, due to the restriction to the movements connected to the Covid-19 pandemic (the whole Italian territory was in lockdown). Nevertheless, April 2020 is not an interesting reference period for tourism, since in principle it should be completely canceled out.

Table 6.1 shows the arrivals and nights spent derived from Vodafone data compared to the Istat figures.

Table 6.1 - Arrivals and nights spent, comparison between Istat and Vodafone (absolute values and percentage differences)

Variable	Territory	Year	Month	Istat	Vodafone	% diff. Vodafone-Istat
Arrivals	Roma	2019	8	810.639	2.664.117	228,6%
		2020	8	188.358	1.036.414	450,2%
	Rimini	2019	8	797.381	1.203.745	51,0%
		2020	8	747.386	1.179.676	57,8%
Nights spent	Roma	2019	8	2.391.301	6.827.457	185,5%
		2020	8	570.568	2.996.464	425,2%
	Rimini	2019	8	4.482.871	3.645.220	-18,7%
		2020	8	3.749.862	4.135.101	10,3%

Sources: Istat, Occupancy in collective accommodation establishments and Vodafone Analytics, years 2019-2020

The comparison between mobile phone data and Istat data shows that:

- Vodafone's estimations of nights spent, both for domestic and inbound, are always higher than Istat data, except for domestic nights spent in the province of Rimini in August 2019.

- The gap between Vodafone and Istat data for the municipality of Roma is wider than the one recorded for the province of Rimini.

The first result is quite expected, since Istat data do not include a part of the private accommodations, particularly in Rome, and all the non-rented accommodations (second homes, homes of friends or relatives, *etc*.), according to the EU Regulation. In addition, it seems that the Vodafone data capture also nights related to other phenomena, like commuters, temporary workers, and other temporary present foreigners. The negative difference in August 2019 the province of Rimini requires further investigations, it could be related to local behaviour during summer nights and occasional data quality issues.

The percentage changes in arrivals and nights spent for the month of August 2020, compared with the same month of 2019, are almost equal between the two sources for the municipality of Roma, but different for the province of Rimini, as shown in Table 6.2.

Table 6.2 – Arrivals and nights spent by origin of the guests, comparison between Istat and Vodafone data (percentage changes)**. August 2020 on August 2019**

Territory	Arrivals			Nights spent		
	Domestic	Inbound	Total	Domestic	Inbound	Total
			Istat			
Roma	-12,68	-89,37	-76,76	-11,61	-88,89	-76,14
Rimini	1,58	-41,81	-6,27	-8,96	-48,81	-16,35
			Vodafone			
Roma	-12,26	-75,71	-61,10	-12,21	-70,57	-56,11
Rimini	17,66	-45,50	-2,00	37,60	-40,22	13,44

Sources: Istat, Occupancy in collective accommodation establishments and Vodafone Analytics, August 2020

To further investigate the reasons for the large differences between Vodafone and Istat data, we considered April 2020 as the reference time, since that month was characterised by the Covid-19 pandemic lockdown. In principle, at that time the tourist flows were expected to be nearly zero, hence a better alignment between the two sources should apply. Table 6.3 reports the results on arrivals and nights spent in April 2020, showing opposite evidence, presenting even higher divergences than in August 2019 and 2020. This finding confirms the difficulties of the tourism definition of Vodafone data, as well as a potential under-reporting of Istat data. A first explanation could be connected to some specific subgroups of SIMSs considered in the algorithm, *e.g.* the

ones of foreign residents in Italy (therefore "not tourists"). The nationality of the foreign SIMs registered in April largely refers to nationalities present in Italy for business reasons rather than for leisure.

This is particularly evident in Rome in which the presence of foreign residents in Italy is greater than in the province of Rimini.

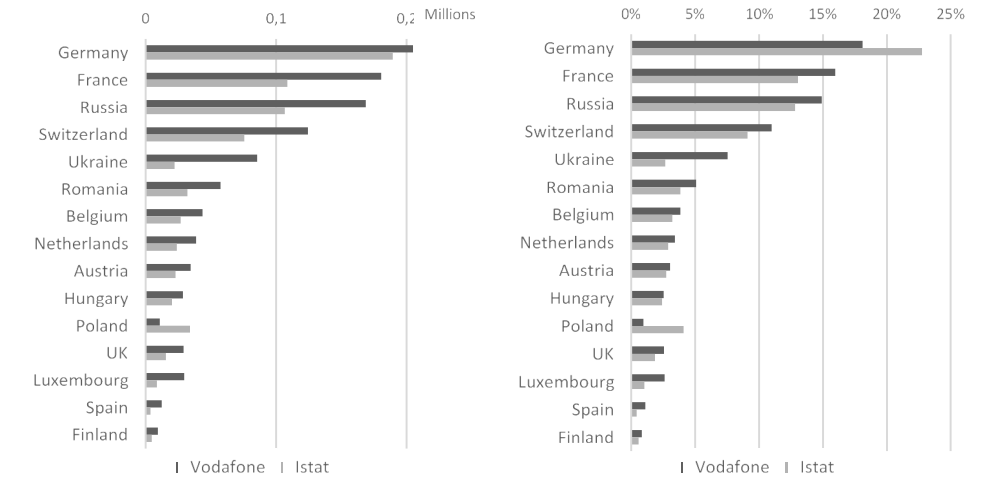
Table 6.3 - Arrivals and nights spent in April 2020, comparison between Istat and Vodafone (absolute values and percentage differences)

Variable	Territory	Year	Month	Istat	Vodafone	% diff. Vodafone-Istat
Arrivals	Roma	2020	4	6.013	154.385	2.467,5%
	Rimini	2020	4	947	18.170	1.818,7%
Nights spent	Roma	2020	4	15.819	1.577.083	9.869,5%
	Rimini	2020	4	47.212	150.032	217,8%

Sources: Istat, Occupancy in collective accommodation establishments and Vodafone Analytics, years 2019-2020

Figure 6.1 reports the estimates on the nights spent by inbound tourists by country of origin for the province of Rimini. It shows a good degree of convergence and consistency in terms of percentage shares between the two sources; Germany and Russia are among the most relevant countries.

Figure 6.1 - Nights spent by inbound tourists in the Province of Rimini by country of origin (absolute values and percentage shares)**. August 2020**

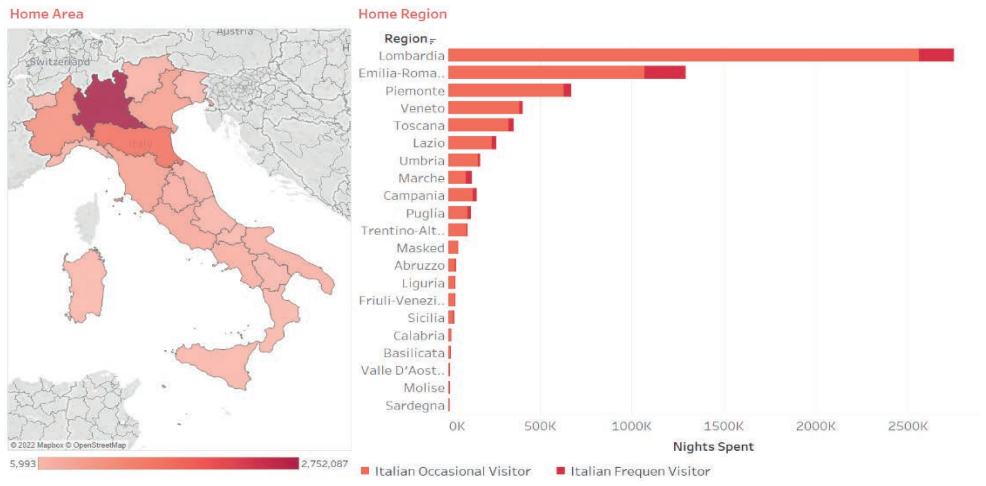


Sources: Istat, Occupancy in collective accommodation establishments and Vodafone Analytics, August 2020

Figure 6.2 shows nights spent by domestic and inbound tourists in the municipalities of the Province of Rimini (absolute values) in August 2020.

Both the charts show the home location of the tourists. In the map, the gradation of colour of the Region is proportional to the number of nights spent by people who are from that Region. The bar chart on the right side of the picture reports the same piece of information, distinguishing between occasional and frequent visitors. People from the first category visit the area at most three times in the analysed period, while people from the other one visit the area at least four times in the analysed period. While tourists are mainly from Lombardia, most frequent visitors come from Emilia-Romagna.

Figure 6.2 - Nights spent by inbound and domestic tourists in the municipalities of the Province of Rimini (absolute values)**. August 2020**

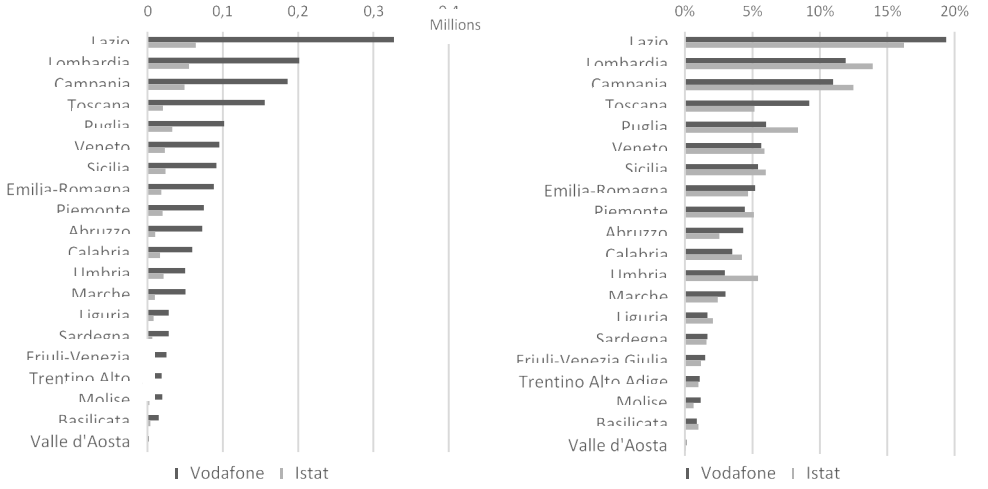


Source: Overnights versus provenience and frequency, Vodafone Analytics, August 2020

Figure 6.3 shows nights spent in Romew in August 2020, by domestic tourists per region of origin. It shows a strong difference in absolute value but a limited one in terms of rank between the two sources. Lazio, Lombardia, and Campania are the territories that originate the highest shares of tourists for both sources.

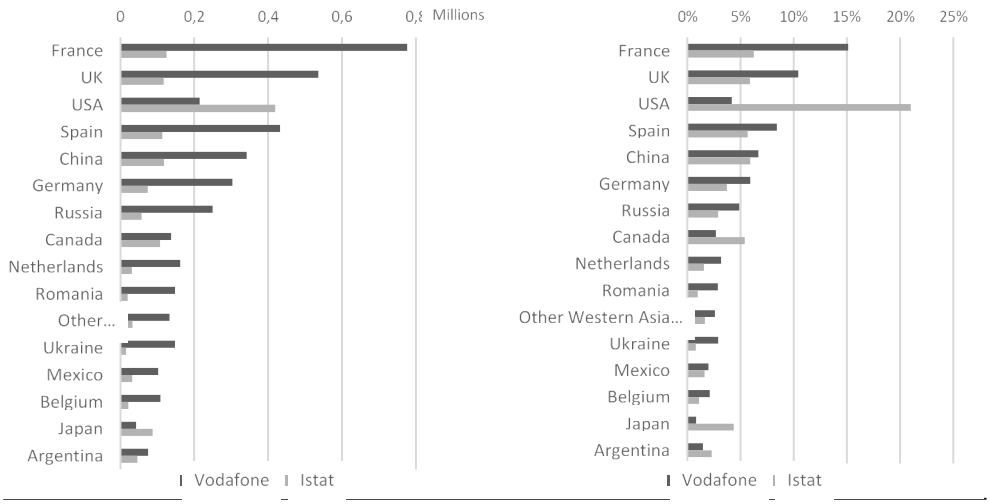
Nights spent by inbound tourists by country of origin are reported in Figure 6.4, showing smaller differences in absolute value but higher differences in the percentage shares for the two main foreign countries of origin (France and the United States) between the two sources. In particular, for the United States, differences are relevant. This is connected to the relevant presence of US citizens temporary living in Rome, working in international organisations, as already underlined above.

Figure 6.3 - Domestic nights spent in the municipality of Roma *(absolute values and percentage shares)***. August 2020**



Sources: Istat, Occupancy in collective accommodation establishments and Vodafone Analytics, August 2020

Figure 6.4 - Nights spent by inbound tourists in the municipality of Roma (absolute values and percentage shares)**. August 2020**



Sources: Istat, Occupancy in collective accommodation establishments and Vodafone Analytics, August 2020

7. Domestic and inbound tourism – Same-day visits and visit routes

Same-day visits are visits without overnight stays. They are measured by Istat through the "Trips and holidays" survey, a focus embedded in the sample Household Budget Survey since 2014. It has the goal to obtain information on the tourist movements of the population through direct interviews with households (tourist demand). The resulting estimates are about the number of tourists, trips, overnight stays, expenses (both per trip and per day), and the number of same-day visits in Italy or abroad (Di Torrice, 2018). The Regulation for Tourism Statistics no. 692/2011, which is also the reference for the "Trips and holidays" survey, defines same-day visits as "*visits without overnight stay carried out by residents outside their usual environment starting from their usual place of residence*". The "Trips and holidays" survey currently estimates the domestic aggregate, whereas it does not provide estimates related to the inbound same-day visits, and therefore the Vodafone estimates may represent a potential enrichment for official statistics.

The European guidelines for statistical surveys recommend identifying a same-day visit by surveying the purpose of the visit, the crossing of administrative borders, the duration (at least 3 hours at the place of destination), and the frequency, thus making it possible to exclude visits that do not include a tourist element. These requirements are explicit in the "Trips and holidays" survey questionnaire, but they are difficult to identify using mobile phone data.

In this Section, we present two insights that the granularity of Vodafone data allows to add to the current official statistical production. The new investigations and the increased level of details provided are however subject to the privacy preservation of routes and trips, which Vodafone guarantees by design via the aggregation of the events. Using Vodafone data, we analysed the potential "same-day visits" of visitors (Italian residents and foreigners) who travel to the province of Rimini (domestic and inbound visits, respectively). The analysis is carried out for the week of Mid-August, 2019 (12-18 August 2019), defining the Italian or foreign user as a same-day visitor who spent at least three hours in a municipality in the province of Rimini, without staying overnight and without having the phone residence there or nearby. Therefore, the algorithm applied both the criteria of the duration of the visit and of the municipality of residence border crossing. On the contrary, it was not

possible to exclude the usual behaviours through frequency, given the limited observation period. As the "Trips and holiday" survey does not ask for the exact date of the same-day visit, so it is not possible to narrow the comparison to a specific week but only to the entire month of August.

Figure 7.1 shows the municipalities (first r-w charts) in the province of Rimini visited by Italians by region of the visitors (second-row charts). The most visited municipality is Rimini doubling Riccione and Bellaria-Igea Marina, the-second and third most-visited areas, respectively. Most Italian same-day visitors are from Lombardia and Emilia-Romagna.

In Figure 7.2 are the visited municipalities in the province of Rimini (first row charts) by the foreign country of origin of visitors (second row charts). The most visited municipality is Rimini. Most foreign same-day visitors are French, Germans, and Russians.

These results let us appreciate the level of territorial detail provided by the mobile phone data compared to the sample survey findings, both in terms of the origin and destination of the visit.

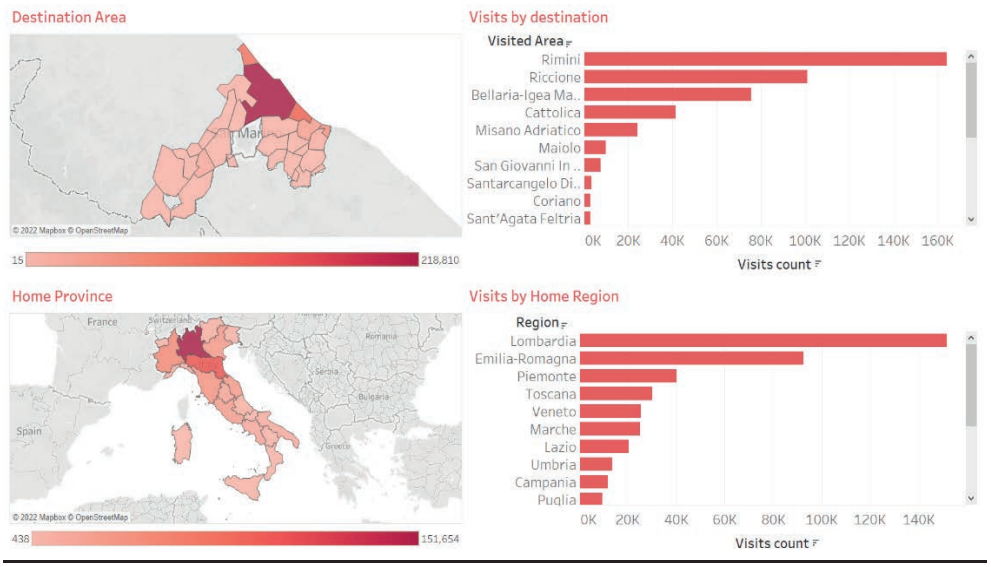
Concerning the origin of the visits of the residents, the results of the sample survey "Trips and holiday" for the same period point out a smaller set of regions, limited to Emilia-Romagna and Marche. However, it must be noted that the presence of regions of origin geographically distant from the province of Rimini provided by Vodafone data indicates that the estimate may be affected by the erroneous inclusion of same-day visits to Rimini as a part of a trip with an overnight stay in an adjacent territory.

This issue is more evident for foreign visitors (not surveyed by the Istat sample survey); even if they come mainly from "neighbouring" countries (France, Germany, and Russia), it is unlikely that they visited the province of Rimini in one day (Figure 7.2). These visits are not strictly defined as tourism, according to the Regulations for Tourism Statistics, as probably they do not originate and end from/in the place of residence.

Also regarding the destination of the visits made by residents, Vodafone data cover a wide set of municipalities, whereas the "Trips and holiday" survey points out only visits to Riccione and Rimini, probably because the tourists may indicate only a municipality in the questionnaire (the main one visited during the day).

All the above considerations can explain the substantial difference between the two sources, as the total amount of visits made by residents estimated by the Istat survey on the whole month is about 22% of the amount provided by Vodafone data, which considers only on the reference week. The most likely reason we identified is the inclusion of same-day visits made during a trip in the Vodafone algorithm, which consequently needs to be refined in order to count only visits starting and ending in the place of usual residence of the tourist (*i.e.* the place of phone residence). In addition, the inclusion of usual same-day visits in the Vodafone estimates can increase discrepancies. Visits made to the same destination on a weekly basis should be excluded as done for frequent trips, but this would be possible only by observing Vodafone data for at least one month.

Figure 7.1 - Same-day visits in the province of Rimini made by residents, by region of origin, and visited municipality (absolute values). Mid-August, 2019 (12-18 August 2019)

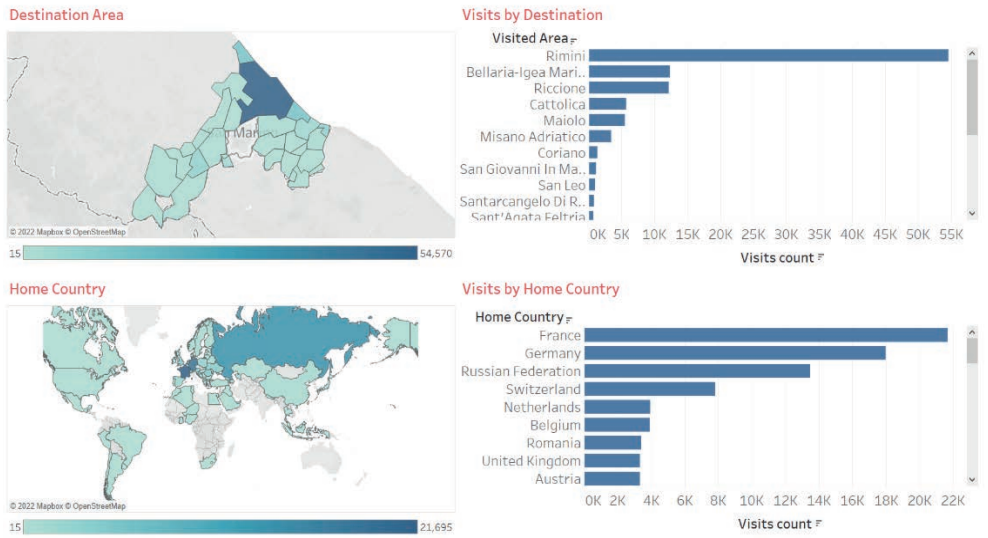


Source: Vodafone Analytics 2019

A further line of analysis that may enrich official statistics is the study of visit routes (or co-visits), namely, different municipalities visited by Italian or foreign tourists during the observed period (August 2019). This analysis provides a significant added value for tourist destinations, as it offers them

the opportunity to know both the most attractive and least frequented "routes" of visit, which, therefore, could be better promoted to decongest the most crowded places. The matrix in Figure 7.3 highlights the shifts between the municipalities (in row and column) co-visited during the observation period, assigning a darker colour to the pairs of those most affected by the phenomenon. The blue map on the left reports the number of foreign tourists per home country, and the red one is the number of Italian tourists per home region.

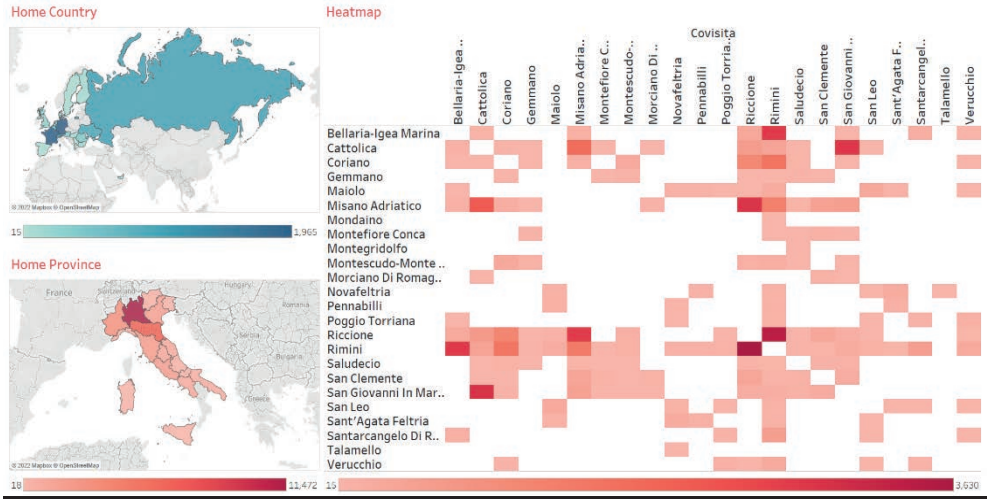
Figure 7.2 - Same-day visits in the province of Rimini made by foreigners, by the foreign country of origin, and visited municipality (absolute values)**. Mid-August, 2019 (12-18 August 2019)**



Source: Vodafone Analytics 2019

Most foreigners who visited at least two municipalities come from France and Germany while Italians who visited at least two municipalities come from Lombardia and Emilia-Romagna. The municipalities most frequently visited by tourists during the same trip in the observed period are Rimini and Riccione. Furthermore, the matrix highlights the close relationship between Riccione and Misano Adriatico, between Rimini and Bellaria-Igea Marina, and between Cattolica and San Giovanni in Marignano and Misano Adriatica.

Figure 7.3 - Visit routes in the province of Rimini by region or country of provenience of the visitor and co-visits among municipalities (August 2019)



Source: Co-visits among municipalities, Vodafone Analytics 2019

8. Outbound Tourism

Istat measures outbound tourism (residents in Italy who travel to foreign locations) through the above-mentioned "Trips and Holidays" survey. Experimental estimates on the use of SIMs during the trips showed that in 80% of trips the travelers used a mobile phone (Dattilo and Sabato, 2017).

It is possible to estimate the size of this type of tourist using the "Welcome SMS" from Vodafone data. When a Vodafone Italy customer connects for the first time to a non-Italian phone network, he/she receives an SMS summarising the contractual conditions and the rates applied. Vodafone Italy logs where (the country) and when (the timestamp) its customers receive these "Welcome SMS". Since these logs contain the SIM identifier, the roaming country, and the first roaming timestamp in that country only, it is not possible to trace any detail of visited location like cities or specific sub-regions (*i.e.* it is not possible to distinguish between "Galicia" and "Catalonia" through these logs since both will be labelled as "Spain"). The same privacy rules previously described in Section 4 about irreversible aggregation and anonymisation of user identification data are applied. To ensure that the Vodafone subscribers are representing all the Italian resident, a similar procedure is adopted, as shortly described in Section 5.

The multi-national dimension of Vodafone Company potentially allows for double checking the outbound tourism flows, for instance an Italian tourist who travel abroad would be addressed via roaming to a foreign Vodafone operational company, if it exists in the visited country. This topic is subject to further developments in the Vodafone international holding.

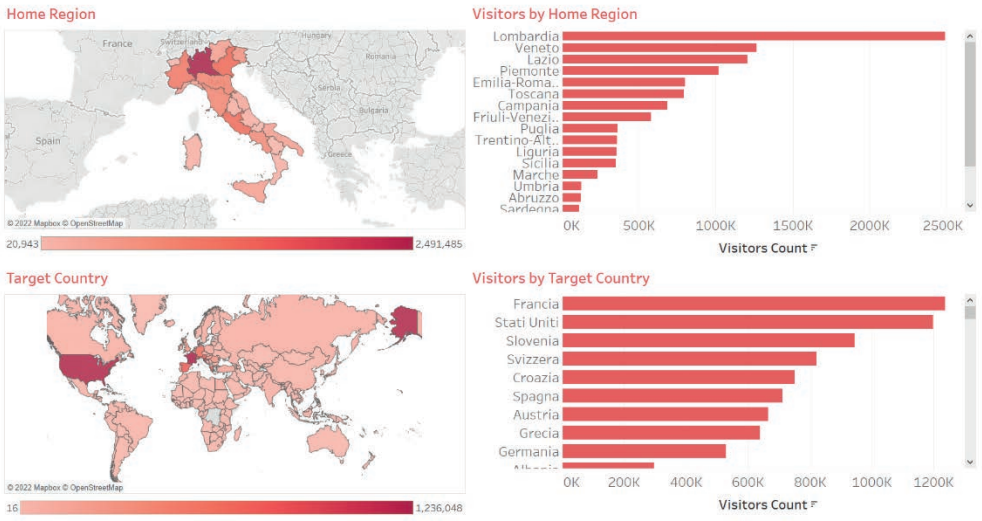
From these data, we developed an analysis aimed to estimating the size of Italian outbound tourism in August 2019 by counting how many Welcome SMS were received per foreign country and then estimating the total outbound Italian population by using the methods described in Section 5. We further enriched the analysis by adding the estimated home location (*i.e.* the phone residence) of SIM at the regional level.

Main results of the trial are in Figure 8.1. The chart at the top shows the distribution of outbound tourists by region of origin. The comparison between Vodafone data, and the "Trips and Holidays" survey estimates, shows a convergence in the top regions of the chart, as both sources indicate

Lombardia and Veneto in the two top positions, followed by Lazio (Vodafone data) or Emilia-Romagna (Istat survey). The chart at the bottom of the Figure shows the distribution of outbound tourists by visited country. France, the USA, and Slovenia appear in the three top positions according to Vodafone data, while the top visited countries are Greece, Germany, and France for the official statistics.

However, results are only partially comparable due to the fact that the official statistics consider only the main destination of the trip, while the Welcome SMS is also delivering information during the intermediate stages of a trip. For example, Slovenia is at the top of the list because is a transit for tourists traveling towards Croatia, and the United States acts as an air stopover for many intercontinental flights. In addition, we need further analysis to differentiate the movements of cross-border workers from tourist flows. This is particularly important concerning the neighboring countries such as France, Slovenia, Switzerland, and Austria, where cross-border workers are common and can cause an overestimation of tourist flows to these countries.

Figure 8.1 - Italian outbound tourists by region of origin and foreign visited country (absolute values)**. August 2019**



Source: Outbound tourism versus region of provenience and country of destination, Vodafone Analytics 2019

9. Conclusions and future developments

The collaboration experience between Istat and Vodafone confirmed the potential of mobile phone data for the production of statistics complementary to the existing official statistics. The big data generated by telecommunications companies may open new perspectives of analysis in the tourism sector, not only for the accurate monitoring on a territorial scale (official statistics are generally at the municipal or regional disaggregation) but also for the greater timeliness and temporal granularity that they can guarantee. Mobile phone data provide the opportunity to monitor the tourism experience in specific areas of interest; while controlling tourism pressure based on the total number of users (people in a given area at a specific time) and time spent in a given destination. Furthermore, it is possible to create groups of users by visit (same-day visit or visit with an overnight stay) and categories (nationality, age classes, types), following their mobility in the area.

The timeliness and the possibility of more detailed temporal breakdowns are growing needs of tourism statistics users, who want information, increasingly close in time to the event, or even estimates in advance. This is especially the case for extraordinarily organised events: short-term events, such as fairs, or longer events, such as the Jubilee or tourism in capitals of culture. This demand emerged clearly during the Covid-19 pandemic, when the need for information became greater while, at the same time, it was difficult to carry on traditional surveys (at least, on households).

A further aspect to underline concerns the coverage, namely the ability of mobile phone data to represent the phenomenon. In terms of overall quantification, the estimate of the total number of tourists (not only those who spend nights in official accommodation) is currently provided by the "Trips and Holiday" sample survey, which also detects tourists in second homes, guests of friends, or those who stay overnight in short-term rentals (for example, Airbnb). However, a sample survey is affected by the limitations of the sample size and the statistical error, which is greater when the event becomes rarefied (as happened to trips during the pandemic). Being able to rely on information derived from MNO data to support and complement traditional surveys can be a significant advantage in maintaining the continuity of tourism statistics and in strengthening their coverage.

However, it should be emphasised that relevant information currently collected by traditional surveys, which is also required by the Regulation on Tourism statistics, cannot be inferred from mobile phone data: for instance, the reason for the trip, the type of accommodation in which the tourist stayed, the quantification of the expenditure incurred. On the other hand, mobile phone data open up the possibility of producing new information and statistics currently not existing, such as co-visits or means of transport upon arrival and departure of the tourist, which would be very useful for users.

In our opinion, the main finding of this collaboration experience is that the mobile phone data and other auxiliary information held by Vodafone represent a valuable contribution to improving the official tourism estimates, forecasts, or eventually for calibrations of sample surveys. The experience described in this work suggests that mobile phone data can supplement and complement the current statistics production, and can provide new insights on emerging topics only partially covered by traditional data sources. It is hard to imagine that mobile phone data can completely replace traditional data sources for the production of tourism statistics, in the short term.

The main challenge remains to refine methodologically and conceptually Vodafone's definitions and data processing algorithms for a greater convergence with the concepts of official statistics (*e.g.* phone residence, resident foreign users, arrivals, outbound tourism), which requires joint testing and verification procedures of implemented algorithms. While in this experience we mainly adopted the Vodafone Analytics perspectives and solutions, discussing and partially adjusting them to the official statistics needs, in future collaborations we envisage the necessity to overcome the current solutions to better identify, for instance, different groups/subpopulations of users, via clustering techniques and AI approaches. In the same spirit, a deeper analysis would be devoted to the assessment of the potential selectivity of the input data, also via dedicated sample surveys, as already used in national statistical offices. An occasional sample survey, with a limited size, would be actually effective for assessing the quality of the results derived from the mobile phone data, as well as from other organically generated data (Zhang, 2019). This way could be useful also to understand how the differences observed in the results are due to misalignment in the reference populations (residents and non-residents) and/or is determined by the ability of the new sources in capturing the tourism flows according to the current official statistics definitions.

Finally, the mandate of official statistics requires the use of transparent algorithms, to this extent, we still require that the international community moves forward with a public-private partnership that guarantees the respect of this principle in a win-win standard agreement for considering data as a public good. Regulated access to statistics derived from mobile phone data is a prerequisite for using this data source in regular tourism statistics production. We hope that the effective collaboration already experienced and described so far will lead to positive results in this regard.

References

Ahas, R., J. Armoogum, S. Esko, M. Ilves, E. Karus, J.-L. Madre, O. Nurmi, F. Potier, D. Schmücker, U. Sonntag, and M. Tiru. 2014. "Feasibility Study on the Use of Mobile Positioning Data for Tourism Statistics (Eurostat Contract No. 30501.2012.001-2012.452)". *Consolidated Report*. Luxembourg: Eurostat.

AirDNA. 2022. *Vacation Rental Data to Set You Apart. Insights to Keep You Ahead*.

Autorità per le Garanzie nelle Comunicazioni – AGCOM. 2022. *Osservatorio sulle Comunicazioni*, N. 2/2022. Roma, Italy: AGCOM.

Calabrese, F., E. Cobelli, V. Ferraiuolo, G. Misseri, F. Pinelli, and D. Rodriguez. 2021. "Using Vodafone mobile phone network data to provide insights into citizens mobility in Italy during the Coronavirus outbreak". *Data & Policy*, Volume 3: e22. DOI: 10.1017/dap.2021.18.

Calabrese, F., L. Ferrari, and V.D. Blondel. 2014. "Urban Sensing Using Mobile Phone Network Data: A Survey of Research". *ACM Computing Surveys*, Volume 47, Issue 2, Article N. 25: 1-20.

Calabrese, F., G. Di Lorenzo, L. Liu, and C. Ratti. 2011. "Estimating Origin-Destination Flows Using Mobile Phone Location Data". *IEEE Pervasive Computing*, Volume 10, Issue 4: 36-44.

Carboni, A., C. Doria, e S. Zappa. 2020. "La produzione statistica nell'emergenza Covid19: la stima dei "viaggi" in bilancia dei pagamenti", *Note Covid-19*. Roma, Italy: Banca d'Italia.

Dattilo, B., and M. Sabato. 2017. "Travelling SIM and Trips: An approach to make mobile phone data usable in tourism statistics". Presented at *New Techniques and Technologies for Statistics – NTTS 2017*, Brussels, Belgium, 11-17 March 2017.

De Meersman, F., G. Seynaeve, M. Debusschere, P. Lusyne, P. Dewitte, Y. Baeyens, A. Wirthmann, C. Demunter, F. Reis, and H.I. Reuter. 2016. "Assessing the Quality of Mobile Phone Data as a Source of Statistics". Paper presented at the *European Conference on Quality in Official Statistics - Q2016*, Madrid, Spain, 31 May - 3 June 2016.

Di Torrice, M. (*a cura di*). 2018. "La nuova indagine sulla domanda turistica". *Letture statistiche – Metodi*. Roma, Italy: Istat. https://www.istat.it/it/archivio/222043.

European Commission. 2019. *Commission Delegated Regulation (EU) 2019/1681 of 1 August 2019*, amending Regulation (EU) N. 692/2011 of the European Parliament and of the Council concerning European statistics on tourism, as regards the transmission deadlines and adaptation of Annexes I and II.

European Parliament, and the Council of the European Union. 2016. *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016* on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation).

European Parliament, and the Council of the European Union. 2011. *EU Regulation N. 692/2011 of the European Parliament and of the Council of 6 July 2011*, concerning European statistics on tourism and repealing Council Directive 95/57/EC.

Eurostat. 2021. *ESSnet Big Data II, Work package I Mobile Network Data* (all the milestones and deliverables). https://ec.europa.eu/eurostat/cros/content/wpi-milestones-and-deliverables_en.

Eurostat. 2015. "Methodological manual for tourism statistics. Version 3.1 - 2014 edition". *Manuals and Guidelines*. Luxembourg: Publications Office of the European Union.

Eurostat, and the European Statistical System - ESS. 2017. *European Statistics Code of Practice. For the National Statistical Authorities and Eurostat (EU statistical authority)*. Adopted by the European Statistical System Committee 16th November 2017. Luxembourg: Publications Office of the European Union.

Guidotti, R., R. Trasarti, M. Nanni, F. Giannotti, and D. Pedreschi. 2017. "There's a Path for Everyone: A Data-Driven Personal Model Reproducing Mobility Agendas". *Conference Paper:* 303-312. 4th Institute of Electrical and Electronics Engineers - *IEEE International Conference on Data Science and Advanced Analytics (DSAA 2017)*, Tokyo, Japan, 19-21 October 2017.

Kuusik, A., K. Nilbe, Tanel Mehine, and R. Ahas. 2014. "Country as a Free Sample: The Ability of Tourism Events to Generate Repeat Visits. Case Study with Mobile Positioning Data in Estonia". *Procedia - Social and Behavioural Sciences*, Volume 148: 262-270.

Nurmi, O., and P. Piela. 2019. "The Use of Mobile Phone Data in Tourism Statistics". In *Proceedings Special Topic Session*, Volume 4: 135-142. 62nd ISI World Statistics Congress 2019, Kuala Lumpur, Malaysia, 18-23 August 2019. Malaysia: Department of Statistics Malaysia, Central Bank of Malaysia, and Malaysia Institute of Statistics.

Pinelli, F., G. Di Lorenzo, and F. Calabrese. 2015. "Comparing Urban Sensing Applications Using Event and Network-Driven Mobile Phone Location Data". *Conference Paper*: 219-226. 16th Institute of Electrical and Electronics Engineers - *IEEE International Conference on Mobile Data Management*, Pittsburgh, PA, U.S.A., 15-18 June 2015.

Ricciato, F., T. Siil, R. Talviste, B. Kubo, and A. Wirthmann. 2021. "A proof-of-concept solution for the secure private processing of longitudinal Mobile Network Operator data in support of official statistics". *Joint Project* conducted by Eurostat and Cybernetica during 2020-2021 (Project Reference number ESTAT 2019.0232). https://ec.europa.eu/eurostat/cros/sites/default/files/unece2021_estat_cybernetica_v6.pdf.

Ricciato, F., F. De Meersman, A. Wirthmann, G. Seynaeve, and M. Skaliotis. 2018. "Processing of Mobile Network Operator data for Official Statistics: the case for public-private partnership". Paper presented at the DGINS Conference 2018, *The European path towards Trusted Smart Statistics*, Bucharest, Romania, 10-11 October 2018. https://ec.europa.eu/eurostat/cros/system/files/dgins2018_mno-so_ricciato_0.pdf.

Saluveer, E., J. Raun, M. Tiru, L. Altin, J. Kroon, T. Snitsarenko, A. Aasa, and S. Silm. 2020. "Methodological Framework for Producing National Tourism Statistics from Mobile Positioning Data". *Annals of Tourism Research*, Volume 81, Issue C. DOI: 10.1016/j.annals.2020.102895.

United Nations Committee of Experts on Big Data and Data Science for Official Statistics - UN-CEBD. 2021. "Mobile phone data". *Task Team of the UN-CEBD*. New York, NY, U.S.: United Nations. https://unstats.un.org/bigdata/task-teams/mobile-phone/index.cshtml

United Nations, Department of Economic and Social Affairs, Statistics Division, and World Tourism Organization - WTO. 2010. "International Recommendations on Tourism Statistics 2008". *Studies in Methods*, Series M, N. 83/Rev.1. New York, NY, U.S.: United Nations Publication.

United Nations, General Assembly. 2014. Resolution adopted by the General Assembly on 29 January 2014, N. 68/261. *Fundamental Principles of Official Statistics*. New York, NY, U.S.: United Nations. https://unstats.un.org/unsd/dnss/gp/fundprinciples.aspx.

Vodafone. 2022. *Vodafone coverage* 2022. Available at https://www.vodafone.it/portal/Privati/Vantaggi-Vodafone/rete-vodafone-copertura-mobile.

Zhang, L.-C. 2019. "On valid descriptive inference from non-probability sample". *Statistical Theory and Related Fields*, Volume 3, N. 2: 103-113. DOI: 10.1080/24754269.2019.1666241.

The joint distribution of income and consumption in Italy: an in-depth analysis on statistical matching

Gabriella Donatiello[1], Marcello D'Orazio[1], Doriana Frattarola[1], Mattia Spaziani[1]

Abstract

This article presents an application of statistical matching methods to integrate the EU Statistics on Income and Living Conditions and the Household Budget Survey with the aim of creating a synthetic dataset that can permit an in-depth multidimensional analysis of households' economic poverty in Italy. The work takes stock of previous experiences done at the Italian National Institute of Statistics - Istat and proposes a modification of a well-known approach to the statistical matching of data from complex sample surveys. The re-designed method permits to create a synthetic dataset that preserves the marginal distribution of both the target variables. The proposed method is more complex than simpler donor-imputation methods and permits taking into account the final survey weights. The higher complexity requires some additional checks when validating the results of the whole application. Preliminary results, presented in this paper, are quite promising also because the work benefits from an accurate ex ante harmonisation strategy of the reference surveys and on the collection of useful data for the application of statistical matching methods.

Keywords: Data fusion, data integration, weights' calibration.

DOI: 10.1481/ISTATRIVISTASTATISTICAUFFICIALE_3.2022.03

1 Gabriella Donatiello (donatiel@istat.it); Marcello D'Orazio (madorazi@istat.it); Doriana Frattarola (frattarola@istat.it); Mattia Spaziani (mspaziani@istat.it), Italian National Institute of Statistics – Istat.

This work is part of the Project for the production of microdata relating to household Income, Consumption and Wealth (ICW project) at national and international level.
Although this article is the result of all the authors' commitment, the Sections are attributed as following: 5, 6 and 7 to Gabriella Donatiello; 3 to Marcello D'Orazio; 1 and 2 to Doriana Frattarola; 4 to Mattia Spaziani.

The views and opinions expressed are those of the authors and do not necessarily reflect the official policy or position of the Italian National Institute of Statistics - Istat.

The authors would like to thank the anonymous reviewers for their comments and suggestions, which enhanced the quality of this article.

1. Introduction

In the last decades, the growing demand to provide data for measuring household economic well-being at the micro level has encouraged the production of integrated statistics on household income, consumption and wealth. In this context, the re-design of the social statistics framework at European and national level facilitated the use of integration techniques to exploit all information from existing data sources, with reduction in terms of costs for National Statistical Institutes and response burden for households. The micro integration is commonly performed by applying *statistical matching* (SM, also known as *data fusion*) methods that consist in an *ex post* integration of data from surveys referred to the same target population. This type of integration exploits variables observed and shared by different data sources for producing statistics on the relationship between variables not jointly observed in a survey. Starting from 2013, at the Italian National Institute of statistics - Istat we have investigated different matching techniques in order to produce joint statistics on income and consumption by integrating EU Statistics on Income and Living Conditions (EU-SILC) data, (from year 2012, with income reference year 2011, to year 2017), and the Household Budget Survey (HBS) data (from 2011 to 2016).

Application of statistical matching techniques is conditional to a series of prerequisites and assumptions. As the integration is based on a suitable set of variables shared by the two data sources, a key prerequisite is their *ex ante* harmonisation. This task was made easy by efforts undertook at national level in order to overcome the core social variables' reconciliation process required at European level. In the Italian EU-SILC and HBS, both conducted by Istat, the re-design of the survey questionnaires was the occasion to harmonise definitions and classifications of many shared variables such as demographic ones, household composition, family relationship, level of education, ILO labour status, *etc*.

The main obstacle to a successful application of statistical matching is the holding of the assumption underlying most of the SM methods, *i.e.* the fact that the relationship between the target variables not jointly observed - income and consumption in our application - could be fully explained by a subset of common variables shared by both the data sources (the *matching variables*). This statement corresponds to assume the independence between

income and consumption conditional on the chosen subset of common variables. Validity of this conditional independence (CI) assumption usually cannot be tested on the available data source, (an additional dataset where all the variables are jointly observed is necessary) and unfortunately it is seldom valid in real world applications. In particular, subject matter experts exclude that income and consumption can be independent conditional on a subset of socio-demographic variables shared by HBS and EU-SILC surveys.

Donatiello *et al.* (2014*a*) performed a statistical matching of EU-SILC and HBS at micro level by applying a *random hotdeck* procedure to impute the observed values for classes of consumption observed in HBS (donor dataset 2011 data) into the EU-SILC survey (recipient dataset 2012 data with income reference year 2011). In that occasion, the CI assumption could not be verified from the matched datasets, but it was assumed to hold as the set of matching variables included a proxy of the income variable (the household monthly income that could be reconstructed in the HBS data and directly observed in EU-SILC) that permitted to improve the accuracy of the SM results.

Donatiello *et al.* (2015) went ahead in the matching of SILC and HBS, focussing on the Renssen's weights calibrations procedure (Renssen, 1998). This is one of the few SM methods that can manage the survey weights and ensure a higher coherence of matching outputs in maintaining the distributions of the considered variables (more details on this method will be given in Section 3). This is a very appealing feature when analysing the data from complex sample surveys, where all the results are derived considering the final survey weights that reflect the probabilistic mechanism used in selecting the sample as well as weights' correction to compensate for coverage or non-response errors. Some interesting results were presented at the EU-SILC best practice workshop held in London in 2015 (Donatiello *et al.,* 2015) and at ITACOSM conference in June 2017 (Donatiello *et al.*, 2017).

In this paper, we investigate all the advantages and disadvantages related to the use of Renssen's SM method by applying it to HBS 2016 data and EU-SILC 2017 data (income reference year 2016). The objective is twofold: first to show further progress in applying the standard methods and a proposal for a modification that facilitates integration at micro level and secondly to highlight the impact of recent improvement done in re-designing the surveys in order to facilitate their *a posteriori* integration. This is the main

lesson learnt from previous matching applications, where all the *a posteriori* exercises of integrating surveys showed that a successful application of SM requires thinking at the integration in the survey design phase.

This work is structured as follows: the next Section will provide a brief description of the surveys and their recent changes also in view of integrating the data. Section 3 gives some insights of the Renssen's SM methods and of the modification introduced here to integrate household surveys at micro level. Section 4 presents the results achieved in applying the method to match HBS and SILC data and highlights the fact that the Renssen's procedure requires additional checks if compared to a "standard" assessment of the results of a SM application. Section 5 shows first results of multidimensional economic poverty and Section 6 makes the point of arrival of our work with some hints on future perspectives. Finally, Section 7 will present the concluding remarks.

2. An ex ante collection of data for micro integration purposes: EU-SILC 2017 data

In Italy, HBS and EU-SILC are carried out by Istat and cover the same population of private households. Both sample surveys use a stratified two-stage probabilistic sampling design. Primary sampling units (PSU) are the municipalities and second stage units (SSU) are the households. Inside each administrative region (estimation domain corresponding also to the primary strata), the PSU are further stratified according to their demographic size and, in order to guarantee self-weighting design, the total of residents in each stratum is approximately constant.

The Italian EU-SILC and HBS show a large number of common variables, and in year 2014, HBS experienced some important methodological improvements aiming at fostering data comparability at European level; in addition, the re-design also involved variables whose definition and observation were aligned as much as possible with that of EU-SILC.

The initial application of SM techniques in Istat (Coli *et al.*, 2005) highlighted the importance of using relevant auxiliary information to make the CI assumption holding and consequently improve the estimation of correlation between target variables. As the 2014 exercise showed (Donatiello *et al*. 2014*a*; 2016*a* and 2016*b*), including a proxy of income (a rough information of household income in classes collected in HBS) in the set of the matching variables, improves markedly the results if compared to application where this set does not include a strong proxy of one of the target variables. It should be noted that the income of Italian HBS was used by us in an experimental way, as it was not disseminated to users due to the well-known difficulties in collecting income data in a consumption survey. Considering that HBS income was not fully reliable we decided to overcome the problem by carrying out a matching based on a modified random hotdeck procedure that allowed to select donors in the same income classes or in those immediately preceding and following that of the recipient unit. Furthermore, in order to have proxy variables for SM purposes, we decided to collect some consumption variables in SILC. For this goal, on a voluntary basis, Istat implemented and tested the rolling module on Consumption & Wealth in EU-SILC 2017, agreed as part of the revision of EU-SILC within the new Framework Regulation on Social

Statistics (IESS)[2]. EU-SILC collection of variables on food consumption and transport, jointly with the already available data on housing costs, was seen as a way to provide enough information to derive a "strong" proxy of the total consumption that could be used in the SM applications. The design of the EU-SILC "Consumption & Wealth" module took stock of our previous SM exercise (Donatiello *et al.*, 2014b), where a detailed analysis of the structure of Italian HBS permitted to identify those consumption components representing good predictors for total consumption. Food and transport, as well as housing costs, were identified as the most important predictors of total consumption expenditures and for this reason these components were finally included into the 2017 EU-SILC Consumption & Wealth module[3].

2.1 Consumption variables in EU-SILC

Consumption & Wealth module in EU-SILC 2017 collected five consumption target variables: food at home; food outside home; public transport; private transport; regular savings. It is worth noting that EU-SILC annually collects an important amount of consumption expenditures, the "Total Housing Costs" (target variable HH070). This variable includes the costs of utilities (water, electricity, gas and heating) and all kind of expenses connected with the household right to live in the accommodation (for owners and tenants it includes mortgage interest payments and rent payments).

Annual target variable on housing costs together with the new variables observed with the module on food consumption and transport represent an important part of the total household expenditures in HBS (Consolini *et al.,* 2018*a* and 2018*b*). These variables gave us enough information to derive a synthetic variable that can be considered a good proxy of the household's total consumption.

2 Regulation (EU) 2019/1700 of the European Parliament and of the Council of 10 October 2019 establishing a common framework for European statistics relating to persons and households, based on data at individual level collected from samples.

3 The Italian module also included several voluntary variables, not provided for in the European Regulation but functional to the SM, such as the use of the monthly income, namely if the household spends the overall income for consumption or if it saves a part or it reduces saving. In fact, the same question was collected by HBS and Bank of Italy's SHIW. In addition, another variable helped to understand how the family finances expenditures that exceed their monthly income.

In order to compare SILC consumption variables with HBS data, we have set-up a list of "derived" consumption variables in HBS, using the same components that we collect in the new SILC module, in addition to total housing costs.

In HBS, most of the components included in the variable HH070 are collected with the exception of the tax on the main residence ("Imu" and "Tasi" taxes in year 2016) and the mortgage interest payments. In order to have a comparable measure of "Total Housing Costs", a modified variable of HH070 is derived in SILC (excluding the costs not covered in HBS) and likewise in HBS.

Then, we have estimated the correlations (Spearman and Pearson on the logarithmic of the two variables) between partial and total consumption. In both cases, we obtained a value of 0.80, remarkably close to one, which confirms a strong correlation between partial and total consumption in HBS.

This harmonised "Total Housing Costs" variable, together with food and transport expenses variables, was used to construct a variable of observed "SILC consumption", which turned out to be a very good predictor of the total consumption (Table 2.1).

Table 2.1 – Comparison between SILC and HBS of partial and total consumption
(Values in euros)

		min	Q1	mean	median	Q3	max
Partial consumption	HBS	0	561	961	862	1237	5317
	SILC	3	593	1001	881	1274	10903
Total consumption	HBS	93	1407	2482	2115	3189	18179

Source: Istat HBS 2016 and EU-SILC 2017

In our first exercises the CI assumption could not be verified from the available data, now with 2016 data on income and "partial" consumption expenditures, both available in EU-SILC 2017, allows us to roughly test the validity of this assertion. In particular, we estimated the Spearman correlation between income and "partial" consumption, controlling for matching variables used in some previous SM exercises; the results presented in Table 2.2, show correlation coefficients well far from 0 and therefore confirm that the independence between income and consumption conditional to some relevant common variables does not hold. In other words, the results of a SM exercise

integrating income and consumption data cannot be considered reliable unless the subset of matching results includes a strong proxy variable of income or consumption (or both).

Table 2.2 – Correlation coefficient between income and partial consumption by matching variables

Macro areas	Durable goods	Correlation
North	up to 5	0.40
	6	0.40
	7	0.48
	8	0.45
Centre	up to 5	0.34
	6	0.46
	7	0.42
	8	0.47
South and Islands	up to 5	0.27
	6	0.40
	7	0.39
	8	0.45

Source: Istat EU-SILC 2017

3. Statistical matching of data from complex sample surveys

As already mentioned, SM methods aim at integrating two distinct data sources, *A* and *B*, referred to the same target population with the objective of exploring the relationship between variables, *Y* and *Z*, not jointly observed in one of the sources. Integration is based on the variables (*X*) shared by the two data sources, and in particular on a suitable subset ($X_M;\ X_M \subseteq X$) of predictors of both *Y* and *Z*, denoted as *matching variables*.

A large part of the SM methods was developed to integrate data originating from simple random samples, where the values of (*X,Y,Z*) are independent random outcomes of the same (unknown) model which describes the relationship between the variables; in other words the observations in the data sources are *independent and identically distributed* (i.i.d.). Unfortunately, the data collected from National Statistical Institutes often originate from complex probabilistic sample surveys carried out on the same finite population that involve multistage and stratified cluster sampling designs, where the i.i.d. assumption is no longer valid. In fact, cluster sampling introduces dependence between units belonging to the same cluster; in addition, complex sampling designs may determine unequal units' inclusion probability.

In literature, there are relatively few statistical matching methods explicitly tailored to handle data from complex sample surveys; one of the most promising is the method suggested by Renssen (1998), based on *weights' calibration*. Calibration is a widespread practice usually employed in sample surveys to improve the precision of the final survey results (also to compensate for non-observation errors). It consists in deriving new weights, as close as possible to the starting ones, which fulfil a series of constraints concerning the totals of a set of auxiliary variables (usually known at population level).

3.1 Renssen's weights calibration procedure

Renssen's SM procedure is particularly suited to manage categorical *X*, *Y* and *Z* variables and is not primarily designed to integrate surveys at microdata level, as the main purpose is the estimation of the contingency table $Y \times Z$. This procedure is articulated in two steps; the first step consists in calibrating weights in both *A* and *B* to align the corresponding estimated

totals of the matching variables X_M (joint or marginal distribution) with known (or estimated) population totals; the second step estimates the two-way contingency table $Y \times Z$. Estimation can be done under the CI assumption:

$$\hat{P}^{(CI)}_{Y=j,Z=k} = \hat{P}^{(A)}_{Y=j|X_M=i} \times \hat{P}^{(B)}_{Z=k|X_M=i} \times \hat{P}_{X_M=i} \quad | \; i=1,\ldots,I \;, j=1,\ldots,J \;, \; k=1,\ldots,K \tag{1}$$

whereas the terms in the formula are obtained by considering the units' weights (w'), modified after the first harmonisation step, *i.e.*:

$$\hat{P}_{X_M=i} = \sum_{a=1}^{n_A} w'_a \, I(x_{Ma} = i) \tag{2}$$

Renssen's approach also allows exploiting a third additional data source *C* in which *Y* and *Z* are jointly observed. A first option consists in estimating $Y \times Z$ directly on *C* after a further calibration step performed on it, aimed at aligning the marginal distributions of *Y* and *Z* with the corresponding ones estimated in respectively *A* and *B* after the initial harmonisation (*incomplete two-way stratification*). In alternative, it is possible to perform the *synthetic two-way stratification*, which consists in "correcting" the estimate obtained under the CI assumption with the additional information provided by *C*. In such a case, *C* must also include the matching variables. In practice, also this alternative procedure consists in a series of calibration steps.

A very appealing feature of the Renssen's procedure is that the marginal distributions of the resulting contingency table $Y \times Z$ are aligned with those estimated on the starting datasets, but after the initial harmonisation step. This is a very important characteristic in official statistics where the coherence of the final statistical outputs is one of the key dimensions of the quality of statistics and plays a crucial role when integrating data from sources referred to the same target population.

Although the whole Renssen's procedure is designed with a macro purpose (estimating the contingency table $Y \times Z$) it also allows to perform imputation at micro level. Micro objective is pursued by generating the predicted values of the *linear probability models* that are fit across the whole procedure. A linear probability model assumes that the probability of an event (falling in a given consumption or income class in our case) can be expressed as a linear combination of a series of explanatory variables (matching variables in our application). These models are not the ones suited

to deal with this case[4], but they are used as "working" models because the corresponding predicted values (the estimated probability of assuming each of the categories of Z (Y) for every unit in A (B)) maintain the appealing feature of the whole procedure. In other words, when used to estimate the marginal distribution of Z (Y) in A (B) they return the same estimated distribution that is achieved by considering the data set B (A) (after the harmonisation). Unfortunately, the estimated probabilities provided by linear probability models should be used carefully, given that these models present some well-known drawbacks (estimated probabilities can be less than 0 or greater than 1; heteroskedastic residuals, *etc.*).

Practically, having the predicted probabilities at the end of SM may not be a viable option for the practitioner that would prefer having imputed categories for the target variable for easing the subsequent analyses. In this sense, a "direct" imputation would consist in adopting a randomised device that generates the imputed category by a random draw with probabilities proportional to the predicted probabilities; this would avoid the well-known negative consequences of getting the "most voted" category (the one with the highest predicted probability), as also shown by Donatiello *et al.* (2016a). Renssen (1998) suggests an "indirect" two-step imputation that consists in a mixed SM micro procedure; in practice, the estimated predictions are the input of a *nearest neighbour hotdeck* procedure (Singh *et al.*, 1993) where the final value to impute is the one observed on the closest donor according to the distance calculated on the predictions. This is one of the possible many variants of the SM mixed methods listed in Section 2.5 of D'Orazio *et al.* (2006). SM mixed methods are mainly developed for continuous target variables but can be easily adapted to handle predicted probabilities for the categories of a categorical target variable, as it will be shown in the next section.

It is worth noting that Renssen's SM allows the introduction of target continuous variables but in this case, some difficulties may arise in the various subsequent calibration steps. We are currently investigating the possibility of applying the micro "extension" of the original proposal to the case of continuous target variables and some preliminary results are quite satisfactory (Donatiello *et al.*, 2017), but they will not be presented in this

4 Some suggestions related to use of models in SM when the target variables is a categorical response variables are in de Waal (2015).

article as additional investigation is deserved. Essentially, in this work we consider continuous *Y* and *Z* target variables, but for the application of the Renssen's SM procedure we categorise them, although the procedure is designed to end up with a synthetic fused dataset with continuous values for the imputed missing variable to be fully used for validation and economic analysis.

3.2 Matching of EU-SILC and HBS

SM exercise aimed at imputing the household consumption variable (*Z*), originally observed in HBS (donor), in the SILC survey (households). This setting is not optimal as contradicts the common suggestion of using the smaller dataset as the recipient ($n_B = 15\,409$ households in HBS vs. $n_A = 22\,226$ in SILC). However, the difference is not so huge and not very relevant as the two-step mixed procedure is used instead of a "standard" SM *hotdeck* procedure. Specifically, two different variants of the Renssen's procedure are proposed for the final imputation of *Z* into SILC. The first step follows the Renssen's recommendations and is common to both the procedures; it consists in the calibration of the survey weights of both the data sources to reproduce the same estimated marginal distribution of the matching variables. Following the previous matching exercises, we identified few relevant matching variables: the geographical areas (5 categories, "North-West Italy", "North-East Italy", "Centre Italy", "South of Italy", "Islands of Italy") and the number of durable goods owned by the households (4 categories). In addition, to have the CI holding the proxy variable of household consumption is included in the set of matching variable; this variable, C^*, is a categorised version (8 categories) of the consumption variable observed in the new SILC rolling module on Consumption & Wealth; for matching purposes, the same consumption variable is derived in HBS using data collected from the survey.

Marginal distributions of the three matching variables are aligned to reproduce the fixed totals. In particular, the reference distributions of the Italian households by geographical areas are achieved by pooling the estimates provided by the starting data sources; the same procedure is used for the number of durable goods, while the reference distribution of households by classes of proxy consumption C^* is estimated on the HBS starting data.

The second step of the matching procedure consists in estimating the contingency table $Y^* \times Z^*$, being Y^* and Z^* the categorised versions of respectively *Y*, the household income provided by SILC, and *Z*, the household consumption expenditures observed in HBS. A first estimate of the $Y^* \times Z^*$ table can be achieved without integration at micro level by applying the expression (1), since the CI assumption can be considered valid as the set of matching variables includes a proxy oh the household consumption. As expected, this table has marginal distributions of both Y^* and Z^* that are equal to the ones provided by the starting data sources after the initial harmonisation step.

Then, as a by-product of the estimation of $Y^* \times Z^*$, we derive predictions of $\hat{p}(y^* = j)$ $(j = 1,2, \dots, J = 8)$ and $\hat{p}(z^* = k)$ $(k = 1,2, \dots, K = 8)$ for the units in both the data sources. Finally, for imputing the values of *Z* in SILC the following additional steps are proposed:

Step 3.1): imputation in SILC of the value of *Z* observed on the closest donor in HBS according to the following distance:

$$\hat{\Delta}_{a,b} = \frac{1}{2}\sum_{k=1}^{K} |\hat{p}(z_a^* = k) - \hat{p}(z_b^* = k)| \qquad (3)$$

In this expression $\hat{p}(z_a^* = k)$ is the predicted probability that the ath unit in *A* $(a = 1, \dots, n_A)$ gets an imputed category equal to *k* for the variable *Z*; similarly, $\hat{p}(z_b^* = k)$ is the same predicted probability for the *b*th unit in *B* $(b = 1, \dots, n_B)$.

The distance function (3) corresponds to the *total variation distance or dissimilarity index* and was chosen as the distance should be calculated on predicted probabilities. In fact, as also noted by de Waal (2015), this specific situation requires the adoption of suitable distance functions aimed at measuring dissimilarity between distributions of categorical variables (for a review of these distances see *e.g.* Cha, 2007). An alternative to the total variation distance can be the Hellinger's distance, but we opted in favour of the total variation distance because it corresponds to ½ of the Manhattan distance and this latter one is already implemented in many statistical packages.

Step 3.2): imputation in SILC of the value of *Z* observed on the closest donor in HBS according to the sum of the total variation distances related to predictions of both *Y* and *Z*:

$$\hat{\Delta}_{a,b} = \frac{1}{2}\sum_{j=1}^{J} |\hat{p}(y_a^* = j) - \hat{p}(y_b^* = j)| + \frac{1}{2}\sum_{k=1}^{K} |\hat{p}(z_a^* = k) - \hat{p}(z_b^* = k)| \qquad (4)$$

The steps (3.1) and (3.2) are alternative implementation of the final step of a SM mixed approach based on predictive mean matching, following suggestions in D'Orazio *et al* (2006). Such a mixed SM procedure joins the advantages of both parametric and nonparametric approaches; in particular, the final phase permits to exclude the matching variables from the computation of distance and is robust to model misspecification. In the presence of several potential donors at the minimum distance from the *a*th recipient unit, for matching purpose just one the donors is picked up completely at random.

As, in the end, the SM procedures described in this Section resemble a SM mixed approach it was decided to compare their results with those of a "standard" mixed approach that does not take into account the survey weights and the constraint of preserving the marginal distributions of the target variables. As shown in D'Orazio *et al* (2006, Section 2.5.1), several variants of the regression-based mixed approach exist when the target variables are continuous, for comparability purposes the choice fell on MM3 and MM5. Specifically, the regression step (step 1) of MM3 consists in fitting in HBS a linear regression model, where *Z* (log transformed) is predicted by the chosen matching variable (the log transformed version of the proxy variable of household consumption *C* is considered). The fitted model is then used to derive in SILC the "intermediate" values of *Z* ($\tilde{z}_a = \hat{z}_a + e_a$) obtained by summing the predicted values of *Z* in SILC ($\hat{z}_a$) with a random error term (e_a) generated from a gaussian distribution with mean 0 and residual standard deviation of the model. The matching step (2nd step) imputes in SILC the value of *Z* observed on the closest donor in HBS according to the distance $d_{a,b} = |\tilde{z}_a - z_b|$. It should be noted that the procedure MM5 also fits in SILC a regression model of *Y* (log transformed) vs the matching variables and then uses it to estimate the intermediate values of *Y* in HBS ($\tilde{y}_b = \hat{y}_b + e_b$). In the matching step of MM5 method the value of *Z* imputed in SILC is the one observed on the closest donor in HBS according to the distance $d_{a,b} = |y_a - \tilde{y}_b| + |\tilde{z}_a - z_b|$ (in the proposed MM5 procedure the matching step is constrained to use donors only once, but this is not possible in our application because SILC is larger than HBS).

Matching exercise was performed in R using the facilities of the **StatMatch** package (D'Orazio, 2022). The results of the proposed method and a comparison with the mixed approach that does not take into account the survey weights are presented in the next section.

4. Main results of the SM application

Typically, the assessment of results of a SM application at micro level consists in analysing whether the synthetic dataset, *i.e.* SILC with imputed household consumption, preserves the marginal distribution of the imputed variable and the relationship of this variable with the matching variables. These checks are necessary because it is not possible to assess the accuracy of the estimates involving the imputed variable obtained at the end of the whole SM procedure by estimating the MSE or just the variance. This is still an open problem of SM applications, where an indirect partial assessment of the variability associated to the final estimates can only be obtained when approaches based on assessment of SM uncertainty are applied (see *e.g.* Conti *et al.* 2012; Zhang, 2015).

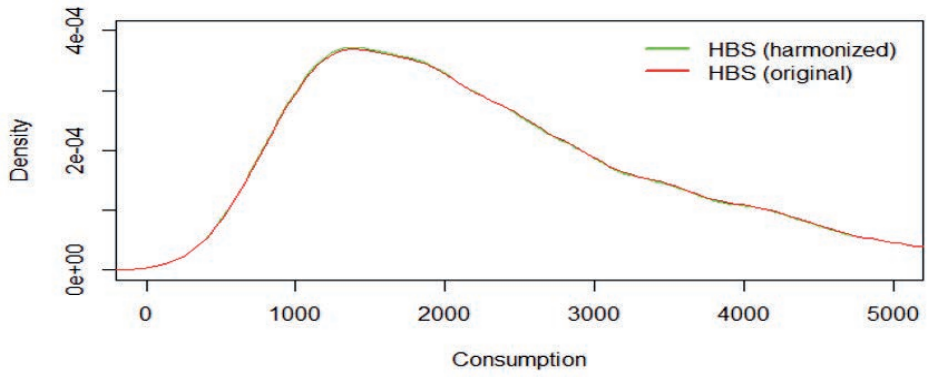
In addition, subject matter experts try to assess also the "plausibility" of the joint relationship between the imputed target variable, the total household consumption in our case, and the other target variable, *i.e.* the household income.

In our application, we believe that additional checks are required as the synthetic data set is the outcome of a complex SM procedure, whose first step modifies (calibration) the starting survey weights with the aim of harmonising the marginal distribution of the matching variables. These modified weights are then the ones to be used when analysing the data starting from the synthetic data set.

In this respect, a first check consists in assessing whether the initial calibration of the survey weights introduces significant changes in the marginal distributions of the target variables.

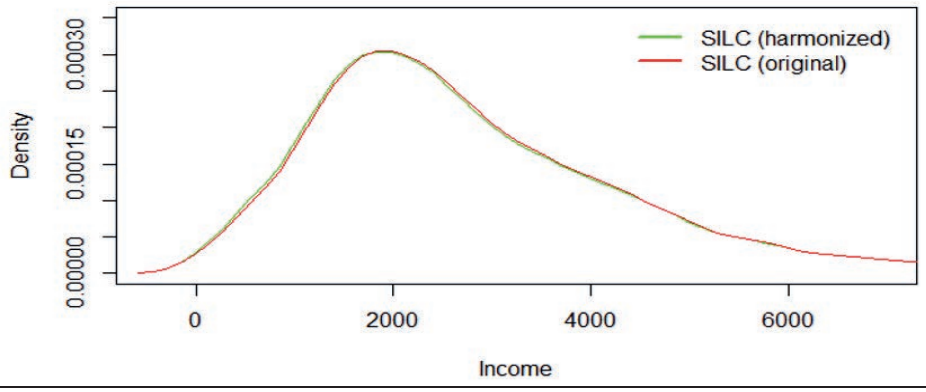
Figures 4.1 and 4.2 show that the estimated distributions of both the target variables remain almost unchanged considering both original and modified weights.

Figure 4.1 - Comparison of total consumption in HBS before and after the initial harmonisation step



Source: Istat HBS 2016

Figure 4.2 - Comparison of total income in SILC before and after the initial harmonisation step



Source: Istat EU-SILC 2017

An additional check consists in investigating the preservation of the joint distribution between the discretised target variable and some of the available common variables. Change in the joint distribution is measured by the Hellinger distance (HD) between the distribution estimated before and after the initial harmonisation step. As shown in Table 4.1, the distances are well below the 5% rule-of-thumb threshold, indicating that the modification of the weights does not affect markedly the joint distribution between the discretised target variable

and each of the considered common variables. The highest value for the HD, but still below the 5% threshold, is observed in EU-SILC for the joint distribution of discretised total income and discretised partial consumption estimated with the data collected in the new module. This is somehow expected since this partial consumption variable derived in SILC cannot be considered as accurate as the corresponding one observed in HBS[5]. For this reason, in the harmonisation step the weights were modified to return the distribution estimated from the HBS rather than, as usual, the pooled estimate.

Table 4.1 - Hellinger Distance of the joint distribution of income and consumption classes before and after the harmonisation step by common variables

	HBS	SILC
Partial consumption	0.5%	4.5%
Durable goods	1.8%	1.9%
Macro areas	0.3%	0.8%
Sex	0.3%	0.9%
Education	0.5%	1.0%
Citizenship	0.4%	0.9%
# people in household	0.4%	1.2%
Tenure status	0.4%	1.1%
# employed people	0.4%	1.0%
Household type	0.4%	1.2%

Source: Istat HBS 2016 and EU-SILC 2017

Finally, as HBS is the main reference for estimating the poverty, we have compared the relative and absolute poverty incidence estimated in HBS before and after the harmonisation step. Results (see Table 4.2) are quite satisfactory as there is a very slight change in the fraction of poor households. This result, however promising, is to be taken with caution, as HBS is the sole survey entitled for the dissemination of estimates on relative and absolute poverty in Italy. It should be noted that Table 4.2 also reports the estimate of absolute poverty derived on the synthetic data set at the end of the SM procedure. In this case, the estimates provided by step (3.1) show a small change compared to the reference ones estimated in HBS that however is about 0.2%.

Analysing more in detail the synthetic data set, Figure 4.3 shows the estimated density functions of total consumption imputed in SILC with respectively the

5 There are some relevant differences such as the method of data collection, in fact HBS uses a diary while the consumption collected in SILC is obtained through few direct questions. Furthermore, the HBS is a continuous quarterly survey while SILC is annual.

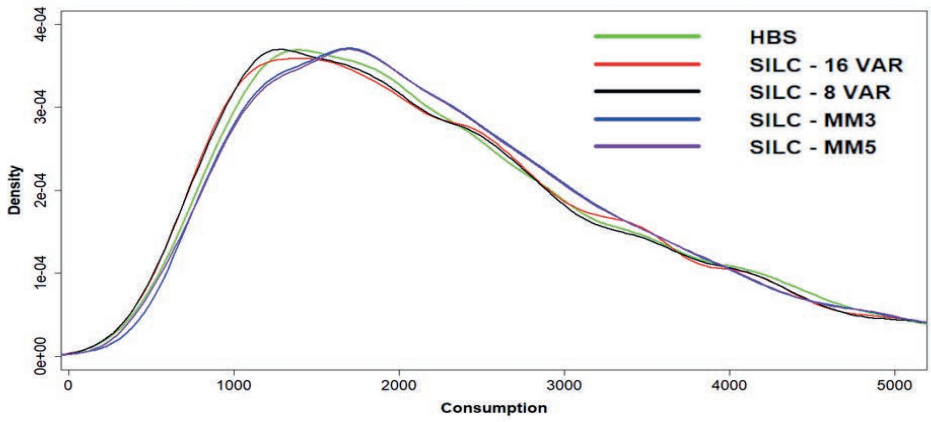
Table 4.2 – Poverty indicators by type of weight before and after the harmonisation step in the synthetic data set (percentage values)

	Origin survey weights	Weights after harmonisation	
	HBS	HBS	FUSED
Relative poverty	10.60	10.64	10.82
Absolute poverty	6.28	6.30	6.12

Source: Istat HBS 2016 and EU-SILC Fused 2017

proposed procedure with slight modification of the final step, (3.1) (nearest neighbour donor with distance calculated on the predicted probability of falling in one of the categories of the variable *Z*, denoted as "8 VAR" in Figures 4.3 and 4.4) and (3.2) (nearest neighbour donor with distance calculated on the predicted probabilities for both *Y* and *Z* variables; denoted as "16 VAR" in Figures 4.3 and 4.4), and the procedures MM3 and MM5. All these estimated distributions are compared to the consumption distribution measured in HBS. It is worth noting that the distribution estimated with the imputed values provided by step (3.1) is closer to the original distribution than the one estimated using imputed values given by step (3.2). On the contrary, the procedure MM3 and MM5 return an imputed consumption whose distribution is shifted toward the right, that tends to overestimate the overall consumption. These results clearly show that the additional step of harmonising the weights through the Renssen's method improves the final estimates.

Figure 4.3 - Comparison of original HBS and imputed total consumption in SILC by our method (distance function 8 or 16 dummies) **and mixed procedures** (MM3 and MM5)



Source: Istat EU-SILC 2017

Moreover, another "traditional" check to evaluate the outputs of the SM application consists in comparing the joint distributions of the imputed variable and each of the most relevant common variables.

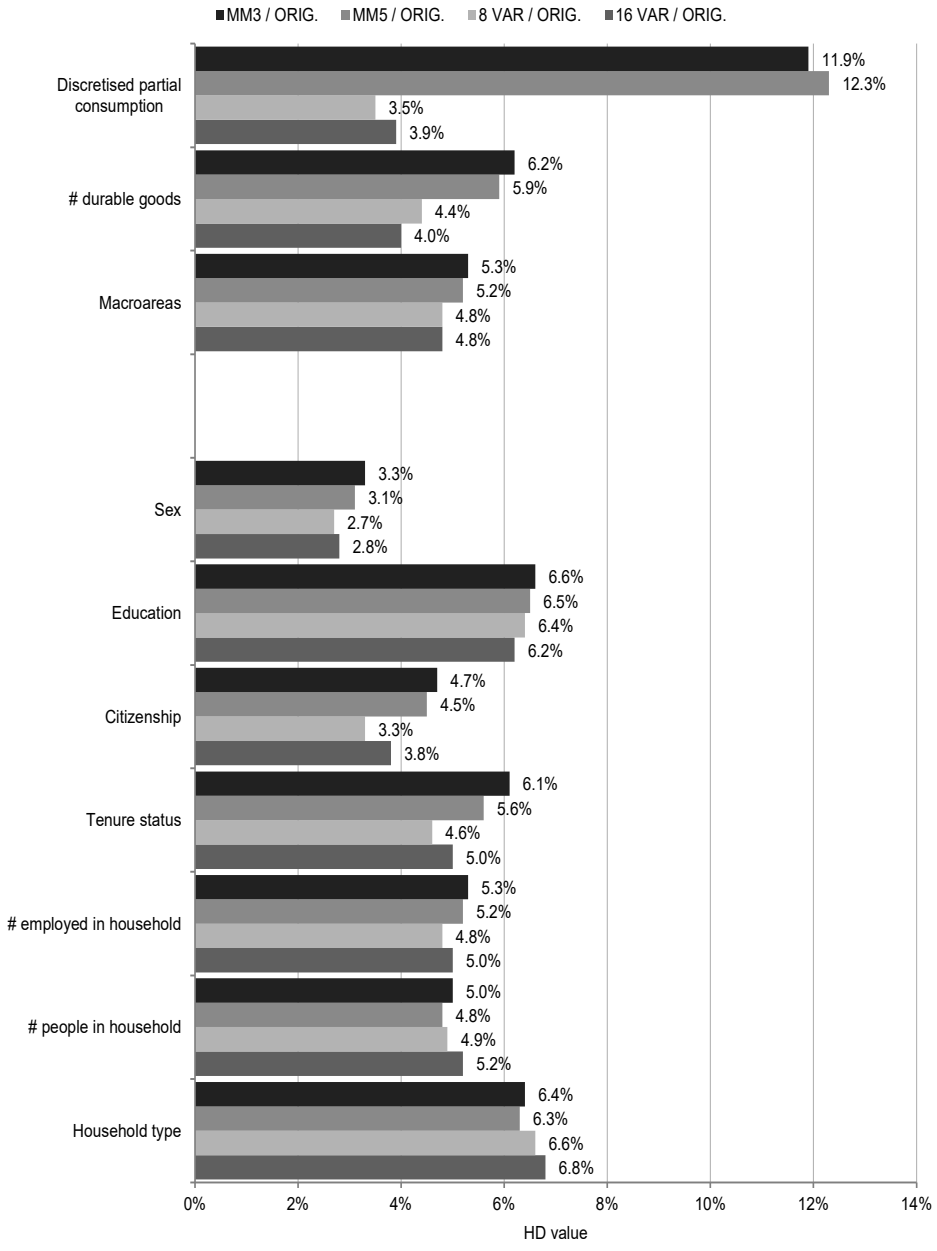
Figure 4.4 shows the HD between the distributions estimated respectively from the synthetic data set and from HBS for each combination of the quintiles of total consumption and some of the common variables (including the matching variables). Cases with the HD below 5% (the chosen threshold) indicate closeness between the distribution estimated on the fused dataset and the reference one (estimated on HBS). Distances over the threshold are observed when crossing the consumption quintiles with the household type, the level of education and with the proxy of consumption but only with procedures MM3 and MM5. In the first two cases the outcome is presumably due to the fact that the variables household type and education level were not used as matching variables as well as to differences in how they are observed in the two surveys. The failure of MM3 and MM5 in preserving the joint distribution of the overall consumption with its proxy in SILC, seems mainly explained by the some difficulties in imputing reliable low levels of overall consumption.

In general, Figure 4.4 shows that imputation using the proposed method with distance on the predicted classes of the discretised overall income (step 3.1) tends to perform better in preserving the relationship of consumption and some of the relevant common variables in SILC. On the contrary, methods MM3 and MM5 are often performing worse.

As the post matching checks show that the proposed procedure with final step (3.1) ("8 VAR" in Tables or Figures) tends to perform better that the other applied methods, all subsequent analysis will consider only the imputed SILC data set at the end of this procedure.

Table 4.3 presents means and medians of the imputed consumption in SILC at the end of step (3.1) and of the observed consumption in HBS by matching variables and some common variables; the relative differences between 5% and 10% are highlighted in light red and the ones over 10% in dark red. As expected, the larger differences are observed in correspondence of the same variables identified when calculating the HD between estimated and reference joint distributions (Figure 4.4). In general, the comparability is high with many differences under 5% and a difference for the total distribution close to 1%.

Figure 4.4 - Hellinger distance comparing estimated joint distribution crossing income quintiles and common variables in the synthetic data set and in HBS



Source: Istat HBS 2016 and EU-SILC Fused 2017

Table 4.3 - Comparison of imputed and original total consumption (Z) by common variables (Values in euros)

	Imputed Z in SILC		Z in HBS		Imputed Z / Z	
	Mean	Median	Mean	Median	Mean	Median
DURABLE GOODS						
up to 5	1370	1165	1448	1257	95	93
6	2063	1789	2151	1850	96	97
7	2774	2466	2743	2464	101	100
8	3563	3210	3581	3290	99	98
MACRO AREAS						
North	2589	2236	2713	2384	95	94
Centre	2535	2134	2554	2156	99	99
South and Islands	2144	1841	2028	1812	106	102
SEX						
Man	2652	2321	2642	2300	100	101
Female	2053	1663	2143	1792	96	93
EDUCATION						
Less Then Primary	1740	1400	1703	1447	102	97
Primary	2397	2037	2255	1956	106	104
Secondary	2706	2387	2722	2396	99	100
Post-Secondary or Upper	3005	2650	3465	3105	87	85
CITIZENSHIP						
Italian	2490	2137	2536	2185	98	98
Foreign	1807	1502	1640	1305	110	115
TENURE STATUS						
Owner	2059	1747	1857	1574	111	111
Rent	2617	2288	2674	2327	98	98
Usufruct	2015	1669	2130	1785	95	94
# EMPLOYED PEOPLE						
0	1805	1484	1991	1649	91	90
1	2487	2159	2410	2097	103	103
2	3340	3005	3391	3091	98	97
3 or more	3891	3478	3794	3584	103	97
# PEOPLE IN HOUSEHOLD						
1	1644	1378	1760	1467	93	94
2	2351	2060	2555	2196	92	94
3+	3182	2843	3023	2685	105	106
HOUSEHOLD TYPE						
Single	1644	1378	1690	1514	97	91
Couples Without Children	2372	2098	2528	2151	94	98
Couples With Children	3226	2875	2948	2607	109	110
Single Parent	2487	2179	2432	2119	102	103
Others	2755	2455	2598	2218	106	111
TOTAL	2437	2080	2471	2107	99	99

Source: Istat HBS 2016 and EU-SILC Fused 2017

Table 4.4 – Comparison of Propensity to consume by data source and common variables

	FUSED	HFCS
	APC	APC
DURABLE GOODS		
up to 5	0.78	-
6	0.86	-
7	0.86	-
8	0.83	-
MACRO AREAS		
North	0.81	0.73
Centre	0.82	0.78
South And Islands	0.91	0.80
SEX		
Man	0.82	0.75
Female	0.87	0.79
EDUCATION		
Less Then Primary	0.80	0.81
Primary	0.94	0.81
Secondary	0.86	0.74
Post-Secondary or Upper	0.71	0.69
CITIZENSHIP		
Italian	0.83	-
Foreign	1.06	-
TENURE STATUS		
Owner	1.09	-
Rent	0.79	-
Usufruct	0.88	-
# EMPLOYED PEOPLE		
0	0.82	-
1	0.93	-
2	0.77	-
3 or more	0.69	-
# PEOPLE IN HOUSEHOLD		
1	0.89	-
2	0.78	-
3+	0.85	-
HOUSEHOLD TYPE		
Single	0.89	-
Couples Without Children	0.75	-
Couples With Children	0.84	-
Single Parent	0.92	-
Others	0.80	-
QUINTILES OF INCOME		
1° quintile	1.61	1.21
2° quintile	1.03	0.90
3° quintile	0.93	0.83
4° quintile	0.83	0.75
5° quintile	0.59	0.64
TOTAL	0.84	0.76

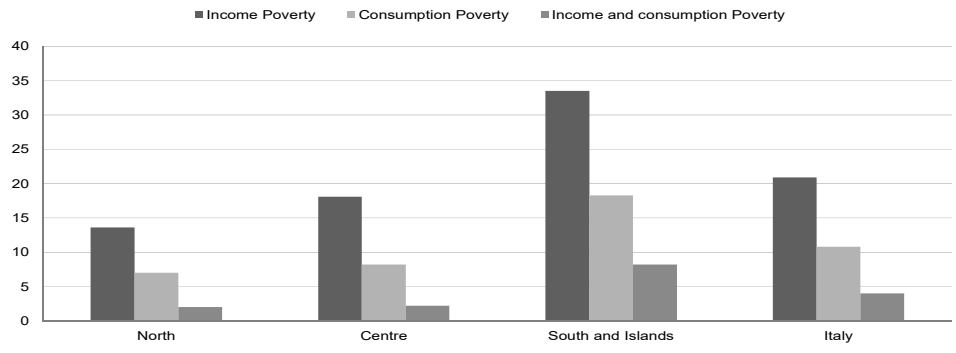
Source: Istat EU-SILC Fused 2017 and Bank of Italy HFCS 2016

Finally, to summarise the relationship between income and consumption in the synthetic data set we have estimated the propensity to consume by some common variables (Table 4.4). Results are between 0 and 1 for most cases with the overall propensity to consume equal to 0.84. This is a very indicative first result, to be considered rather as an exercise and not as a reliable estimate. In any case, as expected, the propensity to consume decreases as the income increases: when crossed with the quantile of income, the propensity to consume ranges from 1.61 for the first quintile to 0.59 for the richest quintile. Moreover, for households who have to pay the rent, the propensity to consume is significantly lower than that of the owners (0.79 *vs*. 1.09). Propensity to consume estimated by Bank of Italy with the European Central Bank's Household Finance and Consumption Survey (HFCS) is presented for comparison, as it represents the unique source in which household income and consumption expenses are jointly observed. Overall propensity is 0.76, therefore lower than our findings, but it is worth considering that comparability is affected by the different way of estimating the consumption based on the consumption expenses collected through the survey and the fact that usually HFCS provides lower consumption levels if compared to HBS, that provides the reference estimates.

5. Initial comparative analysis of poverty

In general, all the various checks done on synthetic data set obtained at the end of the Renssen's SM procedure, with modifications suggested in this paper, indicate that the results go in the desired direction of providing a reliable picture of joint distribution of income and consumption. For this reason, although at a very early stage, it is possible to draft some considerations on one of the main objectives that the analysts would like to achieve by analysing the joint distributions of income and consumption that is to have more insights on economic poverty[6].

Figure 5.1 - Income and consumption poverty by macro areas (percentage values)



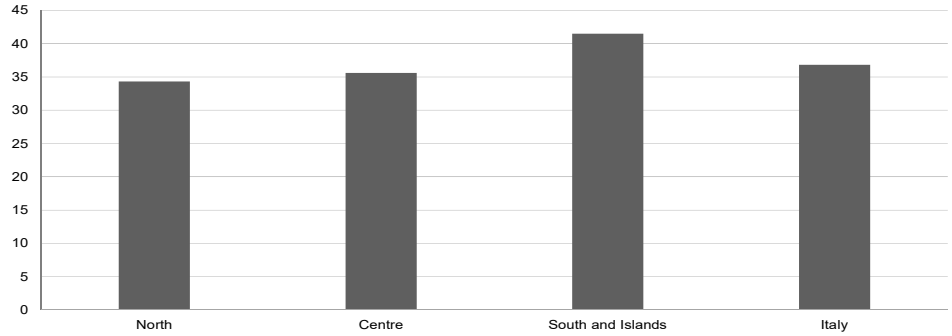
Source: Istat EU-SILC Fused 2017

Preliminary data suggest consumption poverty lower than those for income with overlap between income and consumption poverty around 4% of households (Figure 5.1).

Share of households with expenditure higher than income is an important indicator as the households that are unable to finance consumption entirely from income may face financial difficulties and may reduce their assets. Estimates calculated on the synthetic data source are in the Figure 5.2 and indicate that more than a third of households have consumption above income, especially in the South and Islands. These are findings to be interpreted very carefully as, especially for low-income households, it is known that income is more frequently underestimated while consumption is usually reported more accurately.

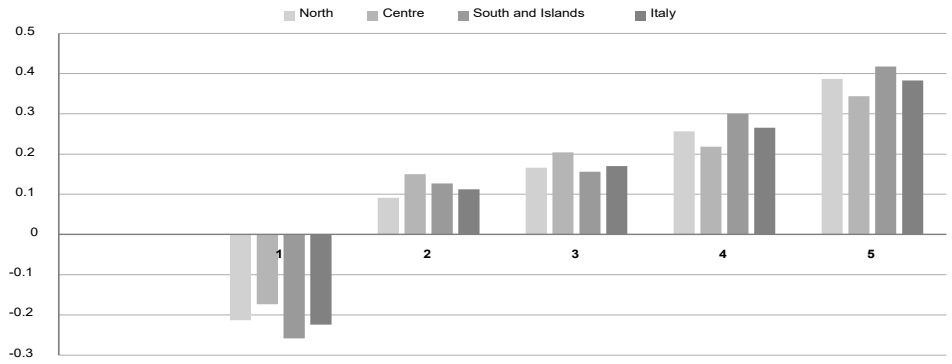
6 Scientific community has long agreed that income can be considered a good proxy for living standards but becomes a better measure when it is associated with data on household consumption and wealth. See: Stiglitz-Sen-Fitoussi Report 2009 and OECD 2013.

Figure 5.2 - Households with income less than consumption by macro areas (percentage values)



Source: Istat EU-SILC Fused 2017

Figure 5.3 - Median income saving rate by income quintiles and macro areas (percentage values)



Source: Istat EU-SILC Fused 2017

In Figure 5.3, data show a strong relationship between saving rates and income quintiles with low saving rates in Q1 that seem to reflect temporary low incomes but, as mentioned above, also potential under-reporting of low incomes/ high expenditures. A significant proportion of households have negative saving rates that, although in line with Eurostat's experimental estimates[7], need to be interpreted with caution.

7 Eurostat's experimental estimates on the joint distribution of income, consumption and wealth are available at: https://ec.europa.eu/eurostat/web/experimental-statistics/income-consumption-and-wealth.

6. Where we stand and the way forward

The construction of a data set containing the joint information on household income and consumption and the estimates provided by it are part of the project for the production of microdata relating to household income, consumption and wealth in Italy (ICW project). The current achievements reflect the work done over the past few years and take stock of all the findings of the various SM exercises done at Istat starting from 2005. Essentially, the proposed SM approach relies on three key choices: (*i*) an *ex ante* harmonisation of the social survey EU-SILC and HBS, (*ii*) the collection through an *ad hoc* module of information relating to the dimensions of consumption and wealth in the income survey EU-SILC, and finally (*iii*) the application of statistical matching techniques that take into account the survey weights.

Harmonisation of the Italian EU-SILC and HBS, started in our Institute in 2011, nowadays is quite extensive covering the sample design, concepts, definitions and consistent treatments and classifications. From the beginning, the harmonisation of the two surveys was also made to facilitate the application of micro-integration methods. Furthermore, the SILC modular approach allowed collecting variables relating to consumption and wealth, which, although limited in number and generally less accurate than data collected using an *ad hoc* survey, have proved to be important as hook variables in the matching procedures.

Statistical matching techniques we used are based on conditional independence assumption. It is well-known that such integration methods are considered a *second best* choice since it is not possible to fully capture the relationships between all the variables of interest conditional on a relatively small set of common matching variables. However, the SILC consumption and wealth module enabled us to test the efficacy of the use of proxy variables of the targets as matching variables capable of justifying the CIA.

As known, the results stemming out from micro integration techniques must be carefully evaluated in order to assess the validity and plausibility of the synthetic dataset, before they can be used for policy purposes and to design measures to fight poverty and material deprivation. Hence, for the validation of our results we used all the criteria suggested in literature (Rassler, 2002) and decided to add some additional checks.

Looking ahead, further efforts should be made to improve the SM method presented in this paper. In particular, care is needed in the initial weights calibration to harmonise the marginal (or joint) distributions of the chosen matching variables. This is not a straightforward step because it requires the estimation of the reference marginal distributions (reference totals) and may not end successfully, *i.e.* the calibration may not converge or may return negative weights (can happen with some calibration functions). This problem generally worsens by increasing the number of matching variables; therefore, also in this case, the main recommendation is to select few relevant matching variables. In addition, this initial weights' calibration may affect the results of the whole SM procedure if it introduces marked changes in the marginal distributions of the target variables. This issue often is not taken into account and we believe that it deserves special attention in the assessment of the results of this specific SM procedure. In our case, the initial harmonisation step came out to be a "cosmetic" weights' calibration which in fact has not introduced significant changes in the distributions of the relevant variables, as well as in the traditional poverty indicators. Furthermore, we have verified that the Renssen weight harmonisation step, together with the variants proposed by us, provide better results than the more traditional mixed methods. The comparison between our method and the MM3 and MM5 procedures clearly demonstrates that the distributions obtained with our method are closer to the reference one of HBS. The traditional mixed procedures, without the additional weight harmonisation step, tend to substantially overestimate the total consumption, as they likely fail in estimating the low incomes of the distribution.

In this work, we have essentially investigated the possibility of modifying the origin Renssen's method to create a synthetic dataset by imputing a continuous rather than a categorical variable. For this purpose, we passed through a discretisation of the target variables but we believe that in the future this step can be skipped, although some additional investigation is deserved.

Future plans also consider the possibility of producing a synthetic data set that includes the dimension related to the household assets. These data are well observed in the Bank of Italy's Survey on Household Income and Wealth but the exploitation of this additional survey in the SM exercise presents a number of methodological challenges. This is an unprecedented activity, which, however, can benefit of the experience accumulated in our previous matching exercises.

The final goal of our work is the production of microdata on the joint distribution of household income, consumption and wealth, which will represent the basis for producing experimental statistics, expected to be disseminated in aggregate tables but only after an accurate phase of validation of their reliability. We are aware that the latter issue is not straightforward due to complexity and underlying assumptions in the application of statistical matching techniques. For these reasons, particular attention will be paid to informing external users on the nature of the synthetic data produced, on the model applied and on the quality assessment.

7. Concluding remarks

Availability of joint micro data on income and consumption is fundamental to measure the poverty and the living conditions of households, overcoming the measures used up to now based on the observation of a single dimension (income or consumption). Statistical matching methods presented in this work seem particularly promising although show some methodological constraints and require additional checks for assessing the plausibility of the final estimates. It is therefore necessary to point out that our results are experimental and here are presented to highlight their potential advantages for the economic analysis. Although the general assessment is based on metrics recognised in the literature, we believe it is important to deepen the validation phase of the experimental statistical outputs.

Initial analysis of our results shows that the reproduction of the marginal distribution of the imputed consumption variables by the statistical matching turned out to be satisfactory. Comparison of the estimated probability density functions for the synthetic consumption variables and the original ones shows a good overlap especially on the tails of the distribution. In addition, we have also analysed the correlation structure between the target variables observed in the original and fused datasets. An effective match should lead to similar relationships between common and target variables in the donor and the matched file; the results obtained in our application show that the sign and order of magnitude of the correlation/association remain almost the same in both datasets. Moreover, the preservation of the joint distribution of the quintile of total consumption with each of the considered common variables show a level of comparability, which can be considered quite good. Synthetic dataset obtained at the end of the SM procedure permits a first multidimensional analysis of poverty and makes it possible to highlight areas of vulnerability for households. Data that have been analysed in this work refer to 2016 and, even with numerous caveats, allow us to make some initial analyses on economic poverty by studying jointly two dimensions of household conditions. The economic situation like the current one has certainly increased the request of multi-dimensional measures to assess the ability of households to support their living standards and to cope with important economic shocks. For this reason, we intend to apply the SM procedures presented here to the data referred to year 2020 to

analyse the impact of the pandemic on the resilience and saving capacity of households, examining joint information on income, consumption and the role of household wealth in mitigating the effects of the crisis.

References

Cha, S.-H. 2007. "Comprehensive Survey on Distance/Similarity Measures between Probability Density Functions". *International Journal of Mathematical Models and Methods in Applied Sciences*, Volume 1, Issue 4: 300-307

Coli, A., F. Tartamella, G. Sacco, I. Faiella, M. Scanu, M. D'Orazio, M. Di Zio, I Siciliani, S. Colombini, e A. Masi. 2005. "La costruzione di un archivio di microdati sulle famiglie italiane ottenuto integrando l'indagine Istat sui consumi delle famiglie italiane e l'indagine Banca d'Italia sui bilanci delle famiglie italiane". *Documenti*, N. 12/2006. Roma, Italy: Istat. https://www.istat.it/it/archivio/219083.

Consolini, P., G. Donatiello, D. Frattarola, and M. Spaziani. 2018*a*. "The Consumption and Wealth data in IT-SILC 2017". *Proceedings of the Workshop on Best Practices for EU-SILC Revision*, Warsaw, Poland, 16th – 17th October 2018.

Consolini, P., G. Donatiello, D. Frattarola, and M. Spaziani. 2018*b*. "The IT-SILC measurement of the household finance, wealth and consumption". *Proceedings of the 35th IARIW General Conference*, Copenhagen, Denmark, 20th - 25th August 2018. http://old.iariw.org/copenhagen/consolini.pdf.

Conti, P.L., D. Marella, and M. Scanu. 2012. "Uncertainty Analysis in Statistical Matching". *Journal of Official Statistics - JOS*, Volume 28, N. 1: 69-88.

de Waal, T. 2015. "Statistical matching: Experimental results and future research questions". Statistics Netherlands – CBS, *Discussion Paper* 2015/19.

Donatiello, G., M. D'Orazio, D. Frattarola, A. Rizzi, M. Scanu, and M. Spaziani. 2016*a*. "The role of the conditional independence assumption in statistically matching income and consumption". *Statistical Journal of the IAOS*, Volume 32, N. 4: 667-675.

Donatiello G., M. D'Orazio, D. Frattarola, A. Rizzi, M. Scanu, M. Spaziani. 2016*b*. "Statistical matching of EU-SILC and HBS at ISTAT: where do we stand for the production of official statistics". *Proceedings of the DGINS Conference of the Directors General of the National Statistical Institutes, Statistics on income, consumption and wealth*, Statistics Austria, Vienna, 26th – 27th September 2016.

Donatiello, G., M. D'Orazio, D. Frattarola, A. Rizzi, M. Scanu, and M. Spaziani. 2014*a*. "Statistical Matching of Income and Consumption Expenditures". *International Journal of Economic Sciences*, Volume III, Issue 3: 50–65.

Donatiello, G., M. D'Orazio, D. Frattarola, M. Scanu, and M. Spaziani. 2017. "Towards the production of integrated statistics on Income, Consumption and Wealth: pre-requisites and methodological challenges". *Proceedings of ITACOSM 2017, The 5th ITAlian Conference on Survey Methodology*, University of Bologna, Italy, 14th – 16th June 2017.

Donatiello, G., D. Frattarola, A. Rizzi, and M. Spaziani. 2015. "The Role of the Available Information in Statistical Matching It-SILC and HBS". *Proceedings of the Workshop on Best Practices for EU-SILC Revision*, Imperial College, London, UK, 16th–17th September 2015.

Donatiello, G., D. Frattarola, A. Rizzi, and M. Spaziani. 2014*b*. "Statistical Matching of IT-SILC and HBS: Some Critical Issues". *Proceedings of the Workshop on Best Practices for EU-SILC Revision*, Banco de Portugal, Lisbona, 15th – 17th October 2014.

D'Orazio, M. 2022. "StatMatch: Statistical Matching or Data Fusion". *R package version 1.4.1*. https://CRAN.R-project.org/package=StatMatch.

D'Orazio, M., M. Di Zio, and M. Scanu. 2006. *Statistical Matching: Theory and Practice*. Chichester, UK: John Wiley & Sons.

Eurostat. 2021. "Income, Consumption and Wealth". *Experimental statistics*. Luxembourg: Eurostat. https://ec.europa.eu/eurostat/web/experimental-statistics/income-consumption-and-wealth.

Istituto Nazionale di Statistica/Italian National Institute of Statistics – Istat. 2017. "Indagine sulle condizioni di vita (EU-SILC) - Dati Trasversali: File per la ricerca". *Microdati*. Roma, Italy: Istat. https://www.istat.it/it/archivio/212385.

Istituto Nazionale di Statistica/Italian National Institute of Statistics – Istat. 2016. "Indagine sulle spese delle famiglie: File per la ricerca". *Microdati*. Roma, Italy: Istat. https://www.istat.it/it/archivio/180341.

Organisation for Economic Co-operation and Development - OECD. 2013. *OECD Framework for Statistics on the Distribution of Income, Consumption and Wealth*. Paris, France: OECD Publishing.

Rässler, S. 2002. *Statistical Matching. A Frequentist Theory*, *Practical Applications and Alternative Bayesian Approaches*. Cham, Switzerland: Springer, Lecture Notes in Statistics.

Renssen, R.H. 1998. "Use of Statistical Matching Techniques in Calibration Estimation". *Survey Methodology*, Volume 24, N. 2: 171-183.

Singh, A.C., H.J. Mantel, M.D. Kinack, G. Rowe. 1993. "Statistical Matching: Use of Auxiliary Information as an Alternative to the Conditional Independence Assumption". *Survey Methodology,* Volume 19, N. 1: 59-79.

Stiglitz, J.E., A. Sen, and J.-P. Fitoussi. 2009. *Report by the Commission on the Measurement of Economic Performance and Social Progress*. Paris, France: French Government, CMEPSP.

Zhang, Li-C. 2015. "On Proxy Variables and Categorical Data Fusion". *Journal of Official Statistics - JOS*, Volume 31, N. 4: 783–807.

The *Rivista di statistica ufficiale* publishes peer-reviewed articles dealing with cross-cutting topics: the measurement and understanding of social, demographic, economic, territorial and environmental subjects; the development of information systems and indicators for decision support; the methodological, technological and institutional issues related to the production process of statistical information, relevant to achieve official statistics purposes.

The *Rivista di statistica ufficiale* aims at promoting synergies and exchanges between and among researchers, stakeholders, policy-makers and other users who refer to official and public statistics at different levels, in order to improve data quality and enhance trust.

The *Rivista di statistica ufficiale* was born in 1992 as a series of monographs titled "*Quaderni di Ricerca Istat*". In 1999 the series was entrusted to an external publisher, changed its name in "*Quaderni di Ricerca - Rivista di Statistica Ufficiale*" and started being published on a four-monthly basis. The current name was assumed from the Issue N. 1/2006, when the Italian National Institute of Statistics – Istat returned to be its publisher.

La Rivista di statistica ufficiale pubblica articoli, valutati da esperti, che trattano argomenti trasversali: la misurazione e la comprensione di temi sociali, demografici, economici, territoriali e ambientali; lo sviluppo di sistemi informativi e di indicatori per il supporto alle decisioni; le questioni metodologiche, tecnologiche e istituzionali relative al processo di produzione dell'informazione statistica, rilevanti per raggiungere gli obiettivi della statistica ufficiale.

La Rivista di statistica ufficiale promuove sinergie e scambi tra ricercatori, stakeholder, policy-maker e altri utenti che fanno riferimento alla statistica ufficiale e pubblica a diversi livelli, al fine di migliorare la qualità dei dati e aumentare la fiducia.

La Rivista di statistica ufficiale nasce nel 1992 come serie di monografie dal titolo "Quaderni di Ricerca Istat". Nel 1999 la collana viene affidata a un editore esterno, cambia nome in "Quaderni di Ricerca - Rivista di Statistica Ufficiale" e diventa quadrimestrale. Il nome attuale è stato scelto a partire dal numero 1/2006, quando l'Istituto Nazionale di Statistica - Istat è tornato a esserne l'editore.

ISSN 1828-1982

€ 15,00