rivista di statistica ufficiale

n. 2-3
2019

In this issue:

Istat
Istituto Nazionale di Statistica

rivista di statistica ufficiale

n. 2-3
2019

In this issue:

Istat

rivista di statistica ufficiale
n. 2-3/2019

ISSN 1828-1982

Istituto nazionale di statistica
Via Cesare Balbo, 16 – Roma

Editorial Preface

This double issue of the *Rivista di statistica ufficiale* publishes a selection of scientific articles all sharing the common trait of looking at methodologies for official statistics that are not based on traditional surveys, but make use of alternative sources of data either alone or combined with surveys.

The main results of the first two studies were also anticipated at the 2019 Edition of *ITACOSM - Survey and data science* (Firenze, Italy, 5th – 9th June 2019).

ITACOSM is an international conference organised biannually by the *Survey Sampling Group* of the Italian Statistical Society – SIS.

It aims at promoting the scientific discussion on the developments of theory and application of survey sampling methodologies in the fields of economics, social and demographic sciences, of official statistics and in the biological and environmental studies.

Charlotte Articus, Christopher Caratiola, Hanna Dieckmann, Max Gerhards, Ralf Münnich and Thomas Udelhoven are the authors of the first article dealing with the analysis of the potential of combining high-resolution remote sensing data and official data for small-scale estimation.

Using the example of the city of Cologne and considering data with a resolution of 100 metres or more as high-resolution data, an analysis of well-being is conducted at 100 metre grid cell and city district level.

Since the data come from different sources and have different scales, scaling techniques from the field of geostatistics are introduced as well as small-area estimation, investigating the methodological and technical challenges of combining methods from both disciplines.

This analysis focusses on methodological challenges mainly, in particular on how to deal with different scales. A Special emphasis is put on different upscaling and downscaling methods and their impact on the composite indicator of well-being for the city of Cologne.

Several uncertainty factors in the construction steps are also investigated and quantified using a sensitivity analysis.

The second article is written by Maria Filomeno and Irene Rocchetti with the purpose of illustrating the processing of a unique efficiency measure to be applied to the Italian Courts of first instance, using both a Data Envelopment Analysis and Beta Regression Models on its efficiency results.

This measure allows both a proper assessment and a comparative analysis among the different offices by considering existent indicators together with their organisational resources.

This study analyses also the efficiency outcomes obtained and their relation with the organisational capacity of the judicial offices, so as to verify the existence of a possible impact of the Superior Council of Judiciary (Consiglio Superiore della Magistratura – CSM) activities during the last years and to improve the organisational capacities of the judicial offices themselves.

The third scientific contribution, by Antonella Bernardini, Maria Cristina Casciano, Claudia De Vitiis, Alessio Guandalini, Francesca Inglese, Giovanni Seri, Marco Dionisio Terribili and Francesca Tiero, describes the work carried out by the statistical-methodological working group of the Italian National Institute of Statistics – Istat aimed at revising the sample design of the Consumer Price Survey taking into consideration the availability of scanner data as innovative data sources.

Scanner data represent the starting point for the implementation of the innovation in the Consumer Price Survey, enhancing and unburdening the data collection phase, together with the progressive introduction of more rigorous probabilistic sampling procedures for the selection of outlets and products (series).

The experiments of probabilistic selection schemes of the series are developed in two main phases that reflect Istat choice of making a gradual transition from a fixed to a dynamic approach in the calculation of the Consumer Price Index.

Marco Di Zio, Romina Filippini and Gaia Rocchetti close the present issue of the *Rivista di statistica ufficiale* describing the mass imputation procedure of the level of education applied to the Istat Base Register of Individuals.

This procedure integrates data of different nature: information deriving from administrative sources, the traditional population census data and the new permanent census survey.

It is a complex procedure composed by different steps depending on the information of the sources used. The imputation is based on log-linear models which allow greater flexibility in modelling associations, if compared to classical methods *e.g.* the hot-deck imputation.

This work analyses also the comparisons between the register estimates obtained with the imputation with those of the sample survey of the permanent census, in order to highlight advantages and limits of the proposed procedure.

Measuring well-being at local level using remote sensing and official statistics data

Charlotte Articus, Christopher Caratiola, Hanna Dieckmann, Max Gerhards, Ralf Münnich, Thomas Udelhoven [1]

Abstract

Measuring societal well-being as a multi-dimensional perspective on the conditions of people's life satisfaction has evolved to be an important task of European Official Statistics. Routinely, the focus of analysis is on country-comparisons. As central dimensions of well-being also vary on local level, we complement these insights by measuring well-being on the very low level of city districts and 100 metre grid cells. To achieve this, we combine high-resolution remote sensing data products with data from official statistics. As the data from different sources often have different scales, we discuss several scaling methods both from the field of geospatial research and from small area estimation. We calculate a composite well-being indicator on district and grid cell level for the city of Cologne and assess the influence of scaling methods and other construction decisions in a sensitivity analysis.

Keywords: composite indicator, kriging, small area estimation, sensitivity analysis.

1 Charlotte Articus (articus@uni-trier.de); Christopher Caratiola (caratiola@uni-trier.de); Hanna Dieckmann (dieckmann@uni-trier.de); Max Gerhards (gerhardsm@uni-trier.de); Ralf Münnich (muennich@uni-trier.de); Thomas Udelhoven (udelhove@uni-trier.de), Trier University, Germany. Economic and Social Statistics Department and Environmental Remote Sensing and Geoinformatics.

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.

1. Introduction and motivation

The concept of well-being is gaining in importance within the global indicator framework for the Sustainable Development Goals (SDGs). The United Nations member states adopted 17 integrated SDGs in 2015. Well-being is directly included in the third goal, which is to ensure healthy lives and promote well-being for all at all ages (United Nations, 2019). For a long time, well-being was measured as gross domestic product (GDP) per capita. Easterlin (1974) has induced a paradigm shift from GDP as a proxy for well-being to the concept of relative income and the incorporation of aspects beyond income by showing that higher income leads to a higher perception of well-being only up to a certain point. Other initiatives, such as the Stiglitz-Sen-Fitoussi Commission Report (2009) and the European Commission GDP and beyond communication (2009), have supported the development from GDP per capita as a measure for well-being towards multidimensional well-being measures. Eurostat (2019) defines 8+1 dimensions as an overarching framework for the measurement of well-being:

- Material living conditions;
- Productive or main activity;
- Health;
- Education;
- Leisure and social interactions;
- Economic and physical safety;
- Governance and basic rights;
- Natural and living environment;
- Overall experience of life.

Stiglitz *et al.* (2009) identify material living standards, personal insecurity, social connections and relationships, environmental conditions and political voice and governance as main dimensions in the Stiglitz-Sen-Fitoussi report. The OECD (2020) defines human well-being in terms of eleven dimensions under the themes of material conditions and quality of life. All definitions comprise both individual and place-related factors. Thus, well-being highly depends on the living environment and differs not only from country to country, but rather is affected by local living conditions.

The traditional initiatives to measure well-being have focussed on international comparisons of country-level indicators (Eurostat, 2020; OECD, 2020). More recently, this perspective has been complemented by regional studies, both in research and official statistics. For example, some National Statistical Institutes have started to report well-being at a local level (see Office for National Statistics, 2019 and Istituto Nazionale di Statistica – Istat, 2019). Eurostat, National Statistical Institutes and the European Commission cooperate in a voluntary data collection exercise to build a database for measuring life quality in European cities (a project previously known as Urban Audit; see Eurostat, 2017). Eurofound (2020), the European Agency for the improvement of living and working conditions, recently examined the quality of life in European capitals in comparison to the rest of the country based on its own European Quality of Life Surveys (EQLS). Moretti *et al.* (2019) employ small area estimation techniques to estimate composite well-being indicators on a regional level. More detailed, they employ factor analysis to reduce the dimensionality of complex indicator systems and integrate this approach into multivariate small area statistics to gain estimates at municipality level.

Generally, the availability of data is a challenge for each initiative to measure well-being at a regional level. Data commonly used to assess the quality of life come from survey data such as the European Union Statistics on Income and Living Conditions (EU-SILC). This data cannot reliably be evaluated at a local level. Small area estimation techniques, as employed by Moretti *et al.* (2019), are a possible solution. Additionally, a feasible strategy is to exploit further data sources such as administrative data from local authorities (see *e.g.* Istituto Nazionale di Statistica, 2019). Integrating administrative and other data, however, may suffer from different degrees of granularity. Shuvo Bakar *et al.* (2020) provide a Bayesian approach to model predictions that help compensating overlapping geographical areas. Alternative methods from geostatistics are known as spatial resampling methods. Against this backdrop, we explore the opportunity of combining high-resolution remote sensing data[2] with official data as an efficient strategy to conduct analyses of well-being at local level.

2 In the field of remote sensing, high resolution data is used to describe raw images. In this paper, we use the term to describe products derived from satellite images or other georeferenced sources. These derived products are referred to as remote sensing data products in the field of remote sensing.

Since the late 1950s, geospatial data have been incorporated in the analysis of urban sociology. In his pioneering work, Green (1957) relates aerial photographic interpretation information with socio-economic data of Birmingham, Alabama and finds that photographic interpretation information can supplement and substitute other socio-economic data sources. Satellite data have since been used to estimate well-being at local level. Lo and Faber (1997) complement census data with satellite data and assess the quality of life in the Athens-Clarke County of Georgia with an environmental perspective. Ghosh *et al.* (2013) evaluate well-being using night-time light data and Engstrom *et al.* (2017) estimate economic well-being by extracting object and texture features from satellite images of Sri Lanka.

In this article, we analyse the potential of combining high-resolution remote sensing data and official data for small-scale estimation (*e.g.* block and district level) using the example of well-being in the city of Cologne. We consider data with a resolution of 100 metres or more as high-resolution data. The analysis of well-being is conducted at 100 metre grid cell and city district level. Since the data come from different sources and have different scales, we introduce scaling techniques from the field of geostatistics as well as small-area estimation and investigate methodological (*e.g.* different scales) and technical (*e.g.* confidentiality requirements) challenges of combining methods from both disciplines. The main focus is on methodological challenges, especially on how to deal with different scales. Special emphasis is put on different upscaling and downscaling methods and their impact on the composite indicator of well-being for the city of Cologne. By means of a sensitivity analysis, various uncertainty factors in the construction steps are investigated and quantified.

2. Data

This article combines INSPIRE conform[3] Census 2011 grid cell data at 100 metre resolution, OpenStreetMap data and Pan-European High-Resolution Layers (HRL). The impact of scaling methods is investigated using the example of the composite indicator of well-being for the city of Cologne, since the City of Cologne (2017, 2014) provides georeferenced and socio-demographic data.

The composite indicator comprises socio-demographic and place-related data. Socio-demographic data are obtained from the Federal Statistical Office and the statistical offices of the Länder (2018) and the City of Cologne (2017). Data on single parents are available at 100 metre grid cell level from disaggregated census statistics. The 100 metre grid cell data do not contain values smaller than three due to disclosure control. In this study, empty cells are treated as zero, resulting in deviations of 2 percent (%) at city level for single parents. The City of Cologne (2017) offers a variety of statistical data at district and municipality level, including information on unemployment with time reference to December 2017 and single parents as of 31 December 2017.

Georeferenced data on schools, museums, play- and sports grounds, hospitals and libraries are published by the City of Cologne (2014) in their open data portal. These data together with OpenStreetMap data are used to conduct network analyses using the QNEAT3 (distance matrices) QGIS plugin (see Figure 1). Residential buildings are taken from OpenStreetMap and include all houses and apartments which are tagged as residential. In order to include distances from each address to primary schools, museums, hospitals, libraries and play- and sports grounds in the analysis of well-being, network analyses are conducted based on the shortest distance using streets tagged as highway[4] from OpenStreetMap (OpenStreetMap contributors, 2019). The OpenStreetMap data are taken from the QuickOSM Plugin in QGIS. Figure 1 shows the distance to primary schools in 250 metre intervals for a part of Cologne[5].

3 INSPIRE conformity means that spatial data are harmonised across Europe and comply with international geomatics standards (European Commission, 2019).

4 Highways include any type of road, street or path.

5 The distances are determined using the QNEAT3 (Iso-Areas) QGIS plugin, which only accepts a projected co-ordinate system. Therefore, the street network is taken from Geofabrik GmbH and OpenStreetMap contributors (2018).

Figure 1 - Distances to primary schools for a part of Cologne

Source: Own illustration based on data from OpenStreetMap contributors (2019), Geofabrik GmbH and OpenStreetMap contributors (2018) and the City of Cologne (2014)

Remote sensing data are also regarded as place-related data. HRL are obtained from satellite imagery by applying automatic processing and interactive rule based classification. Currently, HRL provide information on tree cover density and forest types, grasslands, wetness and water, small woody features and imperviousness (Copernicus, 2019a). In this study we integrate imperviousness data into the analysis of well-being at local level. The imperviousness product gives the percentage of impervious surfaces. Imperviousness data is available in the original 20 metre and 100 metre pixel size for the years 2006, 2009, 2012 and 2015 (Copernicus, 2019b). Furthermore, we consider data on vacant dwellings as place-related data in the broader sense. Results on vacant dwellings[6] at 100 metre resolution are available from Federal Statistical Office and the statistical offices of the

6 The Census 2011 defines an apartment as vacant if it is neither rented out nor used by the owner on the date of the survey and if it is not a holiday and leisure apartment, diplomatic apartment, apartment of foreign armed forces and commercially used apartment.

Länder (2018). The treatment of empty cells as zero leads to a deviation of 11.5 % at city level. Census 2011 results for the city of Cologne are published by Federal Statistical Office and the statistical offices of the Länder (2018). In this article, we analyse well-being at 100 metre grid cell and city district level.

3. Theoretical framework for the construction of a composite indicator for well-being

3.1 Definition of composite indicators

Composite indicators are used to aggregate indicator information to a lower dimension. We refer to Münnich and Seger (2014) for a formal derivation. In the following, we restrict ourselves to a linearly weighted aggregation. Well-being at local level is assessed by a composite indicator, comprising $q = 1, ..., Q$ sub-indicators I, which is calculated as

$$CI_d = \sum_{q=1}^{Q} w_q \cdot I_{qd}, \tag{1}$$

with w denoting the weights and d the area of interest. Our composite indicator at district level includes twelve sub-indicators: (1) single parent households (%), (2) unemployment (%), (3) youth unemployment (%), (4) vacant dwellings (%), the average distance to (5) primary schools, (6) libraries, (7) museums, (8) play- and sports grounds and (9) hospitals, respectively, (10) parks, green areas and sport fields (%), (11) forest areas (%) and (12) water areas (%). At grid cell level the composite indicator comprises 10 sub-indicators. Sub-indicators (1) to (9) are the same as at district level but as absolute numbers and sub-indicators (10) to (12) are summarised as natural areas approximated by the mirror image of impervious surfaces[7].

The indicator is constructed using spatial and official data. In our case, unemployment data are only available at city district level and information about the number of vacant dwellings at 100 metre grid cell level. Both unemployment and vacancy data have to be re-scaled in order to analyse well-being at grid cell and district level, respectively. Therefore, methods for changing the scale are required. In addition, the construction of composite indicators requires normalisation and weighting of sub-indicators. In the following, different scaling methods from spatial research and small area estimation are presented in brief, following Rao and Molina (2015) and Zhang *et al.* (2014). Moreover, various normalisation and weighting methods

7 The imperviousness raster data are summarised within the 100 metre census grid cells using the zonal toolset from the spatial analysis toolbox of ArcGIS Pro.

according to OECD *et al.* (2008) are introduced. In particular, possible sources of uncertainty, which arise from scaling methods, selection of sub-indicators, data normalisation and weighting choices are considered. These sources of uncertainty are regarded as construction steps of composite indicators with each step offering several selection choices, also called triggers. Each possible combination results in a different composite indicator (see equation (1)).

3.2 Scaling of data

As described above, the data come from different data sources. The disaggregated census data on vacant dwellings and single parents are available at 100 metre grid cell level, HLR imperviousness data at 20 metre resolution and socio-demographic data at city district level. In addition, we have georeferenced data on schools, museums, play- and sports grounds, hospitals and libraries. In order to conduct analyses at 100 metre grid cell level and city district level, the scales have to be harmonised using up- or downscaling techniques. In the following, several up- and downscaling methods from the fields of geostatistics and small area estimation are introduced.

3.2.1 Upscaling methods

Upscaling refers to the aggregation of fine-resolution input data to coarse-resolution output data. Information on vacant dwellings is only available at 100 metre grid cell level. In order to conduct an analysis of well-being at city district level, this information needs to be upscaled. The selection of the method is determined by characteristics of the input data. In the following, different upscaling methods are introduced.

Upscaling methods include aggregation using the (weighted mean), block-kriging, the methods of random selection, median rule, mid-point rule, majority rule and reclassification of coarsened images are needed (Yang and Merchant, 1997; Zhang *et al.*, 2014, pp. 219ff.). Random selection assigns randomly a value from the fine-resolution grid to the aggregated coarse-resolution cell. The data at 100 metre census grid cells level are interpreted as point data, which are randomly assigned to represent the district. The selection probability of a fine-resolution grid cell to represent the aggregated coarse-resolution cell is proportional to its occurrence. Thus, the method of random selection is more

likely to preserve the structure than the majority rule, widely which is used. The majority rule chooses the most frequent value from the fine-resolution grid cells within the coarse-resolution area and qualifies the coarse-resolution output accordingly. If two or more classes occur with the same frequency, it is drawn randomly. The median rule attaches the (weighted) median value of the fine-resolution grids within the coarse-resolution area as aggregated value to the district. The degradation-reclassification approach applies an averaging degradation process and reclassifies the resulting images to obtain a coarse-resolution image with the same characteristics as the input images except for the resolution (Yang and Merchant, 1997; Zhang *et al.*, 2014, pp. 219ff.).

In this study the average vacancy rate is determined by calculating the weighted mean and applying block-kriging. In the first method, the vacancy rate is calculated by weighting the number of vacant dwelling and total dwelling in grid cells which are intersected by district boundaries with the high-resolution impervious data. Assuming that the number of dwellings correlates with the impervious surface and is evenly distributed, the dwellings are distributed proportionally to the impervious surface at 100 metre grid cell. The second method is based on Zhang *et al.* (2014, pp. 109ff.) and Zhang and Yao (2008). Block-kriging (or point-to-area kriging) utilises additional information, *e.g.* spatial dependence in the underlying distribution to estimate the mean value of a variable Z for a predefined large area, *e.g.* districts. Block kriging assumes that the mean value of a random variable over a block v_x centred at location x is defined as the average of all n_p random variable points $Z(x_\beta)$ which discretise the block

$$Z(v_x) = \frac{1}{n_p} \sum_{\beta=1}^{n_p} Z(x_\beta).$$

The simplest form of kriging derives kriging weights based on the criteria of unbiasedness and minimum variance of the estimator. Stationarity of the mean and covariance of the problem domain is assumed. The estimator for Z over a block v_x is a linear combination

$$\hat{z}(v_x) = m_Z + \sum_{\beta=1}^{n_p} \lambda_\beta [z(x_\beta) - m_Z] \tag{2}$$

holds, with m_Z being the known stationary mean and λ_β denoting kriging weights (see Zhang and Yao, 2008). The average prediction error $\hat{z}(v_x) - z(v_x)$ is set to be zero

$$E[\hat{z}(v_x) - z(v_x)] = m_Z + \sum_{\beta=1}^{n_p} \lambda_\beta \{E[z(x_\beta)] - m_Z\} - E[z(x_\beta)] = 0.$$

The variance of prediction is given as

$$\begin{aligned}\sigma_Z^2(v_x) &= \text{var}\left[\hat{z}(v_x) - z(v_x)\right] \\ &= \text{var}\left\{\sum_{\beta=1}^{n_p} \lambda_\beta [z(x_\beta) - m_Z] - [z(v_x) - m_Z]\right\}.\end{aligned}$$

Defining

$$\boldsymbol{\lambda^*} = \begin{bmatrix} \boldsymbol{\lambda} \\ -1 \end{bmatrix} = \begin{bmatrix} \lambda_1 \\ \vdots \\ \lambda_{n_p} \\ -1 \end{bmatrix}$$

$$\boldsymbol{Z^*} = \begin{bmatrix} \boldsymbol{Z}(\boldsymbol{x}_\beta) - \boldsymbol{m_Z} \\ z(v_x) - m_Z, \end{bmatrix} = \begin{bmatrix} z(x_1) - m_Z \\ \vdots \\ z(x_{n_p}) - m_Z \\ z(v_x) - m_Z \end{bmatrix},$$

the reduced form can be written as

$$\begin{aligned}\text{var}[\boldsymbol{\lambda^{*T}} \boldsymbol{Z^*}] &= \begin{bmatrix} \boldsymbol{\lambda^T} & -1 \end{bmatrix} \begin{bmatrix} \mathbf{cov}(\boldsymbol{x}_\beta) & \mathbf{cov}(\boldsymbol{x}_\beta, v_x) \\ \mathbf{cov}(\boldsymbol{x}_\beta, v_x)^T & \text{var}[z(v_x)] \end{bmatrix} \begin{bmatrix} \boldsymbol{\lambda} \\ -1 \end{bmatrix} \\ &= \boldsymbol{\lambda}^T \mathbf{cov}(\boldsymbol{x}_\beta) \boldsymbol{\lambda} - 2\boldsymbol{\lambda}^T \mathbf{cov}(\boldsymbol{x}_\beta, v_x) + \text{var}[z(v_x)]\end{aligned}$$

with $\mathbf{cov}(\boldsymbol{X}_\beta, v_x)$ being the point-to-block covariance vector. The weight vector

$$\boldsymbol{\lambda} = \mathbf{cov}(\boldsymbol{x}_\beta)^{-1} \mathbf{cov}(\boldsymbol{x}_\beta, v_x)$$

minimises the prediction variance and solves equation (2). Thus, the block kriging estimator is given as

$$\begin{aligned}\hat{z}(v_x) &= m_Z + \sum_{\beta=1}^{n_p} \lambda_\beta (z(x_\beta) - m_Z) \\ &= m_Z + \boldsymbol{\lambda}^T (\boldsymbol{Z}(\boldsymbol{x}_\beta) - m_Z)\end{aligned}$$

and the kriging variance as

$$\boldsymbol{\sigma}_Z^2(x) = \text{var}[z(v_x)] - \boldsymbol{\lambda}^T \mathbf{cov}(\boldsymbol{x}_\beta, v_x),$$

where var[$z(v_x)$] denotes the average covariance within the block being predicted (Zhang and Yao, 2008, p. 183; Zhang *et al.*, 2014, pp. 109ff.).

3.2.2 Downscaling methods

Downscaling is the inverse of upscaling and refers to the conversion of coarse-resolution data to fine-resolution data. In our case, data on unemployment and youth unemployment are available at district level. In order to conduct an analysis of well-being at 100 metre grid cell level, this information needs to be downscaled. Several downscaling methods from the small area literature and geospatial research are available, which are presented briefly in the following.

The method of area-to-point kriging is in contrast to block kriging as explained above. Assuming that the grid cell is represented by its centroid, point values for each grid cell are estimated based on available data at district level. For the derivation of area-to-point kriging, we refer to Zhang *et al.* (2014, pp. 120ff.) and Kyriakidis (2004). Area-to-point kriging is not used in this study as the quality of the results at 100 metre grid cell level is distorted as very coarse information is converted to very fine information.

Further examples of downscaling methods include geographic centroid assignment, areal weighting, dasymetric mapping and regression methods. Geographic centroid assignment assigns a representative value for the district to its centroid. The values for the large area are then assigned to the 100 metre grid cell centroids using the distances between the large area centroid and small area centroids as weights. Grid cells that are uninhabited are not included in the analysis. Whether a grid cell is uninhabited is determined on the basis of the number of houses classified by OpenStreetMaps. Areal weighting approaches are based on cartographic techniques. Simple area-weighted interpolation determines weights based on the percentage of the overlapping large area and small area assuming that the socio-demographic variable of interest (*e.g.* population, unemployment) is evenly distributed within the large area. Thus, areal weighting methods rely on the assumption of homogeneity within each large area (see *e.g.* Goodchild and Lam, 1980). Dasymetric mapping uses auxiliary data such as remote sensing data (*e.g.* high-resolution imperviousness data). High-resolution imperviousness data

are used to better depict the distribution of the socio-economic variable of interest in the large area and, thus, accounts for the fact that some parts of the area of interest might not be populated assuming that impervious surfaces approximate population and unemployment is evenly distributed across the population (see *e.g.* Eicher and Brewer, 2001). The downscaling methods described above rely on interpolation. Alternatively, the conversion of coarse-resolution data to fine-resolution data can be achieved by regression methods, which establish a relationship between different scales (see *e.g.* Fernandes *et al.*, 2004; Martinez *et al.*, 2009; Wu and Li, 2009).

Alternatively to these methods from the geoscientific research, simple approaches from the discipline of small area estimation can be applied as downscaling techniques. Small area estimation (SAE) generally deals with situations in which survey data is to be evaluated on a highly disaggregated level. In this case, the sample size in some or many areas is typically so small, that traditional direct estimators that only rely on the sample data in a specific area lack accuracy. The strategy then is to use indirect estimation methods, that borrow information from other areas to stabilise estimation.

There is a broad range of different approaches to small area estimation. Which method is suitable, crucially depends on the availability of data, both with respect to the variable of interest and auxiliary information, and the type of the target information. An introduction into the field and a comprehensive overview can be found in the monograph of Rao and Molina (2015). Recent developments are also presented in Pfeffermann (2013). An introductory overview in German language is given by Münnich *et al.* (2013). We focus on approaches that might be of relevance in the context of measuring well-being on a low aggregation level. We first look at some SAE approaches with minimal data requirements on the target resolution level and then present some relevant methods from the broader range of SAE techniques, that opens up if additional information is available. We finally focus on the SPREE-estimator and related extensions because this might prove to be a relevant approach in the context of well-being indicators, which in many cases rely on categorical variables.

Before presenting the selected small area methods, note that small area estimation as a part of survey statistics generally deals with data that was obtained in a random sample in order to obtain reliable statistics for a

larger population from which the sample was taken. Most procedures and expressions presented below can however also be applied in cases where data for the entire target population, *i.e.* register data, is available for a larger area, for example obtained from administrative sources. In these cases the synthetic approaches (while not being *estimators* in the classical sense then) can still be suitable ways to deduce statistics on the level of smaller areas (see *e.g.* Rao and Molina, 2015, Chapter 3.2.3).

If no sample information for the variable of interest is available at the targeted fine-resolution level, the range of feasible approaches is largely restricted and only some very simple synthetic estimators can be considered. Generally, a synthetic estimator uses a reliable estimator for a larger area to derive an indirect estimator for smaller areas within this larger area, relying on the assumption that the small areas share the characteristics of the larger area (Rao and Molina, 2015). If no area information is available, a very simple naive synthetic estimator for the mean (or proportion) $\hat{\overline{Y}}_d$ in a small area d is given by

$$\hat{\overline{Y}}_d^{syn} = \hat{\overline{Y}} \qquad d = 1, ..., D,$$

where $\hat{\overline{Y}}$ is the direct estimator of the large area. This naive synthetic estimator relies on the implicit assumption that the small area means are equal to the large area mean. Obviously, it is highly inadequate when this strong assumption is inappropriate. If a suitable auxiliary variable x is available, the ratio-synthetic estimator for the domain total Y_d can be obtained as:

$$\hat{Y}_d^{rs} = X_d \frac{\hat{Y}}{\hat{X}} \qquad d = 1, ..., D,$$

where X_d is the known area-level total and $\hat{X}$ is the population or larger-area total estimated from the same sample as $\hat{\overline{Y}}$. This estimator relies on the assumption that the rate $R_d = Y_d/X_d$ is approximately equal to the overall ratio $R = Y/X$ and the bias might be large if this assumption is not fulfilled (see Rao and Molina, 2015, Section 3.2). We use the ratio-synthetic estimator to obtain unemployment numbers at 100 metre grid cell. The known area-level totals X_d, here inhabitants, come from the 100 metre census grid cells and Y and X, inhabitants and unemployed people on the district level, from the City of Cologne (2017). It has to be noted that some grid cells can not be allocated

to one district only. In some cases, the 100 metre grid cells are intersected by district borders. We assign the grid cells to the district with the largest intersection.

So far, we presented methods with minimal data requirement on the targeted fine-resolution level. Typically in SAE problems, a sub-sample (albeit small) is available in at least most of the areas. The best-known and regularly applied methods use this information. Synthetic approaches that use sample information on the target level are the regression synthetic estimator and the GREG-synthetic estimator (Rao and Molina, 2015, Section 3.2). More importantly, the most common small area models, which as special cases of a General Linear Mixed Model employ an explicit statistical model to obtain small area estimates, become feasible. A large part of Rao and Molina (2015) is dedicated to these models.

A special problem of SAE is that of estimating cell counts (or proportions) of a categorical variable. Assume that the counts are arranged in a two-way table, where each row contains a vector of frequencies for the p categories of the variable of interest in a given area. Following Hernandez, we call this arrangement a population composition and denote it by $\boldsymbol{Y}$ (Hernandez, 2016). Assume that a sample of the target population is available that – while yielding reliable estimates for the margins of the compositions – is too small to obtain accurate estimates of cell frequencies. Further, some proxy composition $\boldsymbol{X}$ is available. This can, for example, be the result from a previous census that needs updating.

Generally, structure preserving estimators provide estimates of the cell frequencies by adjusting them to the known margins while at the same time in some way preserving the association structure, *i.e.* the relationship between rows and columns, observed in the proxy composition $\boldsymbol{X}$. In this adjustment, several assumptions on the relationship between the association structure in $\boldsymbol{X}$ and $\boldsymbol{Y}$ can be used. For estimation, typically, the method of iterative proportional fitting (IPF) (Deming and Stephan, 1940) is employed. Note that calibration to known margins in these approaches is an inherent feature of the estimation process.

The basic structure preserving estimator (SPREE) was introduced by Purcell and Kish (1980). It makes the simple assumption that the association structure of $\boldsymbol{X}$ and $\boldsymbol{Y}$ is equal. Let $Y_{d,a}$ denote the count for area d, $d = 1, \ldots, D$

and category *a, a* = 1, . . . , *p*. Further, we use Y_{d+} and Y_{+a} to denote the known row- and column-margins, respectively. Assume that a proxy composition $\boldsymbol{X}$ with the same dimensions as $\boldsymbol{Y}$ is available. The aim is, to obtain estimates $\hat{Y}_{d,a}$ that minimise the distance between the cell counts and fitted values under constraints implied by the margins. As this optimisation problem cannot be solved in closed form, an estimate for $\boldsymbol{Y}$ is obtained iteratively by IPF, applying the following procedure (see Hernandez, 2016; Agresti, 2013, Chapter 9.7.2):

Step 1: $$\hat{Y}_{d,a}^{(1)} = X_{d,a} \frac{Y_{+a}}{X_{+a}} \tag{3}$$

Step 2: $$\hat{Y}_{d,a}^{(2)} = \hat{Y}_{d,a}^{(1)} \frac{Y_{d+}}{\hat{Y}_{d+}^{(1)}} \tag{4}$$

Step 3: $$\hat{Y}_{d,a}^{(3)} = \hat{Y}_{d,a}^{(2)} \frac{Y_{+a}}{\hat{Y}_{+a}^{(2)}} \tag{5}$$

Step 2 and 3 are repeated until convergence. The algorithm converges to the optimal solution, *i.e.* it minimises the distance between cell counts and fitted values according to the Kulback-Leibler discrimination information measure $\sum_d \sum_a Y_{ia} \log \frac{Y_{d,a}}{\hat{Y}_{d,a}}$ (Ireland and Kullback, 1968).

A closely related approach is presented by Dostal *et al.* (2016), who suggest an extension of the SPREE with an alternative distance function. This has proven to be a suitable approach in the case of very small domains (Dostal *et al.*, 2016). If sample estimates for the inner cells are available, more elaborate methods become feasible. Zhang and Chambers (2004) propose a Generalised SPREE (GSPREE), that assumes a proportional relationship between the association structure of the target compositions and its proxy. They further present a version of this approach that allows for cell-specific random effects. Hernandez (2016) presents a Multivariate SPREE (MSPREE), an extension of the GSPREE that allows for further flexibility regarding the structural assumption.

Due to restrictions in data availability on the very fine resolution level of districts and grid cells, the more elaborate approaches presented here, cannot be applied in the study at hand. As they open up the opportunity to account for the socio-demographic structure in downscaling, we think that it is worth to pursue them further. This will require close cooperation with official statistics.

3.3 Construction steps of composite indicators

The data need to be normalised prior to any aggregation. Different methods, such as normalised ranking, standardisation or min-max methods, are available. Normalised ranking is the simplest normalisation method. It evaluates the performance of the area of interest in the subsequent dimension according to its relative position

$$I_{qd} = \frac{\text{Rank}(x_{qd})}{D},$$

where x_{qd} is the value of sub-indicator q of area d and D the total number of areas.

Standardisation ensures that the sub-indicators have zero mean and a standard deviation equal to one, *i.e.*

$$I_{qd} = \frac{x_{qd} - x_{qd=\bar{d}}}{\sigma_{qd=\bar{d}}},$$

where $x_{qd=\bar{d}}$ and $\sigma_{qd=\bar{d}}$ denote the average and standard deviation across countries, respectively. Sometimes the composite indicator is adjusted, *e.g.* by weights, as outliers distort the composite indicator.

The min-max methods subtracts the minimum value and divides the difference by the range of the sub-indicator values

$$I_{qd} = \frac{x_{qd} - \min_d(x_q)}{\max_d(x_q) - \min_d(x_q)}$$

and, thus, sub-indicators range between zero and one (OECD *et al.*, 2008).

The relative importance of the sub-indicators is determined by the attached weights. The most frequently used weighting technique is equal weighting. It assigns the same weight to all sub-indicators, implying that all sub-indicators are equally important. However, due to potential correlation between sub-indicators, equal weighting does not guarantee an equal contribution of the sub-indicators to the composite indicator. Alternatively, weights based on statistical models such as the principal component analysis (PCA) or on expert opinions can be applied. The goal of PCA is to determine how different variables change in relation to each other and how these variables are associated. Correlated variables are converted into a new set of uncorrelated

variables using a covariance matrix or correlation matrix. PCA involves finding the eigenvalues of this covariance matrix (OECD *et al.*, 2008; Nardo *et al.*, 2005). Weights based on PCA are constructed following Nicoletti *et al.* (2000). The indicators with the highest squared factor loading are grouped into intermediate composites with the squared factor loadings summed to unity as weight. These intermediate indicators are aggregated by assigning the proportion of explained variance as weights to the intermediate composites. An exemplary application of the construction of weights based on PCA can be found in OECD *et al.* (2008, p. 90f.). The PCA is conducted applying the prcomp function in R.

4. Implementation of the sensitivity analysis

4.1 Theoretical framework for sensitivity analyses

In order to assess the robustness of the resulting composite indicator with respect to the scaling schemes, the normalization method and the choice of weights, a sensitivity analysis is conducted. Sensitivity analyses aim at determining the effect of a change in the input factors on the variable of interest. In this article, the output variation in the composite indicator is caused by the choices in the construction steps. The sensitivity analysis is based on a variance decomposition method as described by Saltelli *et al.* (2000, 2008) as an extension of the original approach proposed by Sobol (1993) and Homma and Saltelli (1996).

The composite indicator for well-being is calculated according to equation (1). The different construction steps (triggers), *i.e.* scaling techniques, normalisation methods and the choice of weights, have a direct impact on the output of the composite indicator. Variance decomposition within the scope of sensitivity analyses indicates how much each construction step contributes to the total variance in the result. Following Saltelli *et al.* (2008), the composite indicator *CI* is understood as a function of the *k* uncertain construction alternatives *T*:

$$CI = f(T_1, ..., T_k).$$

For mutually independent input factors *T*, the total or unconditional variance of the output, *V (CI)*, can be decomposed into

$$V(CI) = \sum_i V_i + \sum_i \sum_{j>i} V_{ij} + ... + V_{12...k},$$

where

$$V_i = V(f_i(T_i)) = V[E(CI|T_i)]$$

and

$$V_{ij} = V(f_{ij}(T_i, T_j)) = V[E(Y|T_i, T_j)] - V_i - V_j.$$

V_i is the first-order or main effect of T_i on *CI*, i.e. the individual contribution of T_i to the variance of the output. The second-order effect V_{ij} represents the joint effect of T_i and T_j. Setting these effects in relation to the total variance, a first-order sensitivity index of T_i on *CI* is obtained:

$$S_i = \frac{V[E(CI|T_i)]}{V(CI)}.$$

S_i is in the interval [0, 1] with values close to 1 indicating input factors with a large effect on the output. Correspondingly, higher order sensitivity indices, *e.g.* the second order effect S_{ij}, are derived by setting the higher order effects in relation to the total variance. Additionally, one might be interested in the total contribution of a specific input factor T_i to the output, *i.e.* the sum of its main effect and all relevant higher-order effects. For T_1 and three insecure input factors a corresponding total sensitivity index is for example given by (Münnich and Seger, 2014; Saltelli *et al.*, 2008)

$$S_{tot_1} = S_1 + S_{1,2} + S_{1,3} + S_{1,2,3}.$$

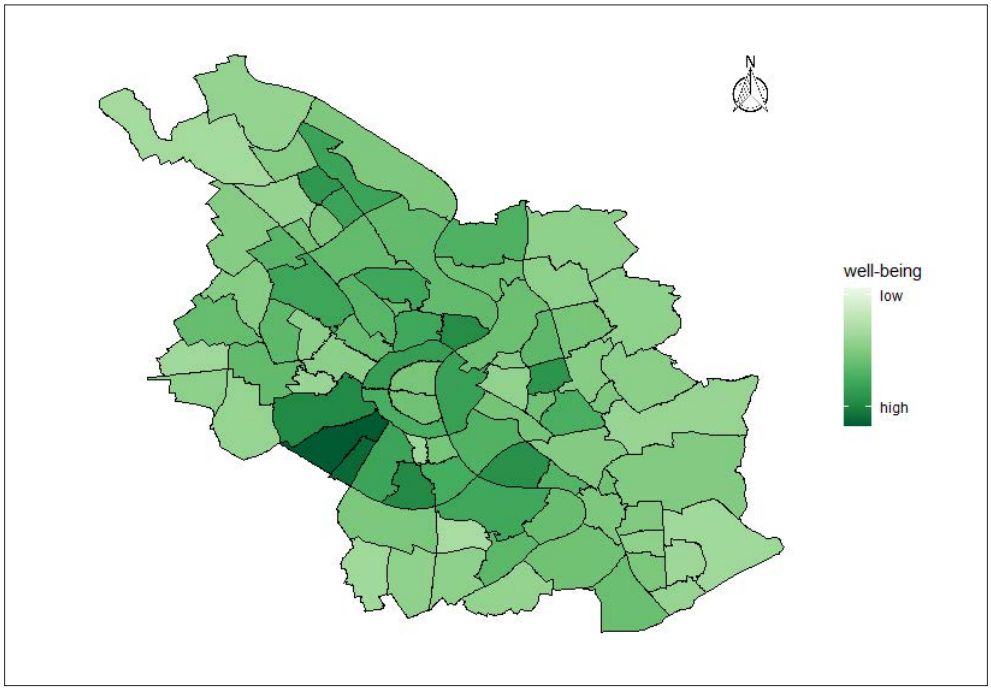
For detailed explanations we refer to Saltelli *et al.* (2008, pp. 20-21 and 155-174). The sensitivity analysis is performed using the R-package multisensi (see Bidot *et al.*, 2018).

4.2 Composite indicator of well-being at district level

As explained above, the construction of composite indicators comprises scaling of the data, selection and normalisation of sub-indicators and the choice of weights. In the sensitivity analysis, we consider two upscaling methods (weighted mean, block-kriging), three normalisation schemes (min-max method, ranking and standardisation), two weighting possibilities (based on PCA and equal weighting) and exclusion of one indicator. Each of the twelve indicators considered at district level is left out once and further all indicators are taken into account, resulting in 13 exclusion options. This results in 156 possible combina- tions per district.

Figure 2 illustrates the composite well-being indicator for the City of Cologne at district level based on all twelve indicators and using the weighted mean as upscaling method, the min-max method for normalisation and weights resulting from PCA.

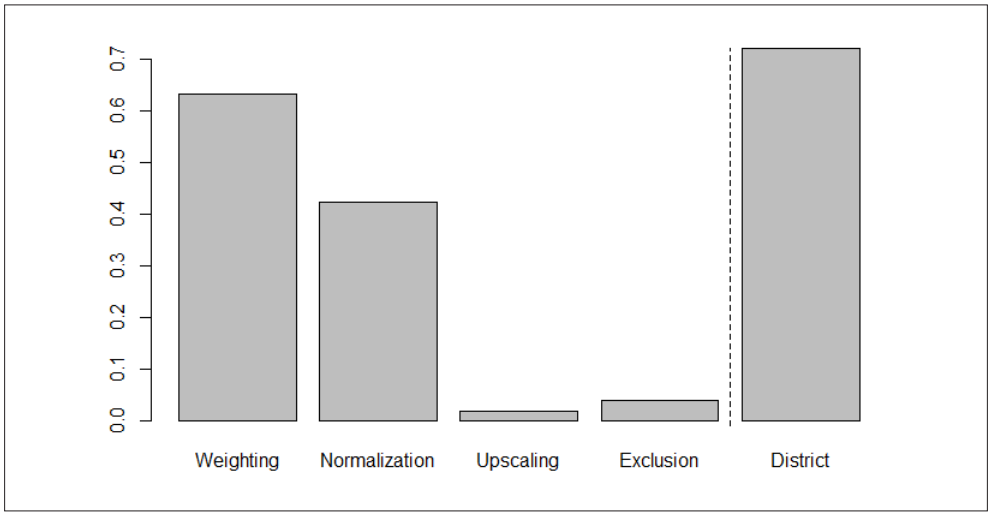
Figure 2 - Well-being indicator values at district level



Source: Own illustration

The uncertainty in the construction of the composite indicator for well-being is assessed by a sensitivity analysis. Sensitivity analyses study how much each source of uncertainty contributes to the output variance. Figure 3 presents the total-order effect with scaling, selection of sub-indicators, normalisation and weighting as input factors. To set these results into relation, we also depicted the variation of indicator results between districts as a reference.

Figure 3 - Total-order effects of the sensitivity analysis at district level

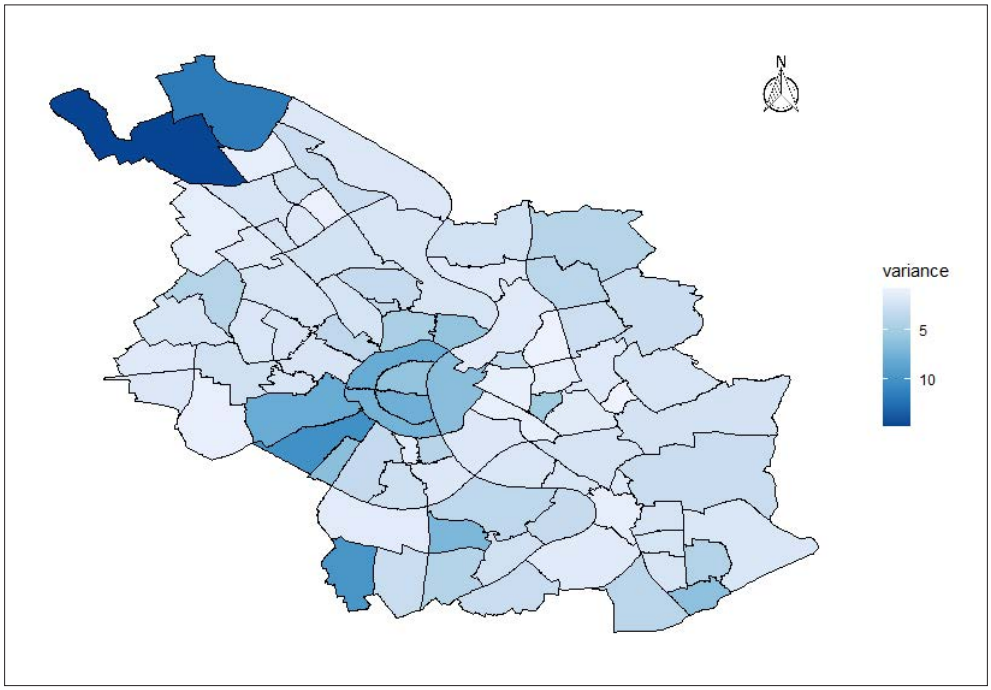


Source: Own illustration

From Figure 3 we observe that the scaling methods have the lowest impact on the output variance and that the construction decision with the largest effect is the choice of the weighting method. It can also be taken from this plot that most of the variability in results is still due to district identity, showing that there is an actual regional heterogeneity of well-being which is not offset by construction decisions.

The variance of the composite indicator depending on the construction is depicted in Figure 4.

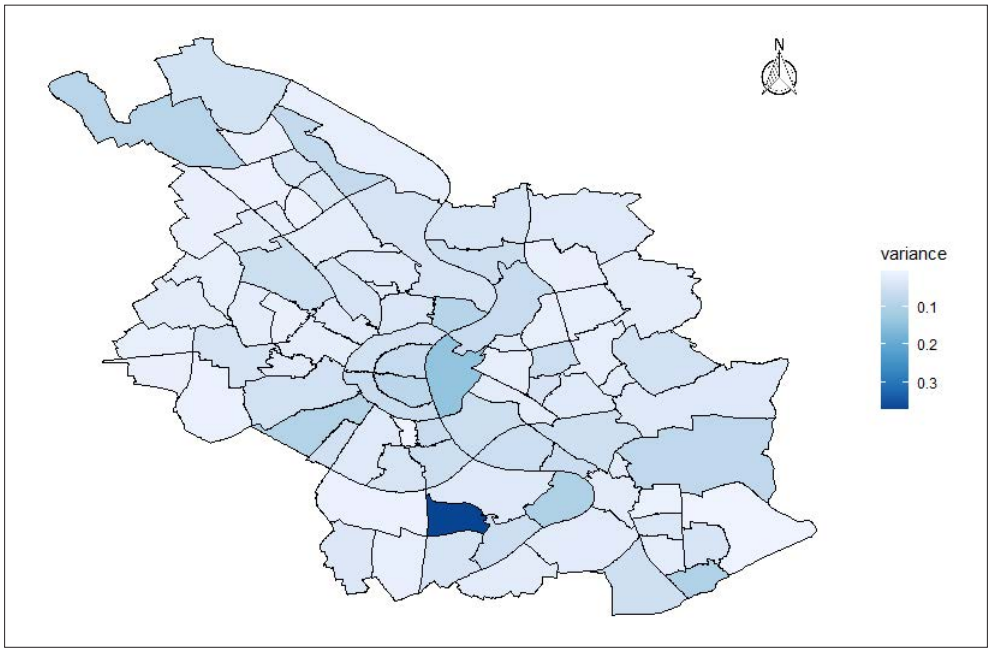
Figure 4 - Conditional variance of the composite indicators at district level



Source: Own illustration

The results of the sensitivity analysis (Figure 3) suggest that the variance within the districts is largely caused by the weighting method. For this reason, the variance conditional on the weighting method (here: PCA) is shown in Figure 5.

Figure 5 - Conditional variance of the composite indicators using weights resulting from PCA at district level



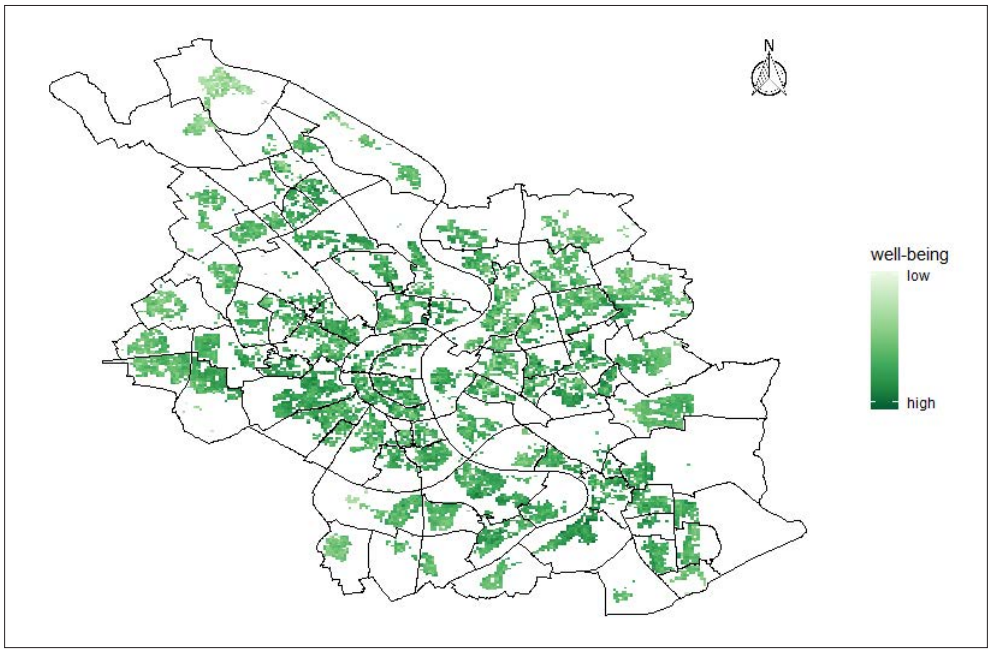
Source: Own illustration

As mentioned above, if various sub-indicator form a group, equal weighting might effectively result in unequal weights. For example, distances to primary schools, museums, hospitals, libraries and play- and sports grounds could be grouped into a single sub-indicator infrastructure. Assigning a weight of $w_q = 1/12$ to each of the q sub-indicators, results in the infrastructure related sub-indicators having a weight of 5/12 in total. This might explain why the variance is particularly high in the north-western and central districts. The infrastructure in the city centre is usually better developed than in the rest of the city. The central distances have the shortest distances and the north-western districts have the highest distances. It should be noted that the results might also be distorted as closer hospitals etc. can be located outside Cologne. However, in this study we concentrate exclusively on the urban area of Cologne. It would be desirable for future investigations to group the sub-indicators before weighting them. Furthermore, it would be sensible to include influential factors beyond the border of the region of interest in the analysis of well-being.

4.3 Composite indicator of well-being at grid cell level

The composite indicator at 100 metre grid cell level is illustrated in Figure 6 based on all sub-indicators normalised by standardisation and weights resulting from PCA. The unemployment numbers are downscaled by dasymetric mapping.

Figure 6 - Well-being indicator values at grid cell level

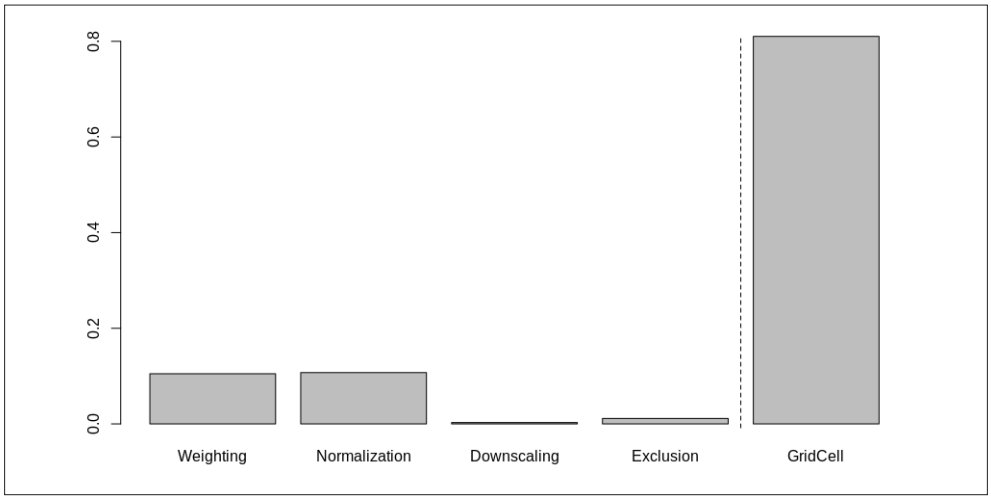


Source: Own illustration

The composite indicator at 100 metre grid cell level has 264 different input combinations per grid cell. We consider four downscaling methods (centroid assignment, dasymetric mapping, areal weighting and the ratio-synthetic estimator), three normalisation schemes (min-max method, ranking and standardisation), two weighting possibilities (based on PCA and equal weighting) and the exclusion of one indicator in each step. This results in a decision matrix of size 2.589.048 5 (number of construction steps). Each row has a different composite indicator as result. In order to be able to handle the size, 2.000 construction possibilities were drawn based on LP-Tau quasi-

random numbers on the interval [0, 1] (Münnich and Seger, 2014; Saltelli *et al.*, 2000). Figure 7 quantifies the impact of the normalisation scheme, weighting choice, downscaling method, exclusion of sub-indicators and the grid cell itself on the output variance. Note again that grid cell identity is not a trigger per se. It was included in the analysis to set the effect of construction decisions in relation to the regional heterogeneity in the results.

Figure 7 - Total-order effects of the sensitivity analysis at grid cell level



Source: Own illustration

As expected, the sensitivity analysis of the well-being indicator at 100 metre resolution reveals that grid cells have the largest impact on the output variance, followed by the weighting and normalisation methods. However, the results of the sensitivity analysis have to be treated with caution as it is based on 2.000 construction possibilities only.

At both district and grid cell level the differences between the areas of interest play a dominant role in the sensitivity analysis. This is even more pronounced on the very fine resolution level of grid cells. Altogether, the results indicate that there is a relevant heterogeneity of well-being at the regional level and, thus, confirm that efforts to measure this multidimensional concept at the local level are worth-wile. A second central conclusion is that, even if composite indicators are sensitive to construction decisions, the effect of these decisions does not "mask" the actual regional differences in indicator results.

5. Concluding remarks

In this article, we explore the potentials of using remote sensing data for local level estimation of well-being. So far, the analysis of well-being is mainly focussed on the country-level. However, differences in central dimensions of well-being, such as material living conditions or the preconditions of social interaction and the quality of leisure time, often exist at street level. Therefore, we combine survey data and remote sensing data to enable an analysis on the very low aggregation level of city districts and 100 metre grid cells. We present different sources of remote sensing data and create infrastructure-related sub-indicators in the composite well-being indicator using tools from the geosciences. Survey data are usually provided at administrative levels, whereas remote sensing data are available at small scale resolutions. Therefore, different upscaling and downscaling methods are introduced. We determine a composite indicator for well-being and quantify the impact of the different scaling techniques and other construction decisions by means of a sensitivity analysis with the scaling techniques, the normalisation scheme, the weighting methods and the exclusion of sub-indicators as uncertainty factors.

At district level, the incorporation of remote-sensing data is very promising. We can show that the upscaling methods only account for a minor proportion of the output variance. Following our application, the weighting scheme is the construction decision with the largest impact on indicator results. This can be attributed to the fact that, among other, equal weights are assigned to each sub-indicator resulting in an actual overweight of the infrastructure-related indicators. For future research, it would, therefore, be desirable to conduct the sensitivity analysis with grouped sub-indicators again.

The analysis at grid cell level is methodologically more challenging. The data at 100 metre resolution contain many empty cells as values lower than three are not published due to confidentiality reasons. Moreover, the change of resolution from district to 100 metre grid cell level is large, which comes at the cost of quality of the estimates at grid cell level. Generally, data availability on this very fine resolution level is largely restricted, so that more complex – and probably better – downscaling techniques could not be applied. For future research, we envisage a close collaboration with official statistics in order to build a data basis that will enable more elaborate approaches from the field of small area statistics. At both district and grid cell

level, the variation of well-being between the areas of interest was included in the analysis as a reference, *i.e.* to set the variability introduced through the construction decisions in relation to the actual heterogeneity between districts. At both levels, this variation between regional entities was the most relevant source of variability. There, thus, is a relevant heterogeneity on these low resolution levels which confirms our motivating notion that the micro-location matters. Further, even if composite indicators are sensitive to construction decisions, the actual differences between regional entities are not offset by these construction decision. This study can be seen as feasibility study showing that further research in this area could open up many possibilities. In particular, the downscaling methods have to be further developed. Moreover, alternative remote sensing data, such as land surface temperature, could be included as an environmental variable to describe human heat stress. All in all, the incorporation of remote sensing data has a huge potential for analyses of living conditions at local level and should be further investigated.

Acknowledgments

The research was conducted within the MAKSWELL project (hiips://www.makswell.eu), funded by the EU under Horizon 2020 (H2020-SC6-CO-CREATION-2017; 2nd and 5th author). Further collaboration took place with the MikroSim project (DFG research unit FOR 2559, hiips://www.mikrosim.uni-trier.de), funded by the German Research Foundation (3rd and 5th author), the REMIKIS project, funded by the Nikolaus Koch Stiftung (1st and 5th author), and the Trier Centre of Sustainable Systems, funded by the Rhineland-Palatinate Research Initiative (4th and 6th author).

We thank Professors Maria Giovanna Ranalli and Li-Chun Zhang for the invitation for this paper. Further, we thank Professor Alessandra Petrucci for inviting the 5th author to the ITACOSM conference.

Finally, we thank the editors and reviewers for the very positive and valuable comments that helped improving the readability of the paper.

References

Agresti, A. 2002. *Categorical Data Analysis*. Hoboken, NJ, U.S.: John Wiley & Sons.

Bidot, C., H. Monod, and M.-L. Taupin. 2018. *A Quick Guide to multisensi, an R Package for Multivariate Sensitivity Analyses*. hiips://cran.r-project.org/web/packages/multisensi/vignettes/multisensi-vignette.pdf.

City of Cologne. 2014. Offene Daten Köln. hiips://www.offenedaten-koeln.de/dataset (September 6th 2019).

City of Cologne. 2017. *Statistische Daten - Thematische Karte*. hiips://www.stadt-koeln.de/politik-und-verwaltung/statistik/statistische-daten-thematische-karte (September 6th 2019).

Copernicus. 2019a. *High Resolution Layers*. hiips://land.copernicus.eu/pan-european/high-resolution-layers (October 21st 2019).

Copernicus. 2019b. *Imperviousness*. hiips://land.copernicus.eu/pan-european/high-resolution-layers/imperviousness (October 21st 2019).

Deming, W. E., and F. F. Stephan. 1940. "On a Least Squares Adjustment of a Sampled Frequency

Table When the Expected Marginal Totals Are Known". *The Annals of Mathematical Statistics* 11(4): 427–444.

Dostal, L., S. Gabler, M. Granninger, and R. Münnich. 2016. "Frame Correction Modelling with Applications to the German Register-Assisted Census 2011". *Scandinavian Journal of Statistics*, 43(3): 904–920.

Easterlin, R.A. 1974. "Does Economic Growth Improve the Human Lot? Some Empirical Evidence". In David, P.A., and M.W. Reder (Eds.). *Nations and Households in Economic Growth*: 89–125. Cambridge, MA, U.S.: Academic Press.

Eicher, C.L., and C.A. Brewer. 2001. "Dasymetric Mapping and Areal Interpolation: Implementation and Evaluation". *Cartography and Geographic Information Science*, 28(2): 125–138.

Engstrom, R., J. Hersh, and D. Newhouse. 2017. "Poverty From Space: Using High-Resolution Satellite Imagery for Estimating Economic Well-Being". *Policy Research Working Paper 8284*, World Bank.

European Commission. 2019. *Inspire Data Specifications*. hiips://inspire.ec.europa. eu/training/inspire-data-specifications (November 6th 2019).

European Foundation for the Improvement of Living and Working Conditions - Eurofound. 2020. "What Makes Capital Cities the Best Places to Live?". *European Quality of Life Survey 2016 series*. Luxembourg: Publications Office of the European Union.

Eurostat. 2019. *Quality of Life Indicators - Measuring Quality of Life*. Luxembourg: Publications Office of the European Union. hiips://ec.europa.eu/eurostat/web/gdp-and-beyond/quality-of-life (August 13th 2019).

Eurostat. 2017. *Methodological Manual on City Statistics*. Luxembourg: Publications Office of the European Union. hiips://ec.europa.eu/eurostat/web/products-manuals-and-guidelines/-/KS-GQ-17-006 (May 5th 2020).

Eurostat. 2020. *Sustainable Development Indicators. SDG 3: Good health and well-being*. hiips://ec.europa.eu/eurostat/de/web/sdi/good-health-and-well-being (May 6th 2020).

Federal Statistical Office and the Statistical Offices of the Länder. 2018. *Ergebnisse des Zensus 2011 zum Download - erweitert*. hiips://www.zensus2011.de/DE/Home/ Aktuelles/DemografischeGrunddaten.html (September 6th 2019).

Fernandes, R.A., J.R. Miller, J.M. Chen, and I.G. Rubinstein. 2004. "Evaluating Image-Based Estimates of Leaf Area Index in Boreal Conifer Stands Over a Range of Scales Using High-Resolution CASI Imagery". *Remote Sensing of Environment*, 89(2): 200–216.

Geofabrik GmbH and OpenStreetMap contributors. 2018. *Download OpenStreetMap Data for this Region: Regierungsbezirk* Köln. hiip://download.geofabrik.de/europe/germany/nordrhein-westfalen/koeln-regbez.html (September 6th 2019). License: Open Database License.

Ghosh, T., S.J. Anderson, C.D. Elvidge, and P.C. Sutton. 2013. "Using Night time Satellite Imagery as a Proxy Measure of Human Well-Being". *Sustainability*, 5: 4988–5019.

Goodchild, M.F., and N.S.-N. Lam. 1980. "Areal Interpolation: A Variant of the Traditional Spatial Problem". *Geo-Processing*, 1: 297–312.

Green, N.E. 1957. "Aerial Photographic Interpretation and the Social Structure of the City". *Photogrammetric Engineering*, 23: 89–96.

Hernandez, A.L. 2016, October. *Multivariate Structure Preserving Estimation for Population Compositions*. Ph.D. Thesis, University of Southampton.

Homma, T., and A. Saltelli. 1996. "Importance Measures in Global Sensitivity Analysis of Nonlinear Models". *Reliability Engineering and System Safety*, 52(1): 1–17.

Ireland, C.T., and S. Kullback. 1968. "Contingency Tables With Given Marginals". *Biometrika*, 55(1): 179–188.

Italian National Institute of Statistics – Istat. 2019. *BES at Local Level*. Roma, Italy: Istat. https://www.istat.it/en/well-being-and-sustainability/the-measurement-of-well-being/bes-at-local-level (May 5th 2020).

Kyriakidis, P.C. 2004. "A Geostatistical Framework for Area-to-Point Spatial Interpolation". *Geographical Analysis*, 36(3): 259-289.

Lo, C.P., and B.J. Faber. 1997. "Integration of Landsat Thematic Mapper and Census Data for Quality of Life Assessment". *Remote Sensing of Environment*, 62: 143–157.

Martinez, B., F. Garcia-Haro, and F.C. de Coca. 2009. "Derivation of High-Resolution Leaf Area Index Maps in Support of Validation Activities: Application to the Cropland Barrax Site". *Agricultural and Forest Meteorology*, 149(1): 130–145.

Moretti, A., N. Shlomo, and J.W. Sakshaug. 2019. "Multivariate Small Area Estimation of Multidimensional Latent Economic Well-being Indicators". *International Statistical Review*, 88(1): 1-28.

Münnich, R., J.P. Burgard, and M. Vogt. 2013. "Small Area-Statistik: Methoden und Anwendungen". *AStA Wirtschafts- und Sozialstatistisches Archiv*, 6(3): 149–191.

Münnich, R.T., and J.G. Seger. 2014. "Impact of Survey Quality on Composite Indicators". *Sustainability Accounting, Management and Policy Journal*, 5(3): 268–291.

Nardo, M., M. Saisana, A. Saltelli, and S. Tarantola. 2005. *Knowledge Economy Indicators. Workpackage 5: Input to Handbook of Good Practices for Composite Indicators' Development*. Roma, Italy: Ispra - Joint Research Centre.

Nicoletti, G., S. Scarpetta, and O. Boyland. 2000. Summary Indicators of Product Market Regulation With an Extension of Employment Protection Legislation. *Economics Department Working Papers No. 226* (ECO/WKP(99)18). Paris, France: OECD.

Organisation for Economic Co-operation and Development - OECD. 2020. *How's Life? 2020: Measuring Well-Being*. Paris, France: OECD Publishing.

Organisation for Economic Co-operation and Development - OECD, Statistics Directorate and Directorate for Science, Technology and Industry, and Joint Research Centre - JRC of the European Commission in Ispra, Italy, Applied Statistics and Econometrics Unit. 2008. *Handbook on Constructing Composite Indicators: Methodology and User Guide*. Paris, France: OECD Publishing.

Office for National Statistics - ONS. 2019. "Personal Well-Being in the UK: April 2018 to March 2019". *Statistical bulletin*.

hiips://www.ons.gov.uk/peoplepopulationandcommunity/wellbeing/bulletins/measuringnationalwellbeing/april2018tomarch2019#personal-well-being-by-local-area (May 5th 2020).

OpenStreetMap contributors. 2019. QuickOSM Plugin (QGIS). hiips://github.com/3liz/QuickOSM. License: Plugin: License GPL Version 2.

Pfeffermann, D. 2013. "New Important Developments in Small Area Estimation". *Statistical Science*, 28(1): 40–68.

Purcell, N.J., and L. Kish. 1980. "Postcensal Estimates for Local Areas (or Domains)". *International Statistical Review/Revue Internationale de Statistique*, 48(1): 3–18.

Rao, J.N.K., and I. Molina. 2015. *Small Area Estimation*. Hoboken, NJ, U.S.: John Wiley & Sons, *Wiley Series in Survey Methodology*.

Saltelli, A., K. Chan, and E.M. Scott. 2000. *Sensitivity Analysis: Gauging the Worth of Scientific Models*. Chichester: Wiley.

Saltelli, A., M. Ratto, T. Andres, F. Campolongo, J. Cariboni, D. Gatelli, M. Saisana, and S. Tarantola. 2008. *Global Sensitivity Analysis: The Primer*. Chichester, UK: John Wiley & Sons Ltd.

Shuvo Bakar, K., N. Biddle, P. Kokic, and H. Jin. 2020. "A Bayesian Spatial Categorical Model for Prediction to Overlapping Geographical Areas in Sample Surveys". *Journal of the Royal Statistical Society: Series A (Statistics in Society*), 183(2): 535–563.

Sobol, I.M. 1993. "Sensitivity Estimates for Nonlinear Mathematical Models". *Mathematical and Computational Experiments*, 1(4): 407–414.

Stiglitz, J.E., A. Sen, and J.-P. Fitoussi. 2009. Report by the Commission on the Measurement of Economic and Social Progress. hiips://ec.europa.eu/eurostat/documents/118025/118123/Fitoussi+Commission+report (August 13th 2019).

United Nations. 2019. *The Sustainable Development Goals Report*. hiips://unstats.un.org/sdgs/report/2019/ (September 20th 2019).

Wu, H., and Z.-L. Li. 2009. "Scale Issues in Remote Sensing: A Review on Analysis, Processing and Modeling". *Sensors*, 9: 1768–1793.

Yang, W., and J.W. Merchant. 1997. "Impacts of Upscaling Techniques on Land Cover Representation in Nebraska, U.S.A". *Geocarto International*, 12(1): 27–39.

Zhang, J., P.M. Atkinson, and M.F. Goodchild. 2014. Scale in Spatial Information and Analysis. Boca Raton, FL, U.S.: CRC Press.

Zhang, J., and N. Yao. 2008. "The Geostatistical Framework for Spatial Prediction". *Geo-Spatial Information Science*, 11(3): 180–185.

Zhang, L.-C., and R.L. Chambers. 2004. "Small Area Estimates for Cross-Classifications". *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(2): 479–496.

Civil justice: a methodological analysis for assessing efficiency

Maria Filomeno, Irene Rocchetti [1]

Abstract

The efficiency of judiciary offices may be connected to different variables. In the aim of deriving a unique statistical measure of the efficiency of each Court of first instance, based on different input and output variables, we applied a Data Envelopment Analysis (DEA), able to produce an efficiency index varying between 0 and 1 (unity representing major efficiency). Given that the resources available in the different judiciary offices are fixed (i.e. the eventual offices organigram variation), we chose to apply an output-oriented DEA model. We further analysed the potential influence of the activity of the Council on the performance of judgmental offices measured in terms of organisation variables, through Beta Regression models on the efficiencies resulted from DEA models.

Keywords: Judicial efficiency, output oriented DEA, beta regression.

1 Maria Filomeno (m.filomeno@cosmag.it); Irene Rocchetti (i.rocchetti@cosmag.it), Consiglio Superiore della Magistratura – CSM, Italy.

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.

1. Introduction

In the last years there has been an increasing interest at the international level for the organisation of the different judicial systems. The need of a compared analysis of the performances of the judicial systems led to the importance of measuring the efficiency of the offices also according to their organisational aspects. Economic literature defines the efficiency meanly as (Landi *et al.*, 2016):

- the capacity, given the available resources, to solve controversies in a reasonable time;
- the quality of sentences, in terms of accuracy and certainty of decisions;
- the independence of the judgment (both in terms of impartial judges' decisions and of independence between the different degrees of judgement).

While the last two efficiency dimensions are not easy to be measured, nowadays the duration of proceedings is one of the indicators mainly used in order to measure the efficiency together to the percentage of ultra triennial pending proceedings, the clearance rate (ratio between resolved and new incoming proceedings), the disposal rate (ratio between resolved proceedings and workload, computed as sum of initial pending proceedings and resolved cases), etc. Just because of the existence of various judicial performance indexes and their different composition, it is not simple to assess uniquely and in a clear and synthetic way the efficiency of a judicial office. The main purpose of this study is the one of producing a unique efficiency measure for the Courts of first instance (in Italy they are 140), basing on the already existent indicators, which allows at the same time a proper assessment of the phenomena and a comparative analysis among the different offices by considering also their organisational resources. The second purpose of the study is the one of analysing the efficiency outcomes obtained by a first analysis and their relation with the organisational capacity of the judicial offices, in the aim of verifying the existence of a possible impact of the activities of the Superior Council of Judiciary (CSM) taken in the last years in order to improve the organisational capacities of the judicial offices. In section 2 materials and DEA method used will be illustrated, in section 3 the effective organisation of the judicial offices and the measured efficiency

is presented (for the whole civil sector and for the two civil sub sectors), in section 4 conclusions are illustrated.

2. Materials and DEA Method

In the aim of deriving a statistical measure of the efficiency of each italian Court of first instance, based on different input and output variables, we applied a Data Envelopment Analysis (DEA). DEA methods consist in linear programming models which locate the best offices and allow locating the other offices by comparison. As alternative we could choose to compute a composite indicator by considering many variables usually accounted for to understand office judicial efficiency; however the choice of proper weights to weigh the composite indicator with would risk to be very arbitrary and to produce politic debates. To assess the efficiency of units, the linear programme is needed to be solved as many times as the number of the units, each time changing the reference unit; the optimisation problem is solved by producing a frontier of efficient units (Courts). These models are non-parametric and the relation existing between performance and endowment of resources is not mathematically exploited; this point is very important in the judiciary context, such as in public administrations, given the difficulty in defining objectively relations among performance indicators and their priorities. DEA models produce efficiency indexes which vary between 0 (no efficiency) and 1 (perfect efficiency) and can be of two different type: *input-oriented* and *output-oriented* DEA models. In the first case, efficient units are the ones using a lower number of resources to obtain a given output, in the second case the efficient units are the ones that with a given number of resources available (input) obtain the greatest possible output. We applied an *output-oriented* DEA model given that the endowment of available resources in the different judicial offices is fixed or can be modified only through measures taken or approved by the Ministry of Justice or/and by the Superior Council of Judiciary; let us consider for example the eventual working staff variation in the judicial offices. In detail, the chosen DEA model produces an efficiency frontier made of a virtual combination of units producing an higher output with the same input; units are dominated, proportionally or not proportionally in terms of output, by units standing on the frontier. The goal is to estimate judicial office efficiency by maximizing the output variables (such as intra triennial proceedings) by remaining fixed the input variables (such as the Judicial personnel coverage rate), see Pardieri *et al.*, 2000. Hence, the efficiency measure of a given civil office obtained by the output-oriented

DEA model indicates how much the office produces given its possibilities: *i.e.* a value of estimated efficiency equal to 0,8 indicates that the office produces at the 80% of its possibilities and basing on the same endowment of resources it could obtain an output greater than 20% of the one obtained.

DEA models are based on data belonging to the quarter monitoring of the Ministry of Justice referring to the year 2018 and concerning both the whole civil sector and the two sub-civil sectors of Civil litigation, whose data are contained in the SICID (Sistema Informativo Contenzioso Civile Distrettuale) register, and of Insolvency Procedures registered in the SIECIC register (Sistema Informativo Esecuzioni Civili Individuali e Concorsuali). The input variables considered in the DEA model are:

- Judicial personnel coverage rate, year 2018 (Judicial present personnel/ judge working staff in courts);
- Number of incoming cases over the judge working staff in courts in the civil sector (standardised variable).

These variables have been chosen because the office, to carry out its judicial activity, needs to have both a proper judicial working staff coverage and a sustainable workload per each judge (represented by the number of incoming proceedings per judge).

The output variables taken into consideration are the ones synthesising the productivity of an office both in terms of duration and in terms of disposal capacity, in detail:

- The percentage of infra-triennial pending proceedings (enrolled in the last three years) over the total of pending proceedings at the 31/12/2018;
- Ratio between ultra-triennial proceedings 2017 and ultra-triennial proceedings 2018;
- Number of resolved proceedings over the working staff assigned to the civil sector;
- Clearance rate (resolved over incoming cases).

These indicators measure the "state of health" of an office and its productivity. Indicators have been computed such that their direction was

coherent with the goals of the study; for example, it has been considered the percentage of infra-triennial pending cases given that our purpose is the one of maximizing the percentage of recent backlog, just because an office to be efficient has to reduce the backlog, in particular the ultra-triennial one.

In the following are reported the results from the model, both for the whole civil sector and for the two considered sub civil sectors, and also the dimensional and geographical distribution of the offices according to the efficiency indexes obtained.

2.1 Results from the DEA model

Table 1 shows the most efficient courts of first instance according to the DEA model (efficiency score equal to 1), together with the dimension, the geographical division judicial offices belong to, the percentage of infra-triennial proceedings and the clearance rate.

As we can notice, the most efficient offices are located mainly in the North of Italy (Biella, Bolzano, Ferrara, Gorizia, Ivrea, Novara and Savona) and are especially small or medium small Courts. There are five offices in the South of Italy (Avezzano, Campobasso, Crotone, Napoli Nord, e Tempio Pausania) and one office of the Centre (Livorno).

Table 1 - Efficient Italian Courts of first instance and their characteristics

Courts	Dimension	Geographical distribution	% **Infra-triennial pending cases**	Clearance rate
Avezzano	Small	South	0.84	0.11
Biella	Small	North	0.75	1.05
Bolzano	Medium Small	North	0.89	1.04
Campobasso	Small	South	0.82	1.05
Crotone	Medium Small	South	0.72	1.28
Ferrara	Medium Small	North	0.93	1.07
Gorizia	Small	North	0.88	1.08
Ivrea	Medium Small	North	0.9	1.07
Livorno	Medium Small	Centre	0.81	1.12
Napoli Nord	Medium big	South	0.93	0.84
Novara	Small	North	0.78	1.15
Savona	Medium Small	North	0.87	1.13
Tempio Pausania	Small	South	0.57	0.99

A particular case among the efficient Courts is just the one of Napoli Nord, an office located in the South of Italy of medium big dimension, with an increased number of ultra-triennial pending proceedings between 2017 and 2018, a number of resolved proceedings per judge not much high and a clearance rate lower than unity but also with the lowest percentage of ultra-triennial pending proceedings due to the fact that, being build recently, it didn't have the time to rack up backlog.

The map below gives a graphic representation of the distribution of the Courts of first instance in Italy according to the efficiency index resulted from the DEA model.

Figure 1 - Distribution of the Italian Courts according to the efficiency indexes obtained through DEA

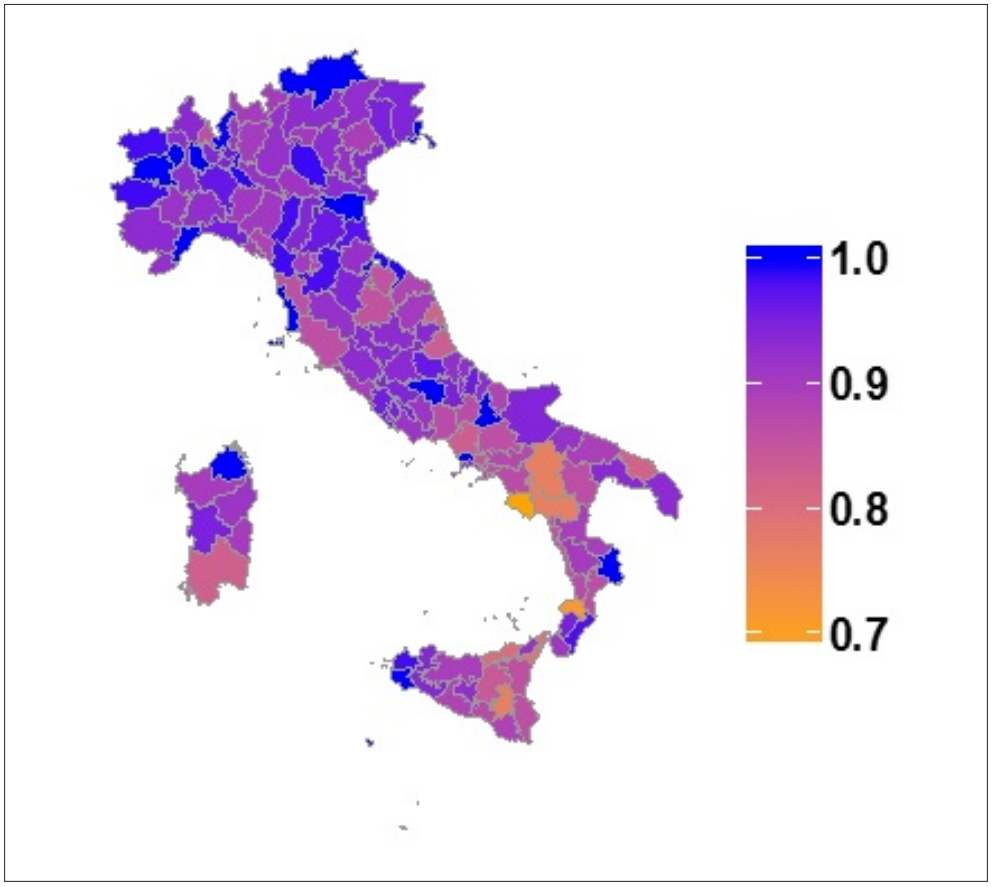


Table 2 shows the number of offices which have increased or not increased the estimated efficiency (through the DEA model) between 2017 and 2018 per dimension of the Courts. The Courts of Gorizia, Ferrara, Ivrea, Napoli Nord and Savona are among the most efficient in both considered years. Generally the 61% of the offices (86 Courts) has increased the estimated efficiency through the DEA model between the 2017 and the 2018 or has maintained the maximum efficiency, equal to unity; for the other offices the efficiency indicator has remained the same or has decreased. The following offices have increased the efficiency in 2018: the big and metropolitan Courts; the 55,6% of the medium big offices, the 63% of the medium small courts and the 56,9% of the small offices.

Tables 3 and 4 show the list of offices estimated as the most efficient according to the DEA models applied to the same input - output variables mentioned before but related respectively to the civil litigations sector (SICID) and to the insolvency procedures (SIECIC). As far as the SICID sector is concerned, the efficient offices and the related output indicators are listed in Table 3. Some offices, such as Ivrea, Napoli Nord, Savona and Ferrara confirm their position of best offices also in the civil litigations sector, characterising especially for a consistent number of infra-triennial pending proceedings and for a good clearance rate (equal or greater to 1).

Table 2 - Distribution of the offices which have increased or not increased the efficiency between the 2017 and the 2018

Dimension	Efficiency increase 2017/2018	No efficiency increase 2017/2018	Total
Big	3		3
Metropolitan	3		3
Medium big	10	8	18
Medium small	41	24	65
Small	29	22	51
Total	86	54	140

Table 3 - Efficient Italian Courts of first instance and their characteristics - SICID

Courts	Dimension	Geographical distribution	% Infra-triennial pending cases	Clearance rate
Aosta	Small	North	0.95	1
Arezzo	Medium small	Centre	0.86	1
Ferrara	Medium small	North	0.96	1
Gorizia	Small	North	0.91	1
Isernia	Small	South	0.62	1.1
Ivrea	Medium small	North	0.95	1.1
Lodi	Small	North	0.9	1
Napoli Nord	Medium big	South	0.95	0.9
Rieti	Small	Centre	0.8	1.1
Savona	Medium small	North	0.96	1.1
Sulmona	Small	South	0.98	1
Tivoli	Medium small	Centre	0.83	0.9
Trieste	Medium small	North	0.94	0.9

Table 4 - Efficient Italian Courts of first instance and their characteristics - SIECIC

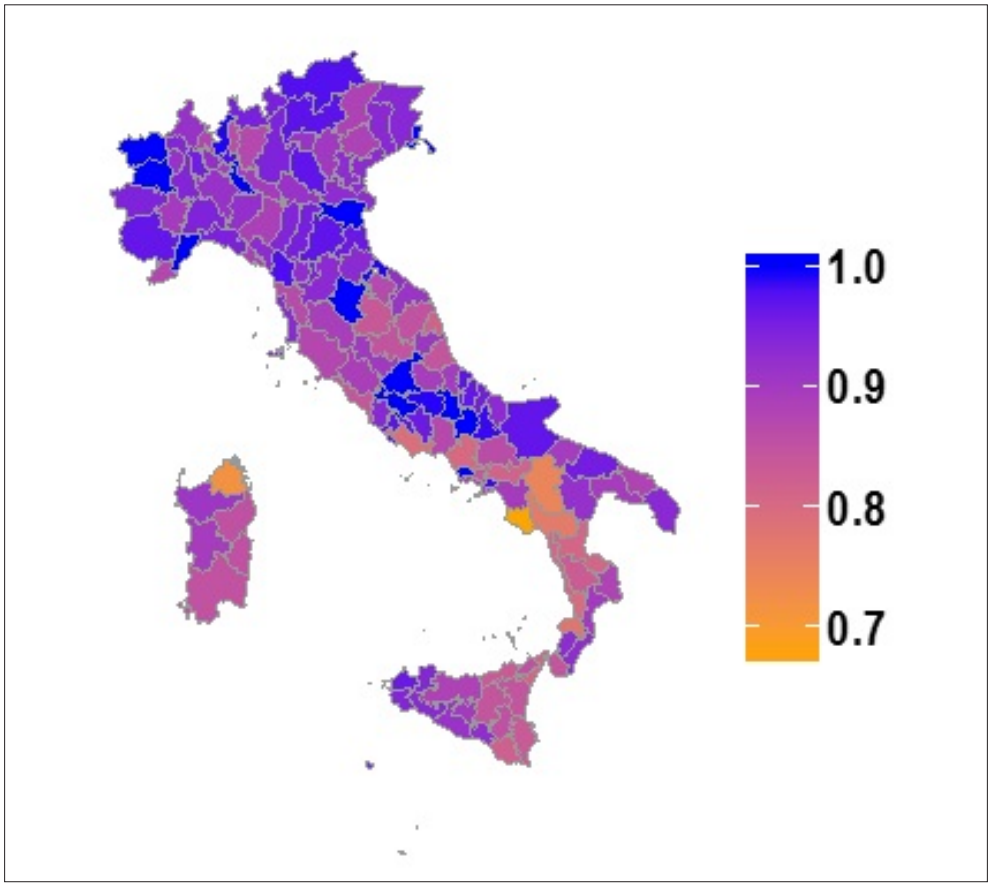
Courts	Dimension	Geographical distribution	% Infra-triennial pending cases	Clearance rate
Biella	Small	North	0.57	1.27
Catanzaro	Medium small	South	0.88	0.87
Ferrara	Medium small	North	0.89	1.21
Livorno	Medium small	Centre	0.62	1.33
Napoli	Metropolitan	South	0.79	1.14
Novara	Small	North	0.68	1.43
Tempio Pausania	Small	South	0.45	1.10

In the insolvency procedures sector, the efficiency index resulted from the DEA model varies more across the territory and medium values of the index are registered also in the North of Italy (see Figure 3).

3. Judicial offices organisation and measured efficiency

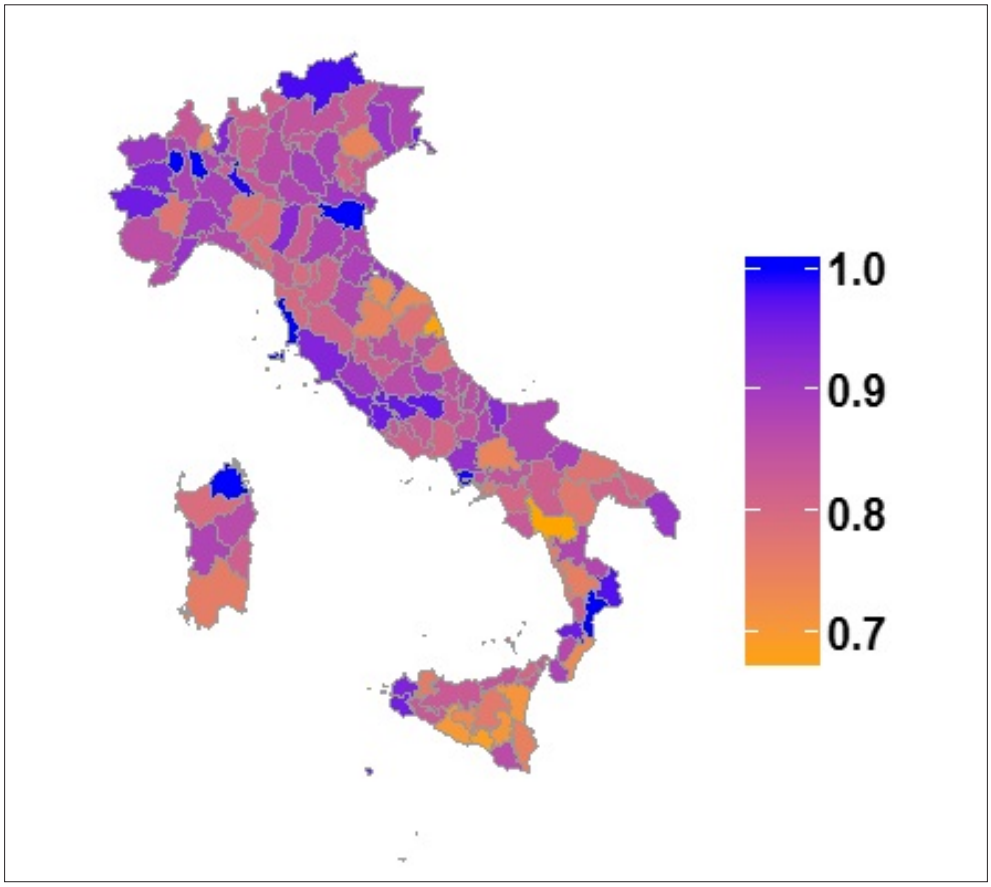
The Superior Council of Judiciary and in particular the VII Commission, has made an effort to spread among the heads of the offices a greater organisational culture, focussing the attention on the reduction of the civil pending proceedings, mainly on the more ancient enrolled cases for whom the probability to incur in economical sanctions (L.Pinto 24 march 2001, n. 89) for the length of the proceedings duration is higher. With the D.L. 98/2011 provisions have been emanated aiming at improving the efficiency of the judiciary system and the quick resolving of the controversies through the introduction of the management programmes. The management programmes, yearly filled by the Courts of first and second instance, have become during the time an important tool of planning and support of the judicial activity of the offices monitored by the Superior Council of Judiciary whose duty is also the analysis of the data resulting from them. In the management programmes deposited within the 31 of January 2018, judgmental offices had to indicate the percentage of ancient pending proceedings (ultra triennial for the Courts of first instance and ultra biennial for the Court of second instance) they would try to dispose of within the 31 December of the year 2018. This kind of data has been taken into consideration for the construction of a variable "Planning capacity" given by the ratio between the real disposal, measured as the difference between the ultra triennial pending proceedings at the end of 2018 and the ones at the end of 2017 and the goal (in terms of pending cases to be dispose of) planned by the heads of the offices in the management programmes for the year 2018. This variable related to the planning capacity has been categorised in three classes, according to the value assumed:

Figure 2 - Distribution of the italian Courts according to the efficiency indexes obtained through DEA - Sicid



- Courts with *adequate target* planning capacity have a value of the index between 0.5 and 1.5 extremes included;
- Courts with *under target* planning capacity have an index higher of 1.5; such offices are considered *prudent* in the definition of the goal, in fact the real decreasing or the backlog is higher (more of the 50%) than the goal of reduction indicated in the management programmes;
- Courts with *over target* planning capacity have an index smaller than 0.5.

Figure 3 - Distribution of the italian Courts according to the efficiency indexes obtained through DEA - Siecic



When it is mentioned in the text *not adequate* planning capacity it is related to Courts with under or over target planning capacity. The planning capacity variable and other indicators directly available or constructed on the base of the Justice Ministry data, have been used in order to analyse the existence of an eventual effect on the efficiency previously measured through the DEA models. The goal is twofold: on one hand the one of verifying the congruity between organisational capacity and real levels of judicial efficiency of the offices and thus the utility of the used planning tools, on the other hand the potential influence of the Council activity on such planning in order to improve the performance of the same offices.

3.1 The Beta regression model

The statistical model chosen for our purpose is the Beta Regression model; this choice is linked to the fact that the rv Beta has support in (0,1) and thus it is adequate to model variables such as ratios and proportions; such variable does not belong to the exponential family but the model based on its distribution is similar to a generalised linear model. We hypothesise $y_1, \ldots, y_n$ independent realisations of the rv $Y_i \simeq Beta(\mu_i, \phi)$, where μ_i and φ are respectively mean and precision parameter unknown; $x_1, \ldots, x_k$ constants known and fixed (k<n) and $\beta=(\beta_1, \ldots, \beta_k)$ vector of unknown parameters ($\beta_k \in$ R).. The formalisation of the model is the following

$$g(\mu_i) = \sum_{j=1}^{k} x_{ij}\beta_j = \eta_i$$

where $g(.)$ is the link function, mapping the mean from a subspace of R to all R, there are possible different choices (logit, probit, log log). In our case, the efficiency outcomes resulting from the DEA model assume value in the interval [0,1] extreme included; therefore the transformation $(y(n\ 1) + 0.5)/n$ (see Cribari-Neto *et al.*, 2010) has been applied, n corresponding to the total number of Courts, hence 140. The precision parameter φ, indirectly proportional to the variance of y, has been modelled by considering as explanatory variable the fact that the Court could be district or territorial, given that these two typologies of offices have a different organisation and this fact affects the variability of the distribution.

The Beta regression has been used to study the relation between the technical efficiency (dependent variable) together with the planning capacity, already mentioned, and other explanatory variables (factors which potentially affect and thus explain the efficiency outcome measured) both organisational, territorial and dimensional such as:

- the proceedings duration;
- the geographical division;
- the dimension of the offices;
- the number of *best practises* adopted by the judicial offices in the organisational context;

- the number of specialised sections (because it is assumed that a higher specialisation can produce a quicker disposal of the pending cases related);
- the number of active enterprises per Court, which could have an impact on the efficiency in the sector of the insolvency procedures (securities and real estate executions);
- the judicial working staff variation, as indicator of an organisational change;
- the turnover index (ratio between the number of judges left and the working staff for the years 2014-2018).

Variables which affect the efficiency in a statistical significant way are the duration of proceedings, the geographical division, the planning capacity referred to the last year and the working staff variation.

Three models have been applied: the first one for the whole civil sector, the other two models are one for each specific civil sector (SICID and SIECIC). Here only the results related to the whole civil sector and to the civil litigation one are reported; the results related to the SIECIC are not statistically significant.

Because of a strong association between the duration and the geographical division and the turnover a two step regression has been applied. The first step analysis is a linear regression of the duration where explanatory variables are the geographical division and the turnover. At the second step the residuals from the first step regression representing the duration net of the influence of turnover and geographical divisions is used as an explanatory variable in the beta regression model.

At the first step analysis it can be noticed that the more the turnover increases the more the proceedings duration increases as well (see Figure 4); furthermore North Italy offices are characterised by a duration of civil proceedings significantly smaller if compared with the offices of the Centre (-104 days), while the South offices have a duration of proceedings higher if compared to the ones belonging to the other two geographical divisions (see Table 5).

Figure 4 - Boxplot of the real duration of the proceedings in the civil sector per turnover classes

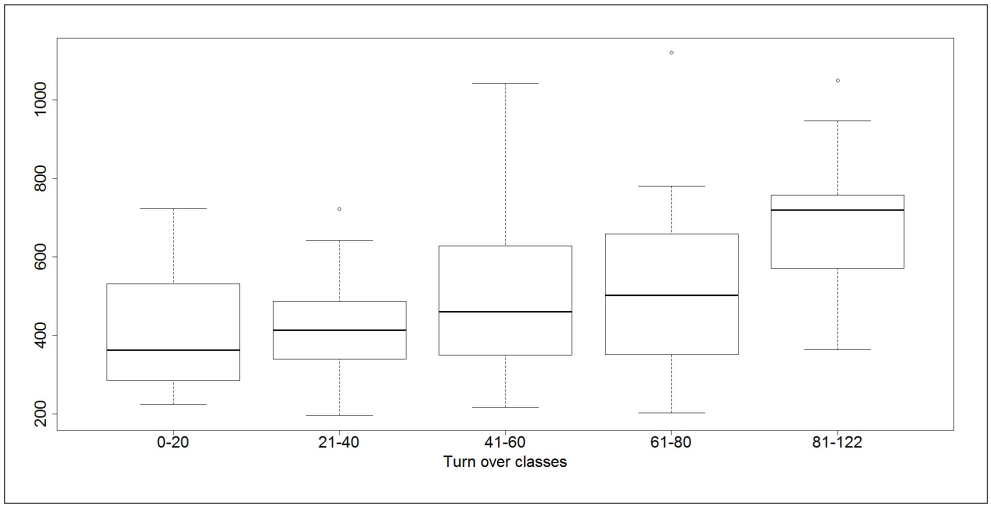


Table 5 - Regression model estimates - first step (a)

Parameter	Estimate	Std. Error	t value	Pr(>|t|)
Intercept	376	35.59	10.57	< 2e-16 ***
North	-104	32.06	-3.23	0.001 **
South	130	30.72	4.23	4.22e-05 ***
Turnover	1.73	0.53	3.24	0.001 **

(a) The statistically significant estimates (at least at the 95%) are the ones to whom correspond the symbols ".", *,**,*** in the column Pr(>|t|).

3.2 Analysis for the whole civil sector

In the following, results derived from the beta regression application (second step regression) on the efficiency measures obtained by using the DEA model, by considering the different exploratory variables already mentioned before are shown; among these independent variables which have the function of proxy of the interventions of the Csm in the aim of improving the organisational capacity of the judicial offices. The results of the statistical model show that, as we can expect, lower is the duration of the proceedings higher is the efficiency, even if the effect of such variable is very low (regression coefficient equal to -0.002, see Table 6) and the offices from the

North of Italy have an efficiency lower than the ones belonging to the other geographical divisions. In detail, the courts of first instance of the North of Italy have an estimated probability of being efficient *versus* not being efficient (ODDS RATIO, OR), higher than 1.6 times if compared to the offices of the Centre and about of 1.8 times if compared to the offices of the South. The interpretation of the estimated parameters through the beta regression is similar to the one of a logistic model (see Agresti, 2017).

Furthermore, offices with a not adequate planning capacity in the 2017 have a minor probability of being efficient in the 2018 *versus* the offices with an adequate target capacity and such estimate is statistically significant. The planning capacity variable related to the considered year (2018) is not significant in the beta regression, however analysing the increase or the not increase of the efficiency between a year and the other (2017-2018), it is clear that the most prudent courts have increased the efficiency more than the ones with an adequate target planning capacity and this fact could be due to the lower tendency to overestimate the goal of reduction of the ultra triennial pending proceedings leading to a more certain achievement of the goal.

Table 6 - Estimates of the beta regression model at the second step (a)

Parameter	Estimate	Std. Error	z value	Pr(>|z|)
Intercept	2.366	0.196	12.098	< 2e-16 ***
Duration (net)	-0.0002	0.0004	-5.183	2.19e-07 ***
North	0.487	0.167	2.914	0.003 **
South	-0.079	0.153	-0.514	0.607
Turnover	-4.079e-05	0.003	-0.014	0.989
Over target planning capacity(last year)	-0.247	0.129	-1.901	0.057.
Under target planning capacity(last year)	0.336	0.269	1.248	0.212
Phi coefficients (precision model with log link)				
Intercept	2.9751	0.1377	21.609	<2e-16 ***
district courts 1	0.6086	0.2961	2.055	0.039 *

.(a) The statistically significant estimates (at least at the 95%) are the ones to whom correspond the symbols ".., *,**,*** in the column Pr(>|z|)

The graphic below shows the trend of the efficiency resulted from the DEA model when the proceedings duration varies; it is possible to notice the decreasing trend of the efficiency of the offices at the increase of the proceedings duration expressed in days.

The geographical distribution besides being linked to the technical efficiency is linked to the planning capacity: the probability of having an adequate target planning capacity (*versus* having a not adequate capacity) of the judgemental courts of first instance of the South of Italy is lower than the one of the offices of the North and of the Centre.

The graphic below shows the distributions of the efficiencies in the different italian geographical divisions (applying a Gaussian Kernel Smoothing); we can notice that for the offices of the North the peak of the curve is closer to 1 (situation of highest efficiency) and the distribution is very concentrated around high values (0.9-1), while for the other geographical divisions the peak is closer to lower values of efficiency and the curves are platykurtic indicating a lower concentration around the maximum value and hence a more dispersed situation embracing both high and low efficiency values.

Furthermore we wanted to verify whether, together with the geographical distribution, the variables representative of the CSM activity, are in some way linked to the planning capacity codified in the three classes already mentioned before. The purpose in this case is the one of verifying whether, the fact that an office has disposed or not the percentage of ultra triennial pending proceedings prefixed and declared in the management programme for the 2017, is linked to other interventions began by the Council always in the aim of improving the judicial efficiency of the judgmental offices (*i.e.* spreading of best practises, number of specialised sections per offices, working staff variation, etc.). Hence a multinomial logistic model has been applied to estimate the probability of having a planning adequate target or under target (prudent offices) capacity *versus* an over target (not adequate) capacity on the base of some variables of intervention and organisation. The variable resulting statistically significant is the one related to the last working staff variation of the offices in the 2016 adopted by the Ministry of Justice, on which the Csm has given a judicial opinion motivated by modifying the initial proposal (see Table 7). The most of the offices which have disposed of how much they had prefixed or more (adequate target capacity or under target offices) has benefited of a variation in positive of the working staff, while the majority of the offices which do not have planning capacity has not obtained an increase or a decrease of the working staff.

Figure 5 - Relation between efficiency and proceedings duration

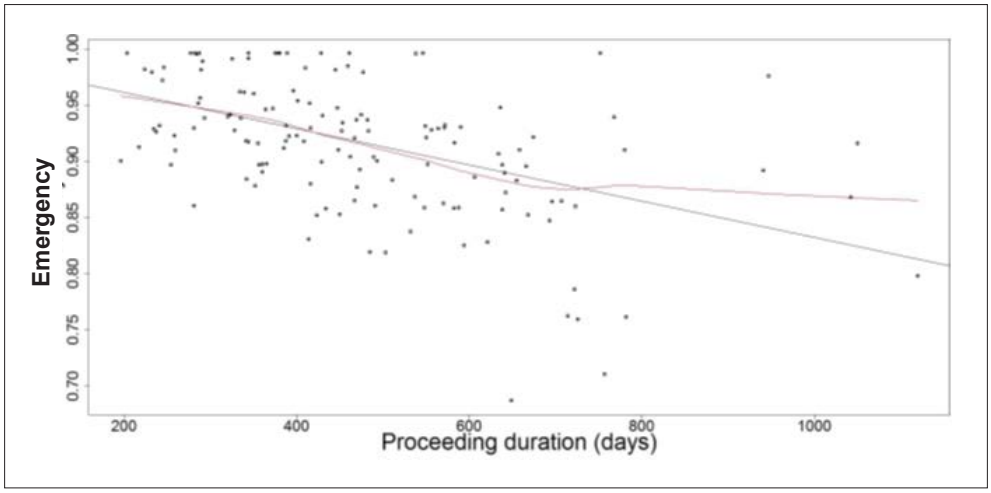


Figure 6 - Efficiency distribution per geographical division

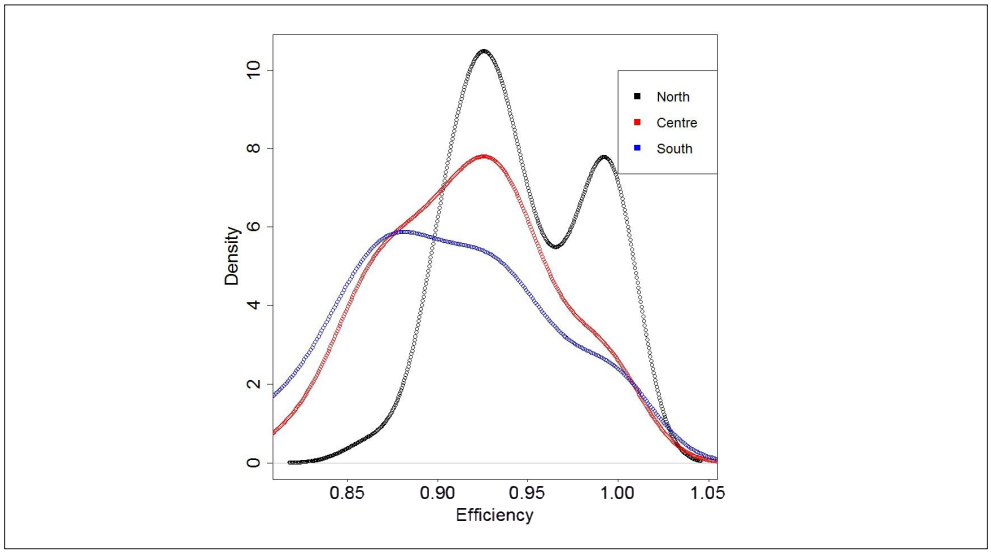


Table 7 - ODDS Ratio of the planning capacity with regard to the variation in terms of increase or decrease of the working staff (multinomial logistic model) (a)

	Intercept	Working staff increase (Ref=No variation)	Working staff decrease (Ref=No variation)
Adequate target capacity	2.12***	2.27**	1.88
Under target	0.19***	2.9	3.55

(a) The statistically significant estimates (at least at the 95%) are the ones to whom correspond the symbols ".", *,**,***

In detail, the Courts in which has been approved an increase of the working staff have a probability of having an adequate target planning capacity (*versus* the Courts which have an over target capacity) of about 2.3 times higher than the offices which did not have obtained any working staff variation: the offices knew how to govern the working staff variation. The positive effects reflect the adequacy of the modifications of the working staff, determined by the Ministry of Justice and by the Superior Council of Judiciary.

3.3 Analysis divided by civil sectors (SICID and SIECIC)

As already mentioned, for the civil litigation sector, the results obtained are similar to the ones got for the whole civil sector (Table 8), both per intensity and per direction of the relations between geographical division, turnover and duration of proceedings; also the influence parameters of these three variables on the efficiency measured through the DEA model are similar (Table 9).

Table 8 - Regression model estimate - first step

Parameter	Estimate	Std.Error	t value	Pr(>t)
Intercept	320.97	39.99	8.03	4.21e-13 ***
North	-106.36	36.02	-2.95	0.00371 **
South	145.19	34.52	4.21	4.68e-05 ***
Turnover	1.632	0.60	2.72	0.00742 **

The statistically significant estimates (at least at the 95%) are the ones to whom correspond the symbols ".", *,**,***

In detail, also for the civil litigation sector, the lower is the proceedings duration (net of the geographical division and the turnover) the greater is the probability of being efficient, even if the intensity of such phenomena is low. Courts from the North of Italy have an estimated probability of being

efficient higher if compared to the offices of the Centre and of the South; then the offices which in the 2017 had a under target planning capacity (and hence are prudent in the definition of the goals) have in the 2018 a probability of being efficient of about 1.6 times higher than the offices with adequate target planning capacity.

Furthermore, an important aspect related to the sub sector of the civil litigation is the one connected to the relation between working staff and efficiencies: offices with a decrease in the number of units in the working staff are characterised for a lower probability of being efficient *versus* the offices which did not have any variation.

As far as the insolvency procedures (SIECIC) are concerned, similar results have been obtained from the two step regressions, while the DEA model does not report statistically significant results but for the geographical distribution (significance at the 90%). For this reason these results are not reported in the following.

Table 9 - Beta regression model estimates at the second step (SICID)

Parameter	Estimate	Std.Error	z value	Pr(>|z|)
Intercept	2.167	0.185	11.695	< 2e-16 ***
Duration	-0.002	0.0004	-5.784	7.3e-09 ***
North	0.572	0.159	3.592	0.000329 ***
South	0.010	0.146	0.069	0.945
Turnover	-0.0003	0.003	-0.116	0.907
Over targ. plan. capacity last year	0.0416	0.130	0.320	0.749
Under targ. plan. capacity last year	0.438	0.208	2.112	0.035 *
Working staff variation +	0.023	0.128	0.184	0.854
Working staff variation -	-0.395	0.173	-2.278	0.023 *
Phi coefficient (*precision model with log link*)				
Intercept	3.107	0.137	22.691	<2e-16 ***
district courts 1	0.496	0.295	1.681	0.0928 .

The statistically significant estimates (at least at the 95%) are the ones to whom correspond the symbols “., *,**,***

4. Conclusions

The results obtained through the application of the DEA models (Data envelopment Analysis), show that the more efficient offices are located mainly in the North and in the Centre of Italy; such results are true both analysing the whole civil sector, the civil litigation and the insolvency procedures sub sectors. Furthermore the increase of the turnover corresponds to the increase of the proceedings duration.

The beta regression models show that the proceeding duration and the geographical distribution affect directly and significantly the first instance judicial efficiency; the decrease of the proceedings duration corresponds to the increase of the efficiency measured through the DEA, furthermore the offices in the North have a significant higher performance if compared to the offices in the Centre and in the South.

The probability of being efficient is lower for the offices with not adequate planning capacity in the 2017, if compared to the offices with an adequate target planning capacity; furthermore the dispersion parameter and thus the variability of the distribution is affected by the kind of Court, district or territorial Court. In the whole civil sector, the increase of the working staff, has not had a significant effect on the measure of the efficiency of the judicial offices yet, but it has contributed to improve the capacity of planning adequately a disposal of the ancient pending cases: the probability of having an adequate target planning capacity is of almost 2,3 times higher in the Courts where has been approved an increase of the working staff if compared to the courts where there has not been any variation. It is clear as the offices knew how to govern the working staff variation, both in the increase, reaching positive prefixed results, and in decrease, not deviating so much from the prefixed goals. The positive effects reflect the adequacy of the modification of the working staff determined by the Ministry of Justice and by the Superior Council of Judiciary. In the sector of Civil litigation (SICID), the beta regression shows similar results as far as the geographical distribution is concerned: furthermore the offices with an under target planning capacity (prudent offices) in the last year have a higher probability of being efficient if compared to the offices with an adequate target capacity.

Furthermore the offices where there has been a decrease in the working staff units have an estimated probability of being efficient lower than Courts where there has been no variation in terms of increasing or decreasing of units. Despite it is not simple to analyse the different reality characterising the organisation of the justice in the Italian Courts, and also the variables which determine directly or indirectly the organisation of the same Courts, this study provides some points of interest. Among these, it has to be underlined the moderate but evident positive impact of the activities taken by the Csm in order to improve the efficiency of the Courts of first instance. In particular, there are evidences of improvement in structuring the management programmes, in the planning capacity of the disposal objectives and in the internal organisation after the working staff variation.

References

Agresti, A. 2002. *Categorical Data Analysis*. Hoboken, NJ, U.S.: John Wiley & Sons.

Cribari-Neto, F., and A. Zeileis. 2010. "Beta Regression in R". 2010. Journal of Statistical Software, Volume 34, Issue 2: 1-24.

Dudley, W.N., R. Wickham, and N. Combs. 2016. "An Introduction to Survival Statistics: Kaplan-Meier Analysis". *Journal of the Advanced Practitioner in Oncology*, Volume 7, N. 1: 91-100.

Landi, L., and C. Pollastri (*a cura di*). 2016. "L'efficienza della giustizia civile e la performance economica". *Focus Tematico* n. 5. Roma, Italy: Ufficio Parlamentare di Bilancio - upB.

McCullagh, P. 1980. "Regression Models for Ordinal Data". *Journal of the Royal Statistical Society*, Series B (*Methodological*), Volume 42, N. 2: 109-142.

Paradi, J.C., H.D. Sherman, and F.K. Tam. 2017. "DEA Models Overview". In *Data Envelopment Analysis in the Financial Services Industry. International Series in Operations Research & Management Science*, Volume 266: 3-39. Cham, Switzerland: Springer Nature.

Re-design project of the Istat consumer price survey: use of probability samples of scanner data for the calculus of price indices

Antonella Bernardini, Maria Cristina Casciano, Claudia De Vitiis, Alessio Guandalini, Francesca Inglese, Giovanni Seri, Marco Dionisio Terribili, Francesca Tiero [1]

Abstract

The availability of scanner data (SD) from the retail modern distribution (food and grocery) is the starting point for the implementation of the innovation in the consumer price survey (CPS) improving and unburdening the data collection phase, together with the progressive introduction of more rigorous probabilistic sampling procedures for the selection of outlets and products (series). This article presents the work carried out by the statistical-methodological working group for the revision of the sample design of the consumer price survey in the light of the new data sources (SD). The experiments of probabilistic selection schemes of the series are developed in two main phases that reflect the choice made by Istat to make a gradual transition from a fixed approach to a dynamic approach in the calculation of the consumer price index (CPI). In 2018, Istat started using the SD to compute the CPI.

Keywords: scanner data, selection schemes, price indices, fixed and dynamic approach.

1 Antonella Bernardini (anbernar@istat.it); Maria Cristina Casciano (casciano@istat.it); Claudia De Vitiis (devitiis@istat.it); Alessio Guandalini (alessio.guandalini@istat.it); Francesca Inglese (fringles@istat.it); Giovanni Seri (seri@istat.it); Marco Dioniso Terribili (terribili@istat.it); Francesca Tiero (tiero@istat.it), Italian National Institute of Statistics – Istat.

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.

1. Introduction

The Consumer Price Survey (CPS) has undergone for some years a review aiming at redesigning different aspects of the data collection procedures and sampling methods. The aim of the project is to introduce progressively in the CPS more rigorous sampling procedures, probabilistic where possible, starting from the selection of outlet and products for the sectors where this is feasible.

The availability of scanner data (SD) from the retail modern distribution (food and grocery) is the starting point for the implementation of the innovation in the Consumer Price Survey. At present food and grocery sector cover 11% of the total and modern distribution the 55% of the total. Scanner data is a big opportunity to introduce improvements in the CPS both for the data collection and the sampling perspective.

In recent years Istat, through a contract with Nielsen and an agreement with the six main retail chains operating in Italy, started, at the end of the year 2014, receiving SD referred to food and grocery market and processing them to experiment the calculus of Consumer Price Index (CPI). This acquisition places Italy among countries using or testing the use of this source of data for compiling CPI. As noted some years ago in the ILO CPI Manual (ILO *et al.*, 2004), "Scanner data constitute a rapidly expanding source of data with considerable potential for CPI purposes" (p. 54); "Scanner data obtained from electronic points of sale include quantities sold and the corresponding value aggregates on a very detailed level" (p. 92); "Scanner data are up to date and comprehensive" (p. 478).

After an experimental phase, since 2018 Istat has started using the SD concerning grocery products (processed food, products for house maintenance and personal care) to compute the consumer price index (CPI). SD are regularly provided to Istat through the market research company ACNielsen, for the main chains present in Italy and a sample of about 2,000 outlets deployed all over the country. For 2018 the compilation of the CPI using scanner data is based on a fixed basket perspective: the use of these data has concerned some channels of the modern distribution, in particular hypermarkets and supermarkets of the main retail chains operating in Italy. Since 2020 Istat has extended the use of scanner data to other channels of the modern distribution (discounts, small sales areas and specialist drug) and has realised the transition

to a flexible basket approach. At the moment, the traditional retail distribution data continue to be collected by the current survey on the field. This choice is determined by the peculiarities of the Italian retail distribution with respect to other countries: the traditional distribution, in fact, is still relevant in many geographical areas of the country.

This paper aims to present the work carried out by the statistical-methodological group for the revision of the sampling design of the consumer price survey in the light of the new data sources (SD). The work is divided into two main lines of research, the first concerning the analyses of the scanner data acquired from 2014 by Istat for the experimental phase, the second involving the study of selection schemes of series (references individuated by EAN and outlet codes) from SD. In the paper, the focus is both on the results of the analysis conducted on the big data sets containing SD and on the experiments in the use of samples from scanner data for the calculus of elementary indices for homogeneous product aggregates (consumption segments, markets).

The analyses conducted on scanner data range from the aggregate level, *i.e.* statistical distributions of turnover by chain and province, to the elementary level, *i.e.* continuity and seasonality of the products and data quality. These analyses have been possible thanks to the information contained in SD. Scanner data files, indeed, contain elementary information referred to single EAN codes (GTIN - Global Trade Item Number, or EAN - European Article Number) for specific outlets consisting of turnover and quantities sold during a week. This information does not provide the "shelf price" of the product individuated by the EAN code and outlet (references or series) but allows us to define a unit value or average weekly price.

For SD experimental framework, firstly, probability and nonprobability selection schemes of series are compared and then different probability sampling designs are examined. The experiments are conducted to test different methods for selecting references by comparing the estimates of different elementary index formulas with the value of the corresponding price indices obtained from the whole set of references for the same elementary aggregate. Sampling designs and price index formulas are studied through a Monte Carlo simulation by first selecting 500 samples for each different sampling design and then calculating variability and bias on the estimated indices in the replicated samples.

Moreover, through a further experiment, the differences between a fixed and a dynamic population approach in the construction of the elementary price indices are highlighted. In this case, the purpose is trying to measure the magnitude of sampling and non-sampling errors for different price index formulas in both approaches. These errors are generated by different causes, among the main ones: disappearing products, ignoring entries of new products and temporary missing products, when a static population is assumed and a fixed basket is used; variability in the number of matched-items between the months and the chain drift of weighted price indices when a dynamic population is considered and a flexible basket is used.

In both cases the experiments were conducted starting from a sampling frame represented by a panel data set that contains permanent series (references), which refers to those series with not-null turnover for at least one relevant week (the first three full weeks) in each month of the considered year, starting from the December of the previous year. The use of data referred to the full weeks (first two, or first three) for each month is advised in the recommendation drafted by Eurostat. This is a restriction inherent to the observable weeks that arises from reasons deriving from operational constraints of the productive process.

In the paper, the use of scanner data in the consumer price index production (Section 2) and the Istat project of redesigning the CPI are presented (Section 3). Then, in Sections 4 and 5, the processing phases and the analyses conducted on scanner data are shown, while in Section 6 the experimental framework for sampling from SD is described. Sections 7 and 8 present a detailed description of the experiments and the main results. Finally, Section 9 describes the sampling design adopted by Istat from 2018 to 2020 to select the sample of outlets from the SD of grocery products and how it came to its definition. Conclusions are reported in Section 10.

2. Scanner data in the Consumer Price Index production

2.1 Practice in the European countries

In a study perspective scanner data from retail stores allows researchers to evaluate how different price index formulas perform at the elementary level. In fact, official CPI is usually constructed in two broad steps. First, elementary price indices are calculated for narrowly defined and relatively homogeneous products, known as elementary aggregates. In a second step, these elementary indices are aggregated into a single consumer price index using expenditure weights. Elementary indices, named also higher-level elementary indices, are therefore the building blocks of price index numbers.

While the aggregation at a higher level is carried out using generally Laspeyres type formulas with weights deriving from national account or expenditure survey data, official practices in elementary price index construction are still not uniform across countries, deserving further investigation in the consequences of different choices (Gábor and Vermeulen, 2014).

The launch of barcode scanner technology has enabled retailers to capture detailed information on transactions at the point of sale. Scanner data is high in volume and contains information about individual transactions or summaries, date, quantities and values of products sold, and product descriptions. As such it is a rich data source for NSIs that can be used both to improve their statistics and to reduce statistical burden and costs.

Scanner data will be increasingly available to statistical agencies and consequently, new methods are needed to work with this new data source. In Europe, countries are rapidly expanding the use of SD for the compilation of the Consumer Price Index (CPI). Norway was the first country using scanner data in regular CPI production (2001) followed by the Netherlands. Other countries have also started using scanner data, for instance, Sweden, Switzerland, Belgium, and Denmark. Scanner data was introduced in the regular production of the CPI from January 2018 in Luxemburg and Italy. From 2016, other NSIs (as CSO of Ireland) have started projects to obtain and analyse scanner data from retailers to research its potential use in the production of statistics and the CPI in particular.

Scanner data can be exploited in different ways. The simplest way is using SD as an alternative source for price collection, replacing collection within the stores, without changing the traditional principles of computing the price indices. This method was applied by the Swiss Federal Statistical Office (Vermeulen and Herren, 2006). Alternatively, as done in Norway and Sweden, SD can be used as the universe from which samples of references can be selected following different methods (Nygaard, 2010; Norberg, 2014).

Finally, all (or almost all) SD can be used to compile price indices, without a strict sample selection, but with consequences on the theoretical definition of the index. In Belgium and Netherlands, the computation method is different and the data are used in a more extensive way to calculate price indices (van der Grient and de Haan, 2010). The method used assumes a dynamic population approach: elementary price indices of homogeneous items are calculated by monthly chained unweighted geometric index (Jevons); no explicit weighting is applied and expenditure information is used just to select a cut-off sample of matched items during two months in a row.

2.2 Impact on Consumer Price Index compilation

Scanner data introduce important advantages compared to data collected through traditional survey. In particular, the availability of turnover and quantity data at the item level offers a real possibility of calculating more accurate indices: it is possible, in fact, to include in the calculus the expenditure share of each product sold. SD also contains descriptive information about items characteristics useful to treat quality change, to identify relaunches of existing products or new products, etc. (Feldmann, 2015; Chessa *et al.*, 2017).

On the other hand, the use of SD in the compilation of CPI must take into account some important drawbacks, as attrition of products, temporary missing products, entry of new products and volatility of the prices and quantities due mainly to sales. These are aspects that need to be addressed from both a theoretical and a practical point of view (de Haan *et al.*, 2016).

To maximise the potential offered by SD it would be necessary to go beyond those methods of price index compilation which do not exploit all the information provided by the data and do not take into account the population dynamics (Chessa *et al.*, 2017). Weighted and chained indices should be

considered to incorporate the overall price trend over a given time, including the prices of new products. Furthermore, the problem of shrinkage over time due to the attrition of a fixed basket of products is solved automatically using chain indices. However, even though in a dynamic approach it is necessary to construct series of chained indices, high-frequency chaining of weighted indices (also superlative Fisher and Törnqvist indices) are affected by chain drift, due to non-symmetric effects on quantities sold and expenditure share of goods before and after the sale (Ivancic *et al.*, 2011; de Haan and van der Grient, 2011).

In recent years, an important debate has taken place among the researchers dealing with the estimate of the consumer price index starting from SD. The focus, above all, has been on the transition from a static population approach (fixed basket) to a dynamic population approach (flexible basket) and it is based on the study of alternative price index formulas based on matched-model methods (matching of products sold during two months in a row) or other methods that are transitive and, therefore, free from chain drift (de Haan *et al.*, 2016).

Other aspects discussed are the quality of SD (completeness and correctness) and the definition of methods to treat appearing and disappearing products, temporary missing products, relaunches, quality change, etc. (Vermeulen and Herren, 2006; van der Grient and de Haan, 2010).

3. Re-Design of the Italian Consumer Price Survey

The aims of the re-design of the Italian Consumer Price Survey are to be ascribed to a reduction of the weight of the traditional data collection in the field to around 50% through expanding the use of the Internet as a data source, widening the use of web scraping techniques and using scanner data as a new source. Scanner data are a great opportunity for introducing probability sampling designs: the selection of elementary items from scanner data allows to implement a sufficiently feasible field procedure and to overcome the potential source of bias of the procedure adopted in the current survey, based on subjective choices[2].

To introduce the use of the SD in the compilation of the CPI, Istat's contract with Nielsen initially provided for the supply of weekly data of turnover and quantities at EAN code (elementary item) and outlet level for six modern retail distribution chains operating in the food and grocery market in 37 Italian provinces (coverage of 55% of the Italian population). For the experimental phase, Nielsen provides backward data for at least one full year and the preceding month of December, starting from December 2013, 2014 or 2015 (depending on the starting point of delivery of each province).

During 2014 and 2015 the scanner data were provided by Nielsen gradually, first the data relating to five provinces, to which were added then 14 and 18 other provinces up to a total of 37 provinces. The release of scanner data by Nielsen follows an informal agreement between Istat and the Association of Modern Distribution, representing the main chains of modern retail trade. The requests of data (with formal and legal meaning) sent by Istat concerned weekly data of turnover and quantities, EAN code, outlet code and others information, for a progressively increasing amount of provinces, for the six

2 The traditional Consumer Price Survey (CPS) carried out at territorial level is based on three purposive sampling stages. The sampling units are respectively the municipalities, the outlets and the elementary items for which the prices are collected. The biggest municipalities are forced by law to participate to the survey. The Municipal Offices of Statistics select the outlets sample, where the prices of a fixed basket of products (including roughly 1,000 products) are collected. The outlets sample is chosen to be representative of the consumer behaviour in the municipality. For each product of the basket the most sold item is selected and the prices of these items are collected throughout the year. At the end of each year Istat refreshes the fixed basket of products and, at the same time, the sample of outlets and elementary items is updated. The elementary price indices are currently obtained at municipality level by unweighted geometric mean. The general price index is calculated by subsequent aggregation of elementary indices, using weights at different levels based on population and national account data on consumer expenditure.

"big chains" (Coop Italia, Conad, Selex, Esselunga, Auchan, Carrefour) covering almost 57% of the total turnover of modern distribution.

Moreover, Nielsen provided the dictionary for the classification of EAN code sold in Italy attributes that allow you to identify the product (manufacturer, brand, possible sub-brand, size, packaging, variety) to GS1-ECR-Indicod product classification (variation of GPC Global Product Classification applies worldwide). Istat ensures internally the translation from ECR to COICOP, the classification of products used for the CPI. Consumption segments, not foreseen by the EU-COICOP, are the most detailed domain of estimate for Italian CPI, constitute groupings of homogeneous products; those defined for the food and grocery are 126 out of a total of 324.

4. Processing phases of scanner data

4.1 Outline

The analyses on scanner data quality constituted an important activity of the statistical-methodological workgroup aimed at guaranteeing the completeness and correctness of the acquired data. Completeness and correctness of the data are two important pillars for the correct use of scanner data (SD) and the compilation of an accurate modelling of the price index over time.

These analyses are part of the processing phase of the collected data in which the quality checks must be performed and the data cleaning must be made.

Indeed, the use of SD implies the definition and the implementation of different checks in both data acquisition and processing phases. The scanner data collection requires the use of formal checks on the flow of collected data but also the development of checks on the quality of the data. The formal checks must be defined to ensure the completeness of the data collected at the provincial level, distribution chains, outlets, products and weeks. The quality checks on loaded data are required to introduce editing rules which identify inadmissible values on the variables of interest as quantities sold, turnover and prices (Rais, 2008; Saidi and Rubin Bleuer, 2010).

As the scanner data file do not provide the "shelf price" of the product individuated by the EAN code and outlet (references or series), the unit prices are defined as a unit value or average weekly price of series starting from the quantities sold and the turnover.

4.2 Formal and quality checks

The continuous flow of acquired scanner data by Nielsen is subject to formal checks both during and after data loading.

The formal checks during the loading of data affecting the presence of a full numeric code for the outlets and products, and the presence of numeric and valid decimal values for the turnover and quantities. Another important check affects the week in which the data must refer to.

The formal checks of the post-loading concern:

i) the presence of duplicates for outlet, reference and week;

ii) the absence of an outlet or a product in the list updated every six months in the first case and every two months in the second case;

iii) the presence of null fields of turnover and quantities;

iv) the presence of unauthorised data such as outlets that do not belong to the authorised provinces or the allowed chains, or outlets not classified as a supermarket or hypermarket.

These checks must ensure that data always refer to the same population (provinces, chains, outlets and products) for each week.

The data quality check follows the phase of formal checks with the aim of introducing editing rules that identify inadmissible values among the variables of interest (quantities sold, turnover and unit prices). First quality checks are implemented to identify and eliminate the problematic occurrences (outlet, EAN code, week) in which:

v) quantity<1 not motivated by unit of measurement;

vi) decimal values on quantities>1 not motivated by unit of measurement;

vii) unit prices <= 0.01 €.

Subsequently, in order to maintain an accurate price index over time, the product prices have been validated considering the data acquired, during the relevant weeks of every month, at the provincial level.

In the following schemes, the formal checks developed during and after the upload are synthesised.

Scheme 1 - Formal checks during the upload

Check	Error type
ID outlet	Not null and numeric
ID product	Not null and numeric
Turnover	Numeric
Quantity (sold packages)	Numeric
Year	Current or previous year
ID week	Not ≥ 53

Scheme 2 - Formal checks after the upload

Check	Error type
Duplication	1 - Duplicated outlets, products or weeks
Outlet	2 - Outlet not in the frame of outlets (updated every 6 months)
Product	3 - Product not in the frame of products (updated every 2 months)
Missing turnover/quantity	4 - Turnover and/or quantity NULL
Quantity	51 - quantity <1 52 - values with decimals
Prices	61 - prices ≤0.01 €
Completeness	71 - Data of the authorised weeks 72 - Data from authorised provinces 73 - Data from authorised chains 74 - Type of outlets (only Hypermarket and Supermarket)

4.3 Identification of inadmissible unit prices

To identify inadmissible unit prices, the price of an occurrence (product*outlet*week) too high or too low concerning an interval built on a synthetic measure of the distribution of the prices of the product sold in the province in the month, several methods have been tested. For the sake of simplicity, also in terms of computational burden, the choice has been reduced between the two methods. Both methods are based on the computation of the median unit prices, considering the quantities sold for each single occurrence (EAN code, week, outlet).

The first method consists of a fixed trimming method and the tolerance interval of prices is:

$$\left(\frac{Median_w}{K_1}, K_2 * Median_w \right); \quad (1)$$

the second can be named moving trimming and the related tolerance interval is:

$$\left(Median_w - \frac{K_1 - 1}{K_1} \frac{Median_w}{log_{10}(Median_w + 10)}, Median_w + (K_2 - 1) \frac{Median_w}{log_{10}(Median_w + 10)} \right). \quad (2)$$

Trimming depends on the values assigned to K_1 and K_2. Assigning values to K_1 and K_2 requires making assumptions on maximum discount and maximum price hike with respect to the median price of the product allowed. For food and groceries, $K_1=5$ (it means that up to 80% discount is allowed) and $K_2=3$ (it means that up to 3 fold increase with respect to the median price of the product is allowed) seems to be plausible values.

However, while in fixed trimming the relative width of the tolerance interval remains unchanged, with moving trimming it narrows as the median price of the product increases (Figure 1 and Table 1).

Figure 1 - Tolerance interval limits of (prices/Median$_w$) with fixed and moving trimming

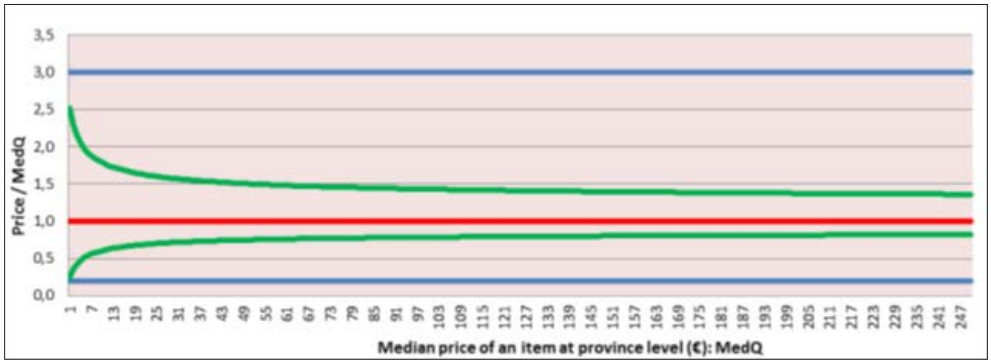


Table 1 - Lower (LB) and Upper (UP) bounds calculated with two trimming methods on a Median price

Median price	Fixed trimming		Moving trimming	
	LB	UP	LB	UP
1	0.2	3	0.246	2.523
2	0.4	6	0.362	2.289
3	0.6	9	0.432	2.147
4	0.8	12	0.48	2.05
5	1	15	0.516	1.979
6	1.2	18	0.543	1.924
7	1.4	21	0.565	1.88
8	1.6	24	0.583	1.843
9	1.8	27	0.598	1.813
10	2	30	0.611	1.786
20	4	60	0.683	1.64
25	5	75	0.702	1.602
50	10	150	0.75	1.504
75	15	225	0.773	1.459
90	18	270	0.781	1.442
100	20	300	0.786	1.432
120	24	360	0.794	1.416
200	40	600	0.814	1.377

The moving trimming is preferable to the fixed one because it narrows down, thanks to the log function, the extremes as the median price of the product increases.

4.4 Removal of inadmissible unit prices effect

In this paragraph, some results obtained from the removal of the occurrences identified with the moving trimming method are presented.

The analysis reported here was conducted on all occurrences by chains available (Conad, Coop, Esselunga, Auchan, Carrefour, Selex) in the years 2013 and 2014, in five Italian provinces. The analysis aims to highlight the impact that the removal of inadmissible unit prices has in terms of the number of occurrences, the number of quantities sold and turnover, and to detect the consumption segments most affected by the deletion of occurrences.

Table 2 presents the removal of inadmissible unit prices effect on the total amount of occurrences, quantities sold and turnover.

Table 2 - Total amount of occurrences, quantities sold and turnover – percentage of references, quantities sold and turnover removed, by year and province

Province	Year	Total amount of			Removal of inadmissible unit prices effect in terms of		
		Occurences	Quantities sold	Turnover	Occurences (%)	Quantities sold (%)	Turnover (%)
Ancona	2013	16,314,683	179,261,691	302,928,869	0.042	0.027	0.024
	2014	16,972,117	183,593,696	309,328,337	0.033	0.021	0.019
Cagliari	2013	11,236,941	145,630,320	256,859,395	0.044	0.027	0.028
	2014	11,383,752	145,612,723	253,942,082	0.044	0.032	0.030
Palermo	2013	10,671,436	139,756,595	222,335,145	0.064	0.030	0.026
	2014	12,094,184	152,094,573	240,512,170	0.098	0.033	0.030
Piacenza	2013	6,474,872	93,362,079	167,777,673	0.021	0.009	0.017
	2014	7,521,222	100,610,744	180,572,592	0.030	0.020	0.027
Torino	2013	45,458,148	671,423,702	1,242,080,444	0.048	0.017	0.023
	2014	48,621,536	679,459,389	1,248,849,275	0.062	0.026	0.026

Generally, the number of removed occurrences is very low (Table 2) in all provinces but not constant in the two years under review, except the province of Cagliari in which the deleted occurrences is 0.044 percent in both years.

The turnover share lost due to the elimination of occurrences is very contained, in fact, is never more than 0.030 percent.

By analysing the problem for a single consumption segment, the situation changes as some consumption segments lost 0.05 percent of turnover.

In scheme 3 are listed, for each province and each year, the consumption segments that suffer a greater loss of turnover share following the removal of inadmissible unit prices related to specific occurrences.

Scheme 3 - Consumption segments (COICOP-6digit) with percentage of turnover removed >0.05%

ANCONA	2013	Mineral water; Other cereal-based products; Non-alcoholic beer, or beer with low alcoholic content; Cured cheese; Dried fruit; Berries; Ice cream; Perfumes and make up products; Electric shavers, trimmers and other electric grooming products; Rice
	2014	Mineral water; Other meat; Other cereal-based products; Non-alcoholic beer, or beer with low alcoholic content; Body, hand and hair lotions; Ready meals with ground meat; Perfumes and make up products; Electric shavers, trimmers and other electric grooming products
CAGLIARI	2013	Mineral water; Other beauty products; Other medical products; Other products for pets; Body, hand and hair lotions; Perfumes and make up products; Electric shavers, trimmers and other electric grooming products; Salt, spices and aromatic herbs; Sugar
	2014	Mineral water; Frozen seafood; Body, hand and hair lotions; Dried, smoked or salted fish or seafood; Perfumes and make up products; Electric shavers, trimmers and other electric grooming products; Pregnancy tests and contraceptives; Sugar
PALERMO	2013	Other alcoholic beverages; Other disposable items for the home; Other perishable items for the home.; Other preserved fish or seafood; Other beauty products; Other products for pets; Detergents and house cleaning products; Nuts; Dried fruit; Ice cream; Body, hand and hair lotions; Fresh pastry; Perfumes and make up products; Electric shavers, trimmers and other electric grooming products; Rice; Brushes, brooms, wipes and sponges; Dried vegetables; Sugar
	2014	Mineral water; Other beauty products; Other house cleaning and upkeep products; Non-electric appliances; Soaps and personal hygiene products; Detergents and house cleaning products; Dried fruit; Body, hand and hair lotions; Packaged bread; Frozen fish; Pizza and quiche; Fresh pastry; Perfumes and make up products; Electric shavers, trimmers and other electric grooming products; Rice; Brushes, brooms, wipes and sponges; Gourmet wines; Sugar
PIACENZA	2013	Fresh Milk
	2014	Mineral water; Non-alcoholic beer, or beer with low alcoholic content; Fresh milk; Paper-based kitchen products
TORINO	2013	Natural meat and mixed-meat hamburgers; Perfumes and make up products; Electric shavers and trimmers; Mineral water; Dried, smoked or salted fish or seafood; Other meat; Dried vegetables; Body, hand and hair lotions; Other house cleaning and upkeep products; Non-electric appliances; Pregnancy tests and contraceptives
	2014	Mineral water; Other products for pets; Other house cleaning and upkeep products; Non-electric appliances; Body, hand and hair lotions; Dried, smoked or salted fish or seafood; Perfumes and make up products; Electric shavers, trimmers and other electric grooming products; Brushes, brooms, wipes and sponges; Alcoholic beverages; Pregnancy tests and contraceptives

From the scheme above it is clear that the consumption segments like *mineral water, perfumes* and *razors* are more problematic than other ones, in fact, in this segment, a greater number of references with inadmissible unit prices can be identified.

4.4 Moving trimming method

In the moving trimming method, as said above, the relative allowed range shrinks contextually with the increase of the median price of the product (Figure 1). In the graphs below (Figures 2-6), in which the density functions of the unit prices of some specific products are represented, the tolerance intervals are delimited by two vertical red lines, while the green line and the dotted green line indicate respectively the weighted and unweighted median price of the product. The values of inadmissible unit prices are indicated by red dots; the green dots indicate the presence of discounts.

Figure 2 - Density function of the unit prices of product "Powdered milk for babies Danone"

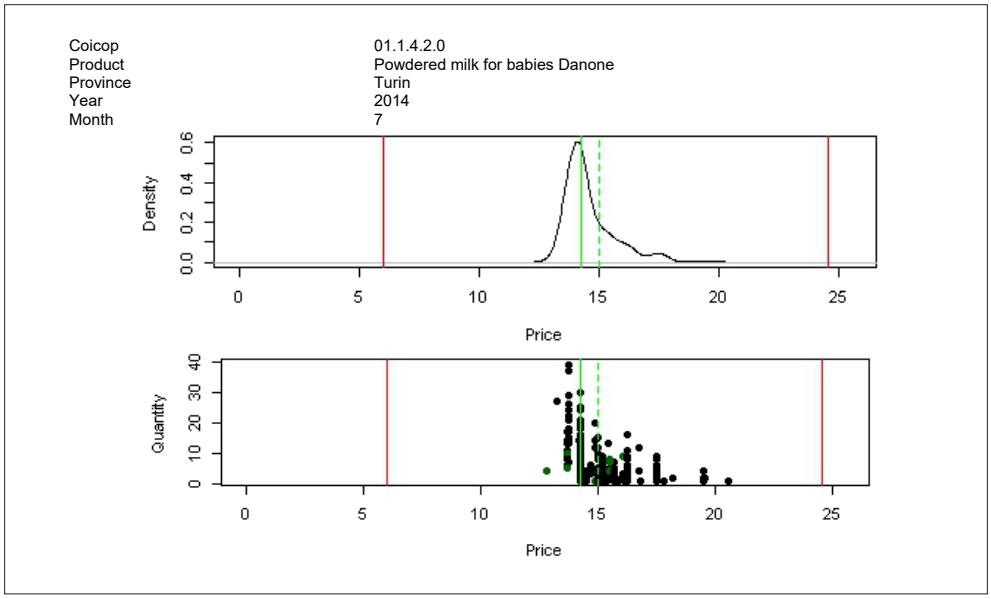


Figure 3 - Density function of the unit prices of product "Aged rum, Havana Club, 700 ml"

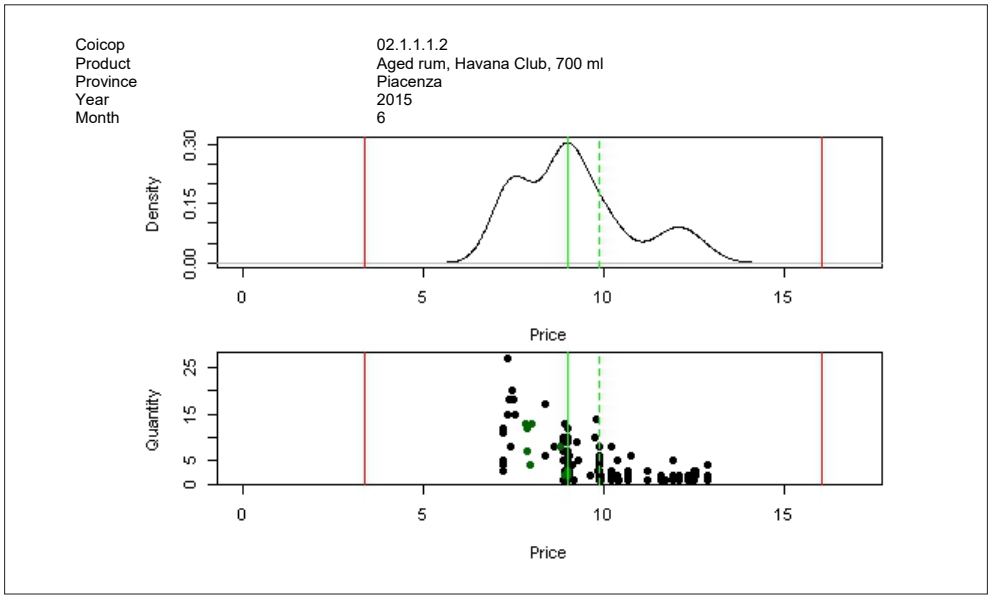


Figure 4 - Density function of the unit prices of product "Rice, Private Label, 1000 gr"

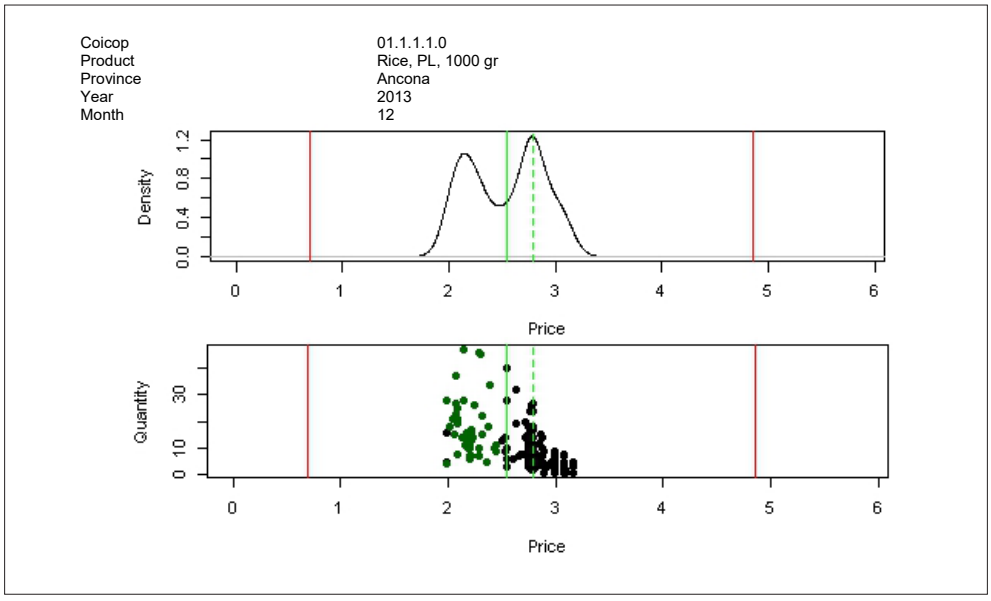


Figure 5 - Density function of the unit prices of product "Sparkling water, Rocchetta, 1.5 lt"

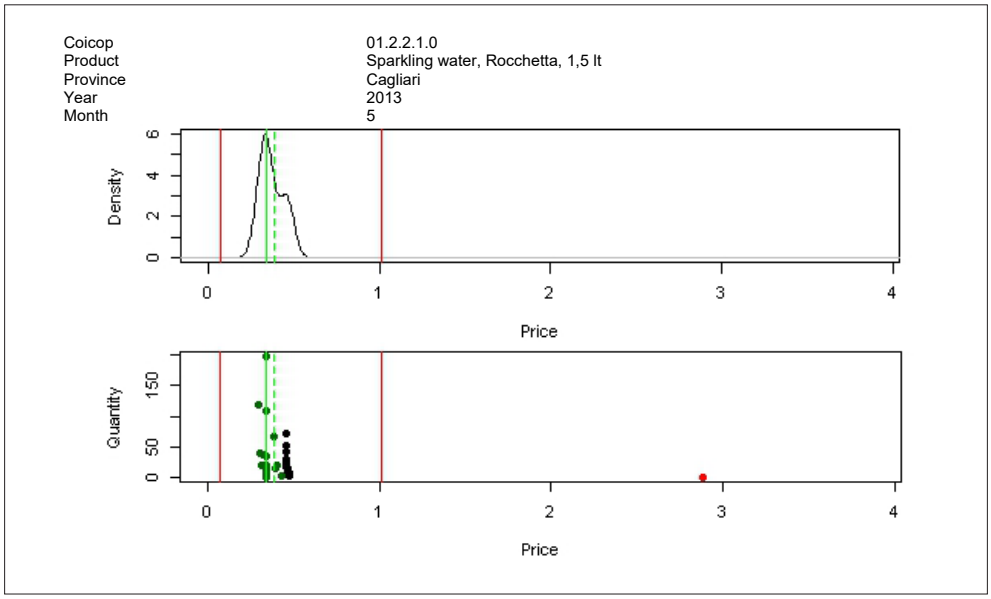
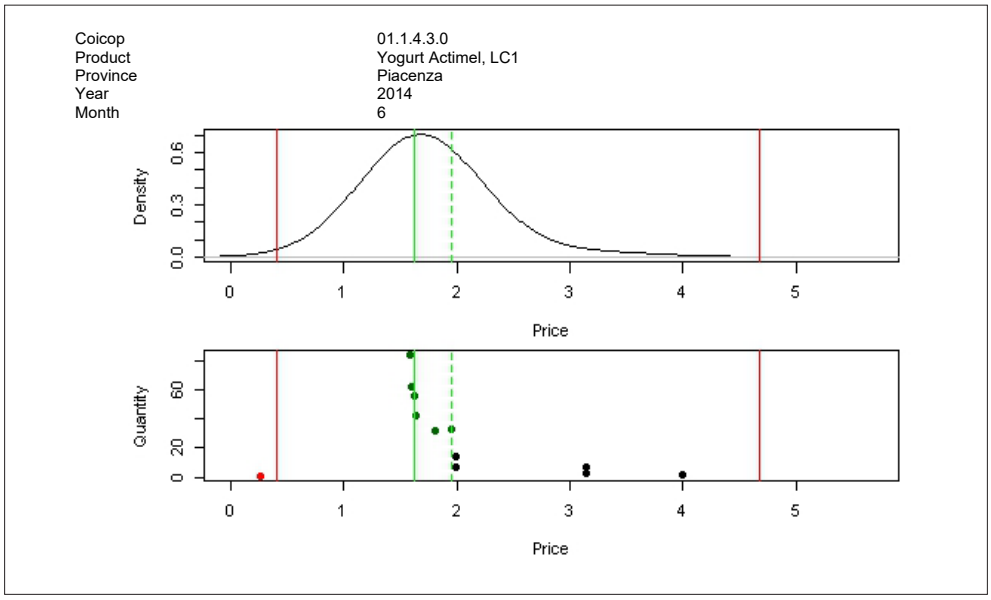


Figure 6 - Density function of the unit prices of product "Yogurt Actimel, LC1"



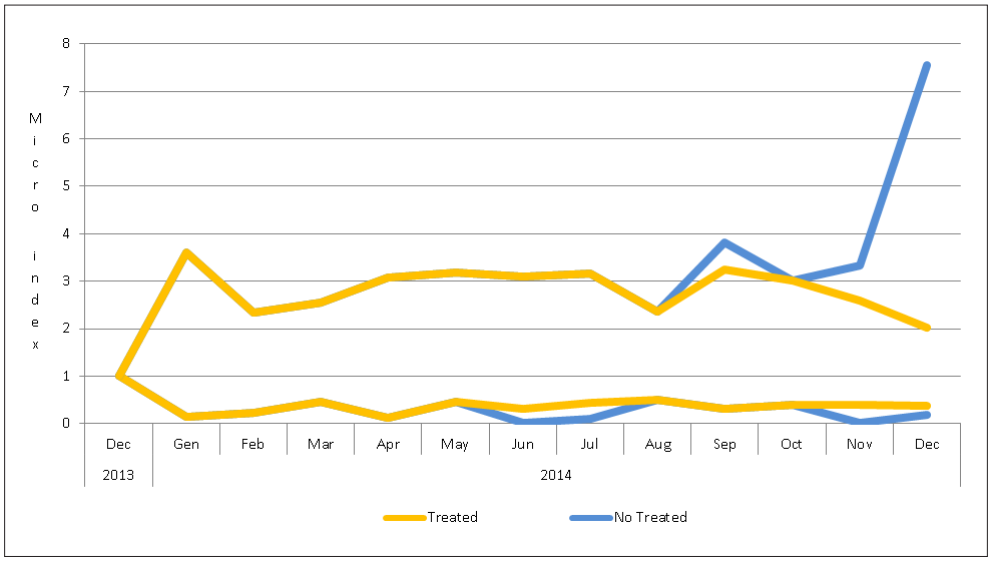
The method does not reveal any inadmissible value in unit prices in graphs 2, 3 and 4, while in graphs 5 and 6 identify the inadmissible values respectively beyond the upper and lower limit of tolerance interval.

4.5 Impact of data cleaning on micro-indices

For an assessment operated by the data cleaning, an analysis on micro-indices - the ratio between the price of the occurrence at time t and the price of the same occurrence at the base time - has been performed. In particular, the micro-indices trend has been observed in the twelve months of 2014.

Figure 7 compares the maximum and minimum value of micro-indices calculated before and after the phase of the data processing in the consumption segment wine of the Torino province.

Figure 7 - Micro-indices trend before and after the phase of the data processing (2014)



The figure above shows that in some months the impact of data cleaning can be quite important, besides its conceptual correctness: in the examined case the difference between the maximum values of the micro-indices of treated and untreated data, in December, is remarkable.

5. Analyses on scanner data

5.1 Framework

The analysis of scanner data has constituted an important line of research conducted by the statistical-methodological workgroup especially in the initial phase of the project. The realised analyses had different objectives:

a) to study the chain and outlet type (hypermarkets and supermarkets) distributions in the first provinces acquired in terms of turnover;

b) to evaluate the attrition problem related to the life cycle of the EAN code and series (or references);

c) to study the continuity of the presence of the EAN codes and temporary availability (such as seasonal, new entry, temporary or definitive absence);

d) to identify seasonal products (products sold just at certain times of the year, following a seasonal pattern) and to study seasonality in consumer prices caused by a variety of influences, some on the demand side and some on the supply side.

The analyses conducted on the big data sets acquired by Istat and analysed herein, cover five Italian provinces (Ancona, Cagliari, Palermo, Piacenza, Torino) and six chains of modern distribution (Conad, Coop, Esselunga, Auchan, Carrefour, Selex). The analyses are focussed on the observed series (EAN+outlet code) belonging to the relevant weeks (the first three full weeks in each month) of each month and on permanent series (panel series SD), as defined above.

The analyses on continuity and seasonality of the products were conducted only in Torino province.

5.2 Distribution of turnover and outlet

In the following analysis, some aspects of the six chains of modern distribution (Conad, Coop, Esselunga, Auchan, Carrefour, Selex) and outlet type (hypermarket and supermarket) distributions in the first five provinces acquired by Istat are highlighted for the year 2014.

Table 3 - Total turnover and number of outlets by chain and province (2014)

Chain		Province					
		Ancona	Cagliari	Palermo	Piacenza	Torino	Total
Conad	Turnover	14,287,546	63,553,952	56,233,579	38,883,578	47,924,071	220,882,726
	Outlet	3	10	15	7	13	48
Coop	Turnover	79,792,735	-	35,976,469	32,823,101	272,443,053	421,035,357
	Outlet	10	-	2	4	26	42
Esselunga	Turnover	-	-	-	58,935,833	108,143,835	167,079,668
	Outlet	-	-	-	2	3	5
Auchan	Turnover	106,363,959	54,247,839	64,424,104	4,890,904	157,247,092	387,173,897
	Outlet	11	2	7	1	8	29
Carrefour	Turnover	11,093,852	38,062,177	64,896,292	2,659,154	424,278,376	540,989,851
	Outlet	1	2	21	1	40	65
Selex	Turnover	74,711,322	77,908,712	-	28,197,290	140,878,602	321,695,925
	Outlet	34	21	-	4	39	98
Total	Turnover	286,249,412	233,772,680	222,097,542	166,389,859	1,152,108,585	2,058,857,425
	Outlet	59	35	46	19	130	289
Coverage turnover		87.00	74.28	69.95	73.68	72.65	73.96

The analysis has been carried on the whole of the 289 outlets of the provinces during the 52 weeks of the year 2014. Table 3 contains the whole turnover and the number of outlets by chain and province and the percent coverage of the six chains with respect to the total turnover of modern distribution for food and grocery at the province level.

The table above shows a heterogeneous situation both among the chains and the provinces: Torino province represents more than 50% of turnover involved; this remark could be influenced by the high number of outlets observed in this province, 130 on a whole set of 289. In the other provinces the number of outlets varies from 19 (Piacenza) to 59 (Ancona), with a turnover between 166 and 286 million euro.

The last row shows the good level of coverage of the six chains, although with a certain heterogeneity among provinces: the coverage is generally close to 72%, with a maximum level of 87% assessed in Ancona province.

In Table 4, likewise above, turnover and number of outlets are reported considering the six chains and two outlet types. This variable is particularly important because connected both with chain (which can have a higher/lower propensity to set up a hypermarket/supermarket) and with the number of sold elementary items, much higher in the hypermarkets than in the supermarkets.

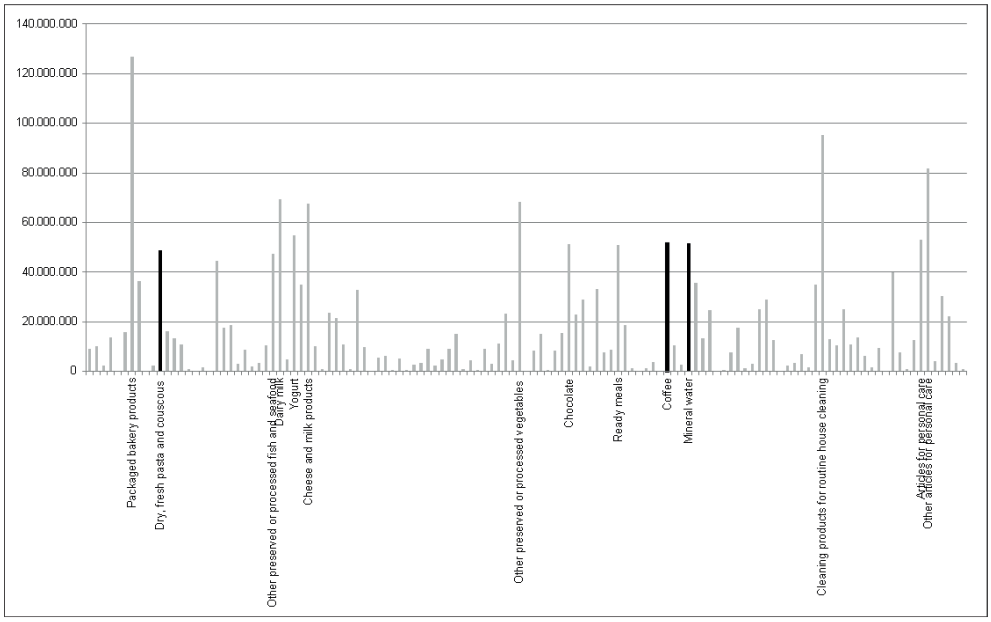
Table 4 shows a greater hypermarkets turnover than supermarkets one, although the number of outlets is reasonably lower (53 hypermarkets and 234 supermarkets, for a total of 289 outlets). Moreover, hypermarkets are more spread in the Torino province (28 outlets) with respect to Palermo and Piacenza provinces (4 outlets). This result, however, has to be linked to the distribution of the local chain, in fact, Esselunga and Carrefour chains have a greater number of hypermarkets than the other chains.

Table 4 - Total turnover, number of outlets and number of items by province and outlet type (2014)

Province		Outlet type		Total
		Hypermarket	Supermarket	
Ancona	Turnover	136,000,290	150,249,122	286,249,412
	Outlet	10	49	59
	Item	63,112	43,000	106,112
Cagliari	Turnover	126,001,578	107,771,102	233,772,680
	Outlet	7	28	35
	Item	53,053	26,418	79,471
Palermo	Turnover	81,250,169	140,129,020	222,097,542
	Outlet	4	41	46
	Item	40,532	33,349	73,881
Piacenza	Turnover	83,803,482	82,586,377	166,389,859
	Outlet	4	15	19
	Item	44,341	50,466	94,807
Torino	Turnover	738,418,729	412,833,155	1,152,108,585
	Outlet	28	101	130
	Item	83,342	54,420	137,762

The following Figure 8 shows the total turnover of all outlets of the five provinces for each of the 126 consumption segments (6-digit COICOP classification) for food and grocery. The wide range of the consumption segment turnover arises from the figure. The black bars highlight the three consumption segments (coffee, pasta and mineral water) on which the first experiments of the selection of the samples were concentrated even if only for the Torino province.

Figure 8 - Total turnover (six chains and five provinces) for consumption segment (2014)



5.3 Distribution of turnover - relevant weeks and permanent series

In this paragraph, the analyses are focussed both on the observed series (EAN+outlet code) belonging to the relevant weeks of each month and on permanent series (panel series SD).

Table 5 shows the whole turnover observed for the five provinces, respectively on the whole set of series (A), on the relevant week series (B) and the panel series (C).

Table 5 - Total turnover for all series, relevant week series and panel series, five Italian provinces (2014)

Province	Turnover			% Coverage		Number of panel series
	All weeks all series (A)	Relevant weeks all series (B)	Relevant weeks panel series (C)	B/A	C/B	
Torino	1,152,108,585	793,433,903	562,758,215	68.87	70.93	7,185,048
Ancona	286,249,412	199,336,585	134,533,396	69.94	67.49	2,331,144
Cagliari	233,772,680	162,008,273	109,160,452	69.30	67.38	1,583,448
Palermo	222,097,542	153,924,950	91,512,573	69.31	59.45	1,388,724
Piacenza	166,389,859	115,388,514	82,736,445	69.35	71.70	1,123,092

Observing only relevant weeks allows one to take into account about 70% of the turnover of all weeks (52 weeks of the year 2014), without relevant local differences. Then, looking at the coverage turnover of the panel series with respect to relevant week series, it ranges from 59.45% (Palermo) to 71.70% (Piacenza).

5.4 Continuity of products

Following the underlying idea that price indices will be computed on the base of a fixed sample (basket) of series during a twelve months period, in this paragraph, the continuity of the presence of the EAN codes in general (elementary items regardless of the outlets where they are sold) in the period is investigated in order to help in determining relevant subsets of items to be included in the basket and in the population to be sampled.

Prices of the series (EAN+outlet code) included in the basket are collected each month and therefore they should be all available at each time unless products acknowledged as seasonal. The cycle of life of the elementary items (EAN codes) is then investigated in order to identify relevant seasonal products other than those already known as seasonal (Fruits and Vegetables for example).

A definition of continuous EAN code is then needed and, on the complementary side of the non-continuous EAN code, it has to be distinguished between different kinds of temporary availability (such as seasonal, new entry, temporary or definitive absence).

The availability of the item/product for price collection can be registered for each of the 52 weeks in a year. Moreover, for each month the first 3 complete weeks are regarded as relevant for price collection.

Four different definitions of continuous items are considered:

a) available at least in 1 relevant week for each month;

b) available at least in 2 relevant weeks for each month;

c) available in all the 3 relevant weeks for each month;

d) available each of the 52 weeks in the year.

It worth notice that the four different definitions define a hierarchy in the sense that the set of continuous products according to the d) is included in the one obtained applying each of the previous definitions, and so on.

In Table 6 the number of different items registered in Torino during 2013 is reported for each of the above definitions. Moreover, the coverage with respect to the total number of items and to the amount of turnover is shown in percentage.

In all four cases the set of continuous EAN codes represents more than 80% of the total amount of the turnover. Going through the definitions from a) to d) the number of different items decreases more than 10% while the coverage of the total amount of the turnover decreases of less than 4%.

The definition a) has been then adopted to identify the set of continuous products whose relevance and characteristics are investigated in the next paragraph. The attention here is focussed on the complementary set of non-continuous EAN codes.

Table 6 - Number of continuous items, coverage of the total number of items and of the total amount of turnover by four different definitions (Torino, 2013)

Availability	N. Item	% Coverage	
		Total item	Total turnover
At least one week each month	47,760	49.7%	87.0%
At least two weeks each month	43,499	45.3%	85.5%
Three weeks each month	38,469	40.1%	83.6%
52 weeks a year	37,450	39.0%	83.1%

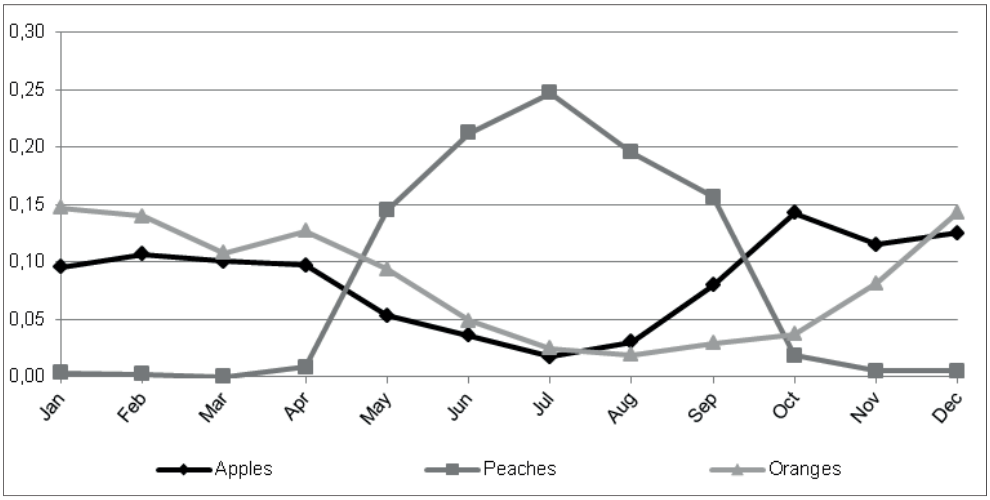
Seeking simplicity and comparability between periods the set of data analysed is reduced to the EAN codes registered in at least one of the relevant weeks in a month. It is implicitly assumed that the sold items only outside the 36 relevant weeks cannot represent valuable products and then can be therefore disregarded. Moreover, the amount of turnover is computed considering only the relevant weeks.

To get a sort of benchmark firstly, the case of well-known categories of seasonal product are investigated. Figure 9 shows the relative amount of turnover (monthly divided by the yearly amount) for Apples, Peaches and Oranges in the 12 months of 2013.

As expected the amount of Peaches turnover is concentrated between May and September while Apples and Oranges are mainly sold in the Fall and Winter period.

It is worth notice that, in order to describe the cycle of life of an EAN code two different aspects have to be considered: (i) if the corresponding product is available for price data collection in a month; (ii) how the amount of turnover is distributed over the whole period. If only the first aspect is considered a product may be classified as continuous because of its availability in each month but the amount of turnover is fully produced in one or two months that induces to classify it as temporary instead.

Figure 9 - Relative amount of turnover for apples, peaches and oranges per 12 months (Torino, 2013)



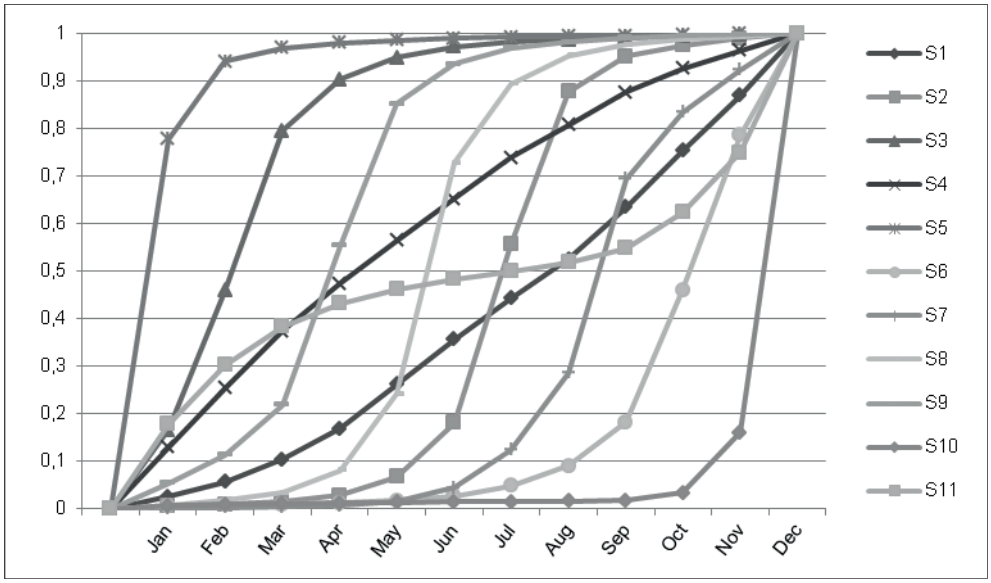
In order to highlight the "standard" different pattern of temporary products, the cumulative over a twelve months period (year 2013) of the relative amount of turnover has been computed. Each elementary item is then characterised by a pattern of 13 increasing values (the first and the last being 0 and 1 for each code). Patterns have been grouped into 11 different clusters according to a *k-mean* model-based procedure[3].

Results are synthetically shown in Figure 10 where the midpoints of each cluster is connected by a line. The purpose of the analysis is to identify standard patterns representing a recognizable cycle of life. The 2 clusters represented by the first line on the left (S5, grey stars) and from the last on the right side (S10, grey rotated squares) can describe the cycle of life of "Christmas products". The line S2 (grey big squares) growing from quite 0 to close to 1 between June and October could describe the pattern of seasonal summer products (as Peaches) while the line S11 (grey squares) growing in the first months and the last ones can be adopted to describe winter seasonal products (Oranges).

3 Software SAS, Fastclus procedure implementing the method called *nearest centroid sorting* introduced in Anderberg, M.R. (1973).

The two lines S1 and S4, growing quite regularly through the year, seem to represent continuous products but, as they have been removed from the data set, may describe "quite continuous" products that are not available in a very few number of months. An EAN code can be therefore classified as seasonal if it presents the same or similar pattern both in 2013 and in 2014.

Figure 10 - Patterns of the cumulative amount of turnover (11 cluster's means) for temporary products (Torino, 2013)



5.5 Seasonality and discontinuity of products

Seasonal products are products sold just at certain times of the year, following a seasonal pattern. According to this definition, it is clear that seasonal products must be detected in the set of *non-continuous* EAN codes, which are sold for less than 12 months during the year.

For the Torino province in the year 2014, this sub-set covers on average only 11.8% of the overall annual turnover, even if it represents 46.5% of series (EAN+outlet code); the peaks of 17.2% and 18.3% of monthly turnover, respectively in April and December, are the first evidence of a seasonal trend in consumptions (Table 7).

Table 7 - Coverage of total turnover per month and total number of EAN codes by item availability (Torino, 2014)

Availability	%Coverage													
	Total turnover per month													Total number of EAN codes (year)
	1	2	3	4	5	6	7	8	9	10	11	12	all	
Less than 12 months	9.3	9.2	9.5	17.2	10.6	11.1	10.8	11.3	11.2	11.0	12.3	18.3	12.0	46.5
12 months	90.7	90.8	90.5	82.8	89.4	88.9	89.2	88.7	88.8	89.0	87.7	81.7	88.0	53.5

Seasonality in consumer prices is caused by a variety of influences, some on the demand side and some on the supply side.

On the demand side, people have different needs depending on the period of the year and on climate conditions. Seasonal consumption patterns arising from this factor normally display about the same behaviour year after year, although the price effects will be modified by supply factors, including possible substitutions. Christmas, Easter, vacation periods and other practices influence the rise and fall in consumption.

On the supply side, the greatest seasonal price changes result from variations in agricultural production, especially among the perishable foods. Consumer demand for these elementary items appears to be quite stable throughout the year, with the result that limited supplies in the seasons when they are not available to determine more elevated prices.

Treatment of seasonal items in CPI is a quite difficult task; their identification requires the availability of at least two yearly collections of weekly/monthly observations in terms of quantities/expenditures, in order to verify whether a periodicity exists.

Analysing data by a group of products is then easy to recognise seasonal movements in the supply and prices of specific *markets* (the lowest level of ECR classification of products, allowing a deeper detail than segments) within consumption segments, although the peaks and the magnitude of seasonal fluctuations may vary widely from year to year: fresh fruits and vegetables, for example, may virtually disappear from the market during certain periods each year. The same problem exists also in some seasonal articles of clothing and in those products which are typically put on the market just for festivities, such as traditional foods or special gift packages.

The identification of seasonal market has been performed on the 2013-2014 collections of monthly total turnover for all products in the Torino province. On the subset of elementary items sold in both years in less than 12 months, the two distinct distributions of the number of contiguous months of sale have been compared in order to isolate, in each segment, the EAN code characterised by a recurring presence approximately in the corresponding period of the considered couple of years.

For those, the annual product turnover in percentage on the total turnover of its market has been calculated and for each market, elementary items have been flagged if they are characterised by a percentage value higher than a determined quantile in the distribution of percentage turnover for all the products in the market.

Finally, the concentration of most flagged EAN codes in a particular market, jointly with a high relative weight of that market on the consumption segment in terms of turnover, identifies it as a "seasonal market".

The described procedure proved to be efficient in the detection of the seasonal component of the annual consumptions curve for the analysed consumption segments. For the group of "Pastry cook products", for example, it has correctly identified the items belonging to the already mentioned market of "Traditional festivity foods" which are sold mainly in Christmas and Easter: while the overall monthly turnover curve depicts a general fall of the consumption from May to August in 2014 (Figure 11), the curve related only to the items in the segment flagged as seasonal clearly shows the opposite trend with a maximum in April and December (Figure 12).

Figure 11 - Percentage of monthly turnover on the total COICOP turnover, COICOP= "Pastry cook products", by all items (Torino, 2014)

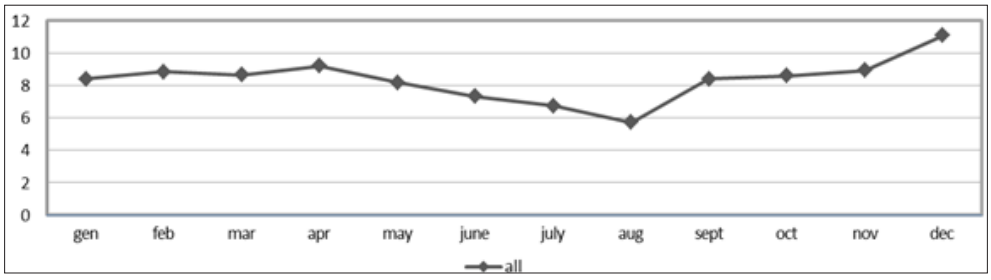
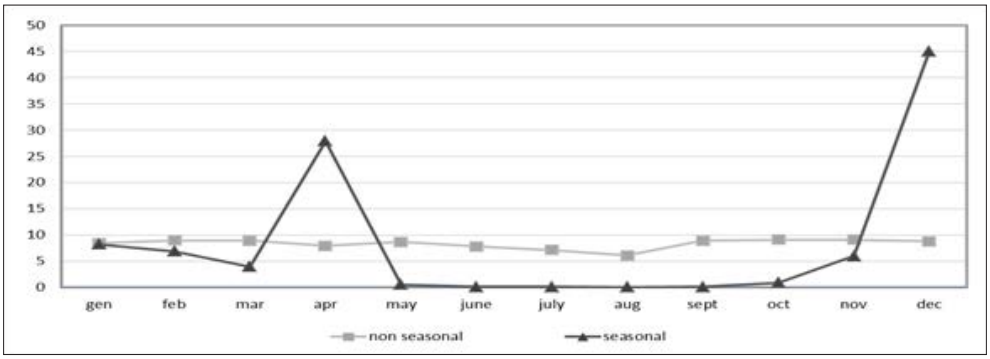


Figure 12 - Percentage of monthly turnover on the total COICOP turnover, COICOP= "Pastry cook products", by seasonality and non- seasonality items (Torino, 2014)



The monthly evolution of expenditures for the complementary set of non-continuous and non-seasonal products is outlined by the grey squares line in Figure 12: here are represented all the items in the consumption segment with a negligible weight with respect to the market or characterised by a certain degree of volatility. Specific assumptions must be made in order to detect the possible sources of volatility, which appears through the fluctuations of sales or by random "entry-exit" patterns during the year.

By applying this procedure, it has been possible to evaluate the coverage of seasonal items for the whole set of product groups: for 2014 in Torino province it consists in 27.7% of the yearly total turnover for the items available less than 12 months, a percentage which grows to 59% in April (Table 8). In the set of non-continuous items, those one flagged as 'not seasonal' count up almost to 87% of the total number of non-continuous series, providing 72.3% of the correspondent turnover.

Table 8 - Coverage of total turnover and total number of EAN codes of seasonal items, by month and item availability (Torino, 2014)

Availability	Seasonality	%Coverage													Total number of EAN code (year)
		Total turnover per month													
		1	2	3	4	5	6	7	8	9	10	11	12	all	
Less than 12 months	yes	24.4	27.9	27.0	59.0	24.0	25.6	16.8	16.5	11.1	13.0	17.0	37.1	27.7	13.1
	no	75.6	72.1	73.0	41.0	76.0	74.4	83.2	83.5	88.9	87.0	83.0	62.9	72.3	86.9
12 months	yes	2.6	3.0	3.0	10.5	2.9	3.1	2.1	2.1	1.5	1.8	2.4	7.2	3.6	2.6
	no	97.4	97.0	97.0	89.5	97.1	96.9	97.9	97.9	98.5	98.2	97.6	92.8	96.4	97.4

While seasonality is fairly known for specific product groups such as the ones already mentioned, seasonal movements are less predictable for particular product sectors: depending on their market share, a periodicity could be not so evident at a first glance just examining the sales trend of the whole segment.

Table 9 figures out the first three COICOP segments, ordered by decreasing percentage, which totalise at least 40% of coverage of the total number of EAN code, respectively for seasonal and non-seasonal items.

The case of low evidence of seasonality is represented by 'Body lotions' in the "Personal care" group; with a 14.4% they are the second most sold seasonal items by number of EAN code covered, even if they totalise just 4.7% of the turnover of its COICOP segment. On the contrary, even for those item groups which have a strong seasonality nature (*i.e.* fresh food or clothing), a component of volatility can always be present due to external events.

Table 9 - Coverage of the most sold items on total turnover and total number of EAN codes for discontinuous items, by seasonality and COICOP segment (Torino, 2014)

Seasonality	COICOP segment	COICOP code	% Coverage	
			Total number of EAN codes	Total turnover
Yes	Pastrey cook products	01.1.1.4.2	15.1	15.9
	Body, hand and hair lotions	12.1.3.3.3	14.4	4.7
	Chocolate	01.1.8.3.0	13.8	26.0
	Total		43.3	46.6
No	Articles for personal care, parfumes, make-up	12.1.3.2.2; 12.1.3.3.1; 12.1.3.3.2	21.1	7.2
	Cleaning and maintenance products	05.6.1.1.1; 05.6.1.1.2	8.3	13.1
	Cakes, tarts; ready-made meals; processed vegetables	01.1.7.3.2; 01.1.9.4.0; 01.1.1.4.2	8.1	10.3
	Total		40.1	35.2

6. Experimental framework for sampling from Scanner Data

6.1 Objective and theoretical context

The experiments carried out by the research group were developed in two phases, assuming in the first one only a static population approach (fixed basket) while, in the second, also a dynamic population approach (flexible basket).

The general aim was evaluating the use of SD for the compilation of the elementary price indices (first level of price index calculus on which the subsequent aggregations are based, in the Italian case the consumption segments) from a sampling perspective.

The elementary price indices are computed considering both closed (fixed basket) and open (flexible basket) population: direct indices are built on a fixed basket of products defined at reference time, ignoring new products; direct chain indices are built on a flexible basket that includes all products that disappear or appear (new products) in a year.

In a dynamic context, the bilateral indices are generally calculated on matched-items: only price relatives of items that are sold in two consecutive months enter in the index formulas (flexible basket) (Ivancic *et al.*, 2011). The comparison of two periods, 0 and t, is based on the chain approach. Chain indices take into account the movements of prices within the considered time interval, thus renewing the basket at each sub-interval and, consequently, solve the base change through the change of weights. So, using chain indices the shrinkage effect over time due to the attrition of a fixed basket of products is solved. On the other hand, in the static population, the loss of representativeness of the basket is addressed through the yearly base change of the index and the renewal of the basket. In this context, the comparison of two periods, 0 and t, or binary temporal index, is based on the direct (traditional) approach.

In the dynamic universe, however, period-on-period chaining of weighted indices introduces chain drift, also for superlative Fisher and Törnqvist indices (de Haan *et al.*, 2016). This source of bias, which increases as the time series grows, is due to the non-transitivity of the weighted price indices.

The transitivity of indices is not important in the static universe, as chaining is not required for direct (bilateral) indices, but is more crucial in the dynamic approach.

6.2 Experimental phases

The goal of the first experimental phase was to evaluate the performance of different sample selection schemes of series (references individuated by EAN and outlet codes) and the use of estimators of weighted and unweighted indices for CPI in a static situation. Following a fixed basket method, different samples of series are selected at the beginning of the reference period and followed during the year. In this phase a simplification was used: the implications of the life-cycle of series, seasonality issues and missing data were not taken into account and only panel series were considered as universe for sampling and price index evaluation. The definition of panel data is based on the permanent series concept, which refers to those series with positive turnover for at least one relevant week (the first three full weeks) in each month of the considered year, starting from the December of the previous year.

The population parameters taken in account are three classic aggregation formulas of monthly bilateral price index: Jevons (unweighted), Fisher (ideal) and Lowe (weights from quantities of previous year). In the static population approach, the use of Fisher (superlative) price index formula is undoubtedly the best way to measure price change: Fisher price index is calculated as the geometric mean of the Laspeyres price index and the Paasche price index. Jevons index is an unweighted CPI that uses price information only (it assumes that expenditure shares remain constant), while Fisher and Lowe use also quantity information. These last indices consider expenditure shares at different times (current and reference period) as weights (Gábor and Vermeulen, 2014). Fisher ideal index is thus preferred by economic theory, it uses quantities at different times and allows for substitution effects. The lack of weighting in the Jevons index is a potential source of bias and the opportunity of weighting items "according to economic importance" is supported by the theory of index numbers (de Haan *et al.*, 2016). In a probabilistic sampling context, it has to be specified that the properties of the estimators must also be considered with the properties of the corresponding indices.

In the second phase of the study, some experiments were carried out to highlight the differences between a static and a dynamic population approach in the construction of the elementary price indices. The goal of the experiments was to analyse how some sources of bias can affect the estimates of different index aggregation formulas in both approaches. In the fixed basket approach, bias can be introduced by the reduction in the size of the sample because of disappeared products (shrinkage), by ignoring new products and temporary missing products. In the dynamic population approach, some sources of bias can be related to the matched model and the type of index aggregation formulas utilised. The matched-model based on the exact matching of items sold in two consecutive months does not explicitly account for unmatched new and disappearing items and does not include temporary missing items. Constructing a time series by chaining period-on-period matched-model Jevons indices can avoid chain drift that affects weighted indices. Chain drift occurs if a chained index "does not return to unity when prices in the current period return to their levels in the base period" (Nygaard, 2010; ILO *et al.*, 2004, p. 445). Furthermore, the lack of weighting, the absence of adjustment for quality change and the lack of imputation of temporary missing items are potential sources of index bias (de Haan *et al.*, 2016). The two approaches refer to different sampling schemes: under the static approach, the series is drawn through a two-stage sampling design (outlets and products), while under the dynamic approach, only the selection of outlets is considered.

As in the first phase, the experiments were conducted starting from a panel series, but in this case, artificial populations were generated with products appearing and disappearing (momentarily and permanently). The population parameters here considered are monthly chained bilateral unweighted (Jevons) and weighted superlative indices (Fisher and Törnqvist[4]).

In both experiments, comparison between alternative selection schemes are made for each price index taking the corresponding universe (panel series SD) index value as a benchmark. Indices performance were evaluated in terms of bias for all selection schemes of series. For probability selection schemes, the accuracy (relative bias and sampling variance) of the price indices has been studied in a Monte Carlo simulation scenario. In this context, 500 samples have been selected, according to different sampling designs. Indices

4 Törnqvist index is the weighted geometric average of the price relatives.

variability is measured considering the estimate of the relative sampling error, computed on the estimated indices in the replicated samples. For the sample selection and weighting of price indices, the total annual turnover was taken as reference.

The study was conducted on data relating to the provinces of Torino and Rome. In particular, in the first experimental phase, the probabilistic and non-probabilistic approaches in the selection of series have been investigated assuming as domains of interest three consumption segments (Coffee, Pasta and Mineral Water - COICOP 6 digits) or 88 markets (ECR groups) belonging to the six consumption segments (Coffee, Pasta, Mineral Water, Olive Oil, Spumante and Ice Cream). The SD reference universe is relative to Torino province, 121 outlets and six retail chains (Conad, Coop, Esselunga, Auchan, Carrefour, Selex) available for the year 2014. Scanner data referring to the previous year have been used for the sample selection and weighting of price indices that were based on the total annual turnover of 2013. In the second experimental phase, the study is carried out on Rome province in 2015 and three consumption markets (Short Semolina Pasta, IGP-IGT Italian White Wine, Laundry Bivalent Washing Machine Liquid + Gel).

6.3 Parameters and unbiased estimators

As described above, the parameters of interest taken into account in the experiments are monthly Jevons, Laspeyres, Paasche, Fisher, Lowe and Törnqvist indices.

According to a static population approach, for sake of simplicity, a formalisation of the population parameters considered and the corresponding unbiased estimators are shown in the following scheme (de Haan *et al.* 1999).

Scheme 4 - Price index and sampling estimator

	Population parameter	Sampling estimator
Jevons	$I_{J}^{0,t} = \prod_{i=1}^{n} \left(\frac{p_i^t}{p_i^0} \right)^{1/n}$	$\hat{I}_{J}^{0,t} = \prod_{i=1}^{n} \left(\frac{p_i^t}{p_i^0} \right)^{w_i \Big/ \sum_{i=1}^{n} w_i}$
Laspeyres	$I_{LA}^{0,t} = \sum_{i=1}^{n} \left(\frac{p_i^t}{p_i^0} \right) \left(\frac{p_i^0 q_i^0}{\sum_{i=1}^{n} p_i^0 q_i^0} \right)$	$\hat{I}_{LA}^{0,t} = \sum_{i=1}^{n} \left(\frac{p_i^t}{p_i^0} \right) \left(\frac{p_i^0 q_i^0 w_i}{\sum_{i=1}^{n} p_i^0 q_i^0 w_i} \right)$
Paashe	$I_{P}^{0,t} = \sum_{i=1}^{n} \left(\frac{p_i^0}{p_i^t} \right) \left(\frac{p_i^t q_i^t}{\sum_{i=1}^{n} p_i^t q_i^t} \right)$	$\hat{I}_{P}^{0,t} = \sum_{i=1}^{n} \left(\frac{p_i^0}{p_i^t} \right) \left(\frac{p_i^t q_i^t w_i}{\sum_{i=1}^{n} p_i^t q_i^t w_i} \right)$
Fisher	$I_{F}^{0,t} = \sqrt{I_{LA}^{0,t} \; I_{P}^{0,t}}$	$\hat{I}_{F}^{0,t} = \sqrt{\hat{I}_{LA}^{0,t} \; \hat{I}_{P}^{0,t}}$
Lowe	$I_{LO}^{0,t} = \sum_{i=1}^{n} \left(\frac{p_i^t}{p_i^0} \right) \left(\frac{p_i^0 \, q_i^z}{\sum_{i=1}^{n} p_i^0 \, q_i^z} \right)$	$\hat{I}_{LO}^{0,t} = \sum_{i=1}^{n} \left(\frac{p_i^t}{p_i^0} \right) \left(\frac{p_i^0 \, q_i^z \, w_i}{\sum_{i=1}^{n} p_i^0 \, q_i^z \, w_i} \right)$
	q_i^z refers to the quantity series in the previous year	
Törnqvist	$I_{T}^{0,t} = \prod_{i=1}^{n} \left(\frac{p_i^t}{p_i^0} \right)^{\frac{1}{2}\left(\frac{p_i^0 q_i^0}{\sum_{i=1}^{n} p_i^0 q_i^0} + \frac{p_i^t q_i^t}{\sum_{i=1}^{n} p_i^t q_i^t} \right)}$	$\hat{I}_{T}^{0,t} = \prod_{i=1}^{n} \left(\frac{p_i^t}{p_i^0} \right)^{\frac{1}{2}\left(\frac{p_i^0 q_i^0 w_i}{\sum_{i=1}^{n} p_i^0 q_i^0 w_i} + \frac{p_i^t q_i^t w_i}{\sum_{i=1}^{n} p_i^t q_i^t w_i} \right)}$

In the formulas, w_i is the direct weight, *i.e.* the inverse of the inclusion probability of the sampling unit deriving from the sampling design.

Under a dynamic approach, analogous formulas can be expressed for the chain indices, by substituting (0, t) with (t-1, t).

A generic chain index obtained as product of index $I^{0,1}$, $I^{1,2}$,…, $I^{t-1,t}$ referred to sub-intervals s (0,1), (1,2), ...(t-1, t) can be expressed as:

$$I_{chain}^{0,t} = \prod_{s=1}^{t} I^{s-1,s}$$

6.4 Accuracy of price indices

Accuracy of the alternative price indices under different selection schemes of series is evaluated on Monte Carlo simulation scenarios considering the estimates calculated on 500 replicated samples.

Bias and relative sampling error formulas shown below are expressed for a generic parameter (price index) and regarding simulation context.

For a generic estimated index in the *c*-th products group, $\hat{\theta}_c$, the bias can be expressed as

$$B\left(\hat{\theta}_c\right) = E\left[\hat{\theta}_c\right] - \theta_c, \tag{1}$$

in which:

$E\left[\hat{\theta}_c\right]$ is the expected value of the estimated index $\hat{\theta}_c$ in the products group *c*, obtained from 500 samples, and θ_c is the corresponding index value computed on the reference universe (panel series SD).

The relative sampling error of a generic estimated index $\hat{\theta}_c$ in the products group *c* can be expressed by

$$RE\left(\hat{\theta}_c\right) = \frac{\sqrt{Var\left(\hat{\theta}_c\right)}}{\bar{\hat{\theta}}_c}, \tag{2}$$

in which mean and variance of $\hat{\theta}_c$ are calculated on the estimates generated from the selection of 500 samples in the products group *c*.

7. Fixed basket approach: first experimental phase and results

7.1 Probability and non-probability selection schemes

In this experimental phase, probability and nonprobability selection schemes of series in each consumption segment (Coffee, Pasta, Mineral Water) are considered: probability two-stage sampling in the first approach; cut-off and representative elementary item samples in the second approach.

Both for probability and nonprobability sample schemes, series are selected from a sample of outlets (primary stage units - PSU) stratified by chain and outlet type (hypermarket and supermarket). In each stratum, the sample has been allocated proportionally to the turnover. The selection of outlets is carried out in each stratum by simple random sampling (SRS). The outlets' sample size has been fixed at a number of 30 out of 121 outlets of retail trade modern distribution in Torino province. In any case, the series constitute the secondary stage units (SSU).

Nonprobability sampling[5] of series was carried out by selecting series based on cut-off thresholds of covered turnover in the previous year, 2013: two samples are formed with all the series covering respectively the 60 and 80 percent of the total turnover in each of the considered consumption segment in the selected outlets. Moreover, considering the currently used fixed basket approach, a reference selection scheme was defined as selecting the most sold EANs for each representative product in the selected outlets.

For the probability sample, the sample size for SSU is fixed by a sampling rate of 5 percent of the number of EANs in each consumption segment in the sampled outlets. Sample series are selected with probability proportional to size (PPS), in terms of total turnover of the previous year, by adopting Sampford sampling (Sampford, 1967) and Pareto sampling (Rosén, 1997a and 1997b).

Sampford's method is an extension of Brewer's method that selects more than two units from each outlet and without replacement. Units for which

5 The selection of elementary items is made starting from a random sample of outlets. This approach introduce a sampling variance of the estimates determined by the selection of the outlets sample. This component of sampling variability was not taken into account, therefore only the indices bias in non-probability selection schemes is considered.

the initially size measure (turnover during year 2013) is larger than a certain threshold turnover are selected with certainty, where after the inclusion probabilities are calculated for the remainder of the elementary items universe in the sampled outlets. Threshold turnover is defined taking into account the average turnover of outlet, the sampling rate (5%) and the *k* coefficient which can take values greater than 0 and less than 1.

Pareto sampling (PAS) is an order PPS sampling based on the definition of two variables:

- the target inclusion probabilities, $\lambda_{ji} = n_j * s_{ji} \Big/ \sum_{i=0}^{N_j} s_{ji}$,

where n_j is the sample size in the *j*-th outlet (j=1,…,30), determined as product between the sampling rate (5%) and the total number of elementary items in the *j* outlet, N_j, and s_{ji} is the size of *i*-th elementary item ($i=1,...,N_j$) in the *j*-th outlet.

For $\lambda_{ji} > 1$ then let $\lambda_{ji} = 1$.

- the uniform random variable *U* between 0 and 1.

The size measure is associated with each sampling unit and a ranking variable is constructed as a function of these two variables as $Q_{ji} = \frac{U_{ji} * (1 - \lambda_{ji})}{\lambda_{ji} * (1 - U_{ji})}$, in which U_{ji} is a permanent random number (PRN) associated to the *i*-th elementary item in the *j*-th outlet.

In each sampled outlet, elementary units are then sorted in ascending order and the n_j units per outlet with the smallest values of the ranking variable are included in the sample.

In the following scheme the adopted probability and nonprobability selection schemes are synthesised.

Scheme 5 - Probability and nonprobability selection schemes

Selection scheme	Sampling unit	Stratification	Allocation	Selection
One stage - cut-off	Outlet Ean-code	Chain – outlet type	Proportional	SRS Cut-off
One stage – most-sold elementary items	Outlet Ean-code	Chain – outlet type	Proportional	SRS Most-sold elementary items
Two stage sampling	1° Outlet 2° Ean-code	Chain – outlet type	Proportional	SRS
			Fixed sampling rate	Sampford Pareto

7.1.1 Scanner data: operational context

The analyses shown below describe the SD operational framework on which the experiments were carried out: three consumption segments, mineral water (01.2.2.1.0), pasta (01.1.1.6.1) and coffee (01.2.1.1.0) in the Torino province.

Table 10 presents, for each consumption segment, the coverage turnover of relevant week series (B/A) and panel series in relevant weeks (C/B), and the number of panel series.

Table 10 - Total turnover for all series, relevant week series and panel series by consumption segment (Torino, 2014)

Consumption segment	Turnover			%Coverage		Number of panel series
	All series (A)	Relevant weeks all series (B)	Relevant weeks panel series (C)	B/A	C/B	
Coffee	28,622,978	19,665,517	15,692,414	68.7	79.8	9,608
Pasta	26,192,517	17,902,061	13,631,744	68.4	76.2	23,636
Mineral water	26,434,572	18,506,760	16,851,559	70.0	91.1	6,990

Coverage turnover is slightly variable in the consumption segments (minimum 68.4%, maximum 70%) when taking into account the percent ratio (B/A) between the turnover of all series in the relevant weeks and the turnover of all series (52 weeks of the year 2014); coverage turnover is different in the consumption segments when the turnover of the panel series is compared to all series in relevant weeks (C/B) - over 75% for coffee and pasta segments, above 90% for the mineral water segment.

Table 11 describes the turnover covered in the consumption segments in the two cut-off samples with thresholds turnover defined at 60 and 80 percent and the average number of elementary units per outlet considered in the two scenarios.

Table 11 - Percentage and average number of elementary items per outlet covering 60 and 80% of turnover by consumption segment (Torino, 2014)

Consumption segment	Percentage of series		Average number of elementary items per outlet		
	Turnover threshold 60%	Turnover threshold 80%	Total	Covering 60% of turnover	Covering 80% ofturnover
Coffee	16.2	36.1	46	8	17
Pasta	23.4	44.8	114	27	51
Mineral water	12.1	26.3	34	4	9

7.1.2 Main results

In this paragraph some results of the experiments are illustrated: for the mineral water segment the estimates of the twelve-monthly indices achieved through five samples selection are presented; estimates of the monthly Lowe indices are analysed for all consumption segments. The emphasis placed on that index is since it takes full advantage of the information found in the SD.

Tables and figures below present the results of the two Monte Carlo simulations (Sampford and Pareto sampling designs) for monthly Lowe, Fisher and Jevons indices. Besides the outlet selection, a probability or non-probability selection of series has been implemented for comparing the estimates, mainly in terms of bias with respect to the real value of indices.

In seek of brevity, only plots on the mineral water segment have been reported (Figure 13). Figure 13 shows the level estimates of the monthly Lowe, Fisher and Jevons indices computed on probability and non-probability samples and the real value (universe panel series SD) of the corresponding index. Figure 14 shows the level estimates of the monthly Lowe indices computed for coffee and pasta segments.

Irrespective of the consumption segment, the most sold is generally far from the real value and sometimes neither able to catch the trend. The cut-off at 80% is always better than the cut-off at 60% and both are closer to the real value of the index lower is the variability of the segments in terms of prices and turnovers of series.

In general, with the probabilistic selection of elementary items within the selected outlet less biased estimates are obtained. This holds especially under Sampford sampling. Instead, under PAS the results are more biased but more cumbersome because they seem to be dependent on the series variability of the segments in terms of price and/or turnover. The Jevons index represents a kind of exception in this context because the cut-off at 80% and at 60% is less biased than the estimates obtained under Sampford and PAS sampling of series.

Figure 13 - Fisher, Jevons and Lowe indices computed with different selection schemes of series for mineral water segment, year 2014

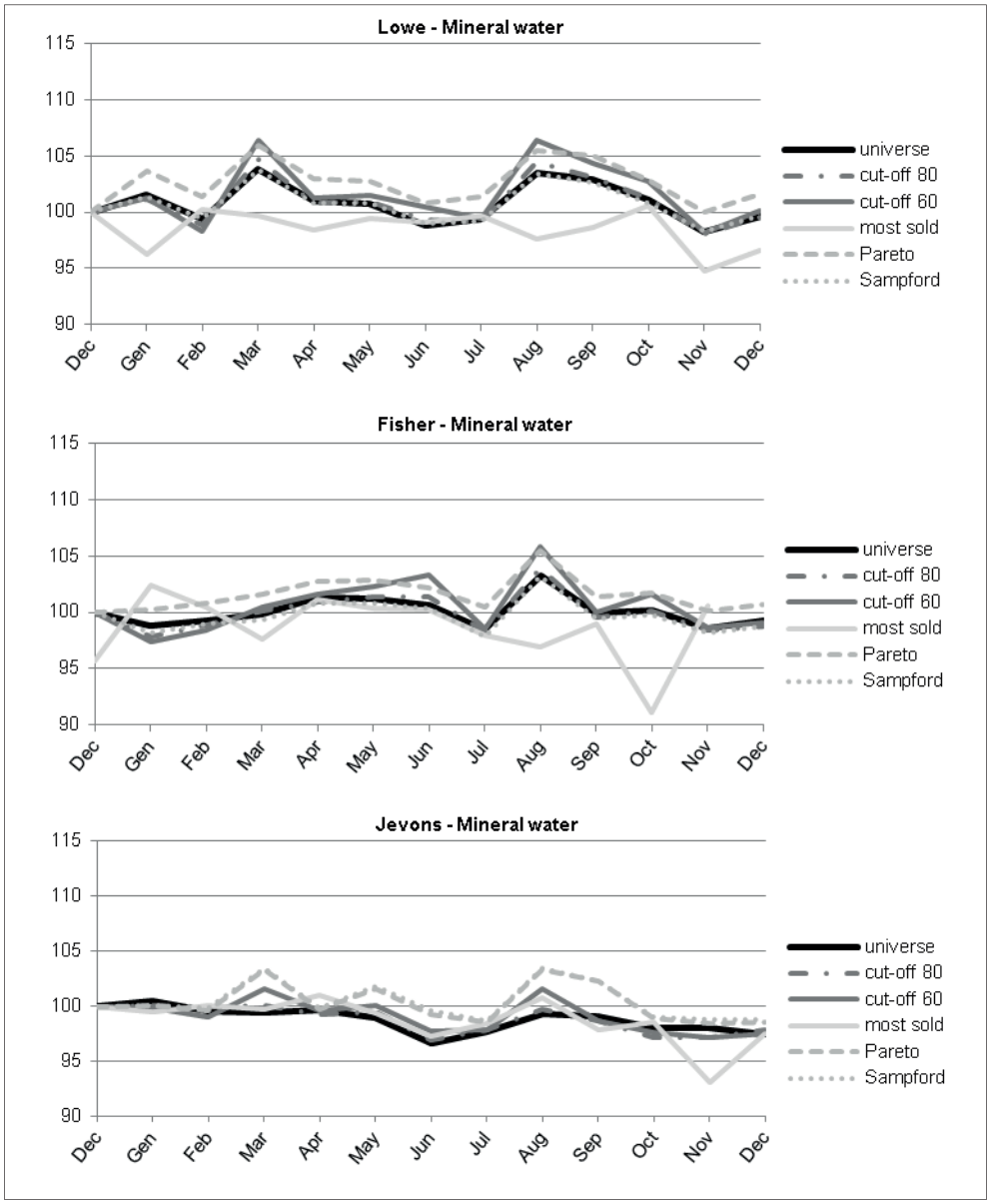
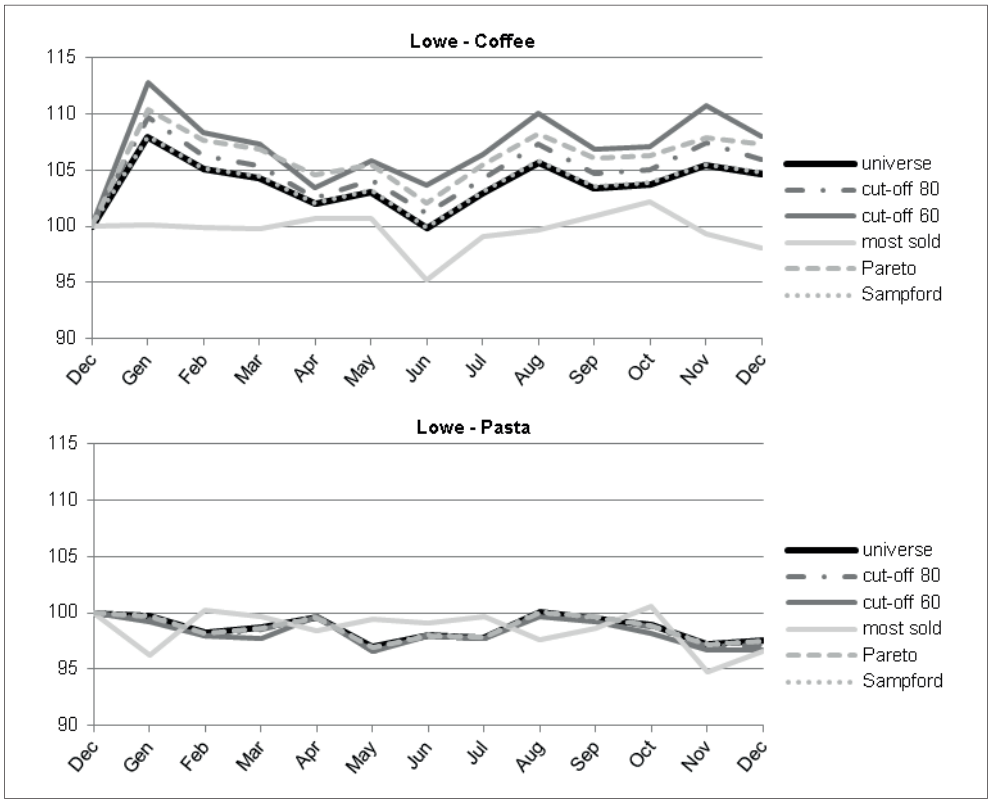


Figure 14 - Lowe indices computed with different selection schemes of series for coffee and pasta segments, year 2014



Looking at Lowe index for the three considered segments, it is possible to see that, under Sampford sampling, Lowe is always unbiased. On the contrary, the most sold item is always far from the real value. Both the cut-offs and also PAS are close or not to the real value depending on the features of the segments, especially concerning the turnover variability of series. In fact, for pasta segment in which the turnover variability is low, the estimates of Lowe index are unbiased irrespective of the method of series selection and apart for the most sold item. In the other two cases the estimates under PAS are in between to those related to the cut-off at 80% and the cut-off at 60%. However, with PAS and with both the cut-off samples the estimates catch the trend of the price index.

In Tables 12 and 13, for each estimated price index, minimum and maximum values assumed by the bias and relative sampling error distributions of the 12 monthly indices are exposed for consumption segment and sampling design.

Table 12 - Bias distribution of monthly Lowe, Fisher and Jevons indices for consumption segment and sampling design

Consumption segment	Sampling design	Love Index		Fisher index		Jevons index	
		Min	Max	Min	Max	Min	Max
Coffee	Sampford	0.05	0.18	-0.40	-0.05	1.99	6.01
	Pareto	2.17	2.60	2.13	2.66	1.95	5.92
Pasta	Sampford	-0.08	0.07	0.09	0.50	-3.12	0.17
	Pareto	-0.11	0.06	-0.34	-0.04	-2.29	0.06
Mineral water	Sampford	-0.26	0.13	-0.71	-0.09	-0.47	4.18
	Pareto	2.74	5.25	1.15	4.81	-0.40	4.77

Table 12 shows that Lowe and Fisher indices present the lowest levels of bias under Sampford sampling in each consumption segment, opposite behaviour can be seen for the Jevons index. Lowe and Fisher indices perform well under Pareto sampling only in the pasta segment.

It is interesting to note that the increase of the bias in PAS design with respect to Sampford, is most conspicuous for Lowe and Fisher indices, mainly in the mineral water segment. The two sampling designs did not show to have a significant impact on the bias of the Jevons index.

Table 13 - Sampling Error distribution of monthly Lowe, Fisher and Jevons indices for consumption segment and sampling design

Consumption segment	Sampling design	Love Index		Fisher index		Jevons index	
		Min	Max	Min	Max	Min	Max
Coffee	Sampford	0.05	0.18	-0.40	-0.05	1.99	6.01
	Pareto	2.17	2.60	2.13	2.66	1.95	5.92
Pasta	Sampford	-0.08	0.07	0.09	0.50	-3.12	0.17
	Pareto	-0.11	0.06	-0.34	-0.04	-2.29	0.06
Mineral water	Sampford	-0.26	0.13	-0.71	-0.09	-0.47	4.18
	Pareto	2.74	5.25	1.15	4.81	-0.40	4.77

The results presented in Table 13 underline very high relative sampling errors for Fischer index and more content for Lowe and Jevons indices in all consumption segments and in both sampling designs.

In the mineral water segment higher relative sampling errors for all indices are found, in particular for Fisher index under Pareto sampling. Considering PAS design is known that the increase of the relative sampling error is most relevant for Lowe index with respect to Jevons index.

Looking both at the bias and relative sampling error distributions of the indices, it follows that, especially in Sampford sampling, Lowe index performs quite well in the coffee and pasta segments but less well in the mineral water segment.

Table 14 shows the coverage probability and width of 95% confidence intervals for Lowe, Fisher and Jevons indices achieved under two sampling designs for each consumption segment.

Table 14 - Coverage probability and width of 95% confidence intervals under Sampford and Pareto sampling with 5% sampling rate for Fisher, Jevons and Lowe indices per consumption segment

Consumption segment	Sampling design	Lowe Index		Fisher Index		Jevons Index	
		Coverage probability %	Width	Coverage probability %	Width	Coverage probability %	Width
Coffee	Sampford	94.70	5.75	95.83	14.58	12.10	6.04
	Pareto	70.17	6.77	92.50	14.40	12.80	9.02
Pasta	Sampford	94.92	4.88	95.38	12.35	68.87	5.41
	Pareto	94.88	4.58	96.00	12.33	76.01	12.20
Mineral Water	Sampford	94.90	10.66	94.35	20.03	78.18	9.29
	Pareto	69.81	11.20	94.68	26.73	77.13	7.60

The confidence intervals (CI) at 95% for Fisher, Jevons and Lowe indices have been derived, through a Monte Carlo method, under Sampford and Pareto with the same sampling fraction (5%).

Crossing the results in Table 14 on the coverage probability and the width of the CIs interesting results are derived. The Fisher index keeps the confidence probability close to the nominal confidence level independently of the consumption segments and the sampling design considered. However, the width of its CIs is usually much wider than those related to the other

two indices. Furthermore, Jevons index has narrow CIs among the considered indices due to its low variability, but its coverage probability level is always greatly below the nominal one because it is biased. Finally, Lowe indices has narrow CIs and good coverage probability levels, especially under Sampford sampling. Instead, under Pareto, it has lower coverage probability levels for coffee and mineral water segments. In fact, it seems that Lowe index is affected by bias when computed for segments with high variability of series both in terms of price and turnover.

7.2 Probability sampling designs

Besides comparing probability and non-probability schemes, a deepening was carried out on some probabilistic designs characterised by the use of different criteria of sample allocation, both for outlets and elementary items (EANs), and different selection methods of the sampling units: 1) one-stage stratified sample of EANs; 2) cluster sample of outlets; 3) two-stage sampling with stratification of outlet (PSU) by chain and type (hypermarket and supermarket) and EAN (SSU).

For each sampling design the size of the final sample of EANs was fixed in average at 7,400 to compare the different sampling strategies on equal sizes. The first sampling design was carried out stratifying the EANs by market (ECR group) in each consumption segment. Sample size is allocated among the strata through a Neyman formula, taking into account the variability of price relatives of the EANs in the markets observed in the reference year 2013. Two selection schemes were considered, a simple random sampling (SRS) and with probability proportional to size sampling (PPS) with size equal to turnover (2013). In both selection methods, some sampling units are selected with certainty.

In the second design, cluster sampling, a sample of outlets (14 out of 121 outlets) is selected. Outlets were stratified by chain and type (hypermarket and supermarket). In each stratum, two different allocation of outlets were tested: optimal allocation (Neyman) and proportional allocation based on the turnover of the strata in 2013. Outlets are selected with both SRS and PPS methods. All the EANs in the selected outlets were included in the sample.

Finally, the two-stage sampling design was characterised by a stratification of both PSU and SSU. The stratifications adopted for the PSU and the SSU are the same of the two schemes described above. The PSUs have been allocated through a proportional allocation and selected with PPS based on the turnover of the previous year. The SSU in the selected outlets have been allocated proportionally with the Neyman allocation defined for the stratified sampling and selected with SRS and PPS. To keep in average around 7,400 EAN in this case the sample of outlets has been fixed at a number of 30 out of 121 outlets. In the following scheme the probability sampling designs are synthesised.

Scheme 6 - Probability selection schemes of series

Selection scheme	Sampling unit	Stratification	Allocation	Selection
One stage	EAN-code	Market	Neyman	SRS
				PPS
Cluster	Outlet	Chain – outlet type	Neyman	SRS
				PPS
			Proportional	SRS
				PPS
Two-stage	1° Outlet	Chain – outlet type	Proportional	PPS
	2° Ean-code	Market	Proportional	SRS
				PPS

The comparison among behaviours of the three index aggregation formulas (Jevons, Fisher, and Lowe) has been conducted with the aim of underline the differences among them under the different sampling strategies considered. The same sampling estimator (*i.e.* a plug-in estimator) has been considered for all the index aggregation formulas and the same overall sample size have been drawn under each sampling strategy.

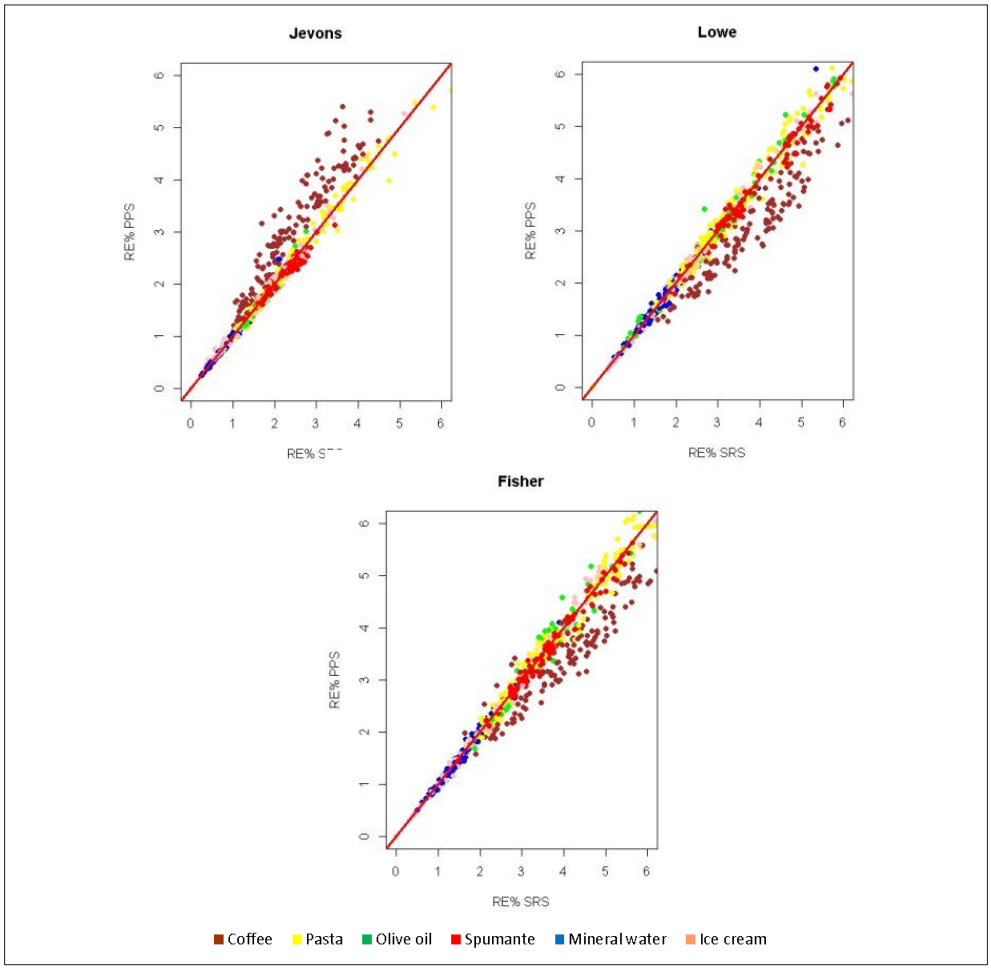
The CPI has been computed at market and consumption segment level for 13 months (from December 2013 to December 2014) for the province of Torino in the 88 markets related to the 6 consumption segments already listed above.

7.2.1 Main results

In the following analyses, relative errors (RE%) of Jevons, Lowe and Fisher estimates are presented under different sampling designs (Scheme 3).

From Figure 15, the estimates, for all the aggregation formulas, are in both cases unbiased (RB –relative bias is approximately equal to 0). With respect to the variability, it possible to see a slight difference among the indices.

Figure 15 - Relative error (RE%) of Jevons, Lowe and Fisher estimates under stratified sampling - SRS and PPS selection methods. Markets in coffee, pasta, olive oil, spumante, mineral water, ice cream. Torino 2014



Looking at Figure 16 the point cloud related to Jevons is usually below those related to Lowe and Fisher, which are almost at the same level.

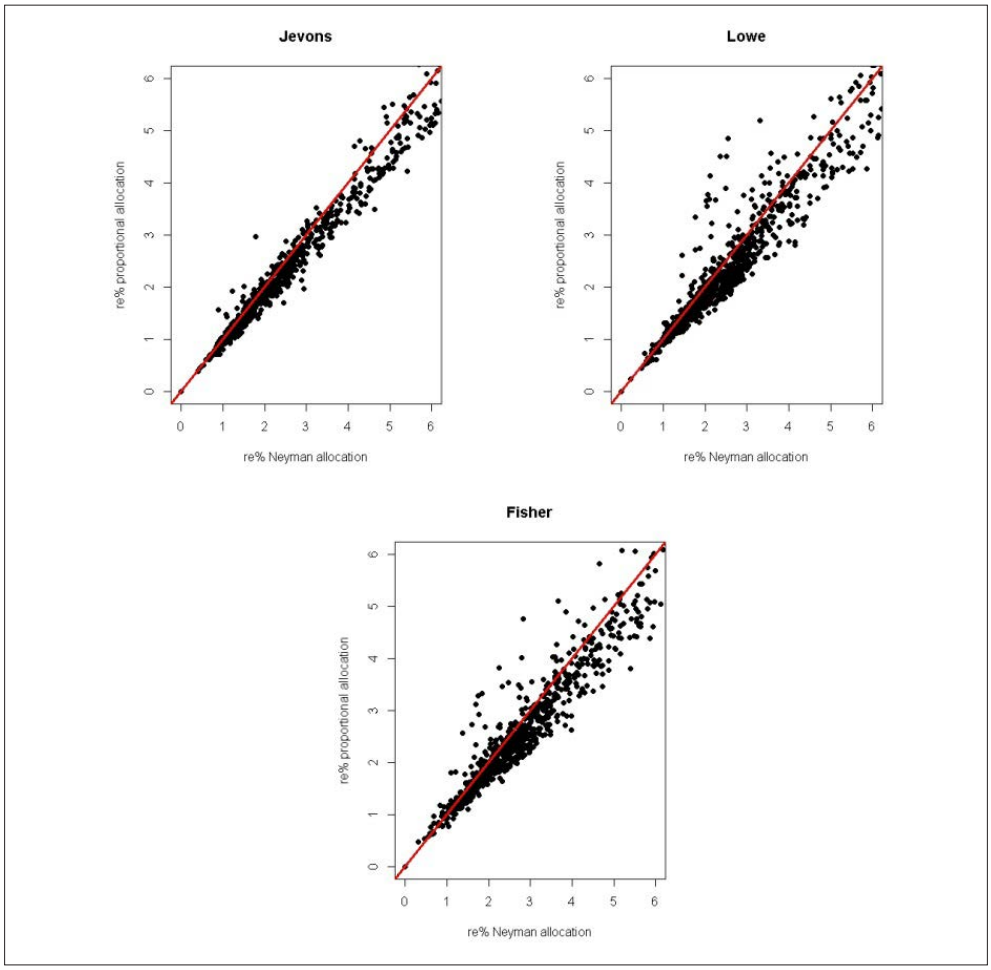
In the figure below, it is possible to notice also a different behaviour of the estimators of the indices with respect to the consumption segments. In particular, for the coffee segment, the estimator of Jevons index is more efficient when EANs are selected under SRS, whilst those for Lowe and Fisher are more efficient when the selection of units is based on PPS. For the other segments, there is no significant differences between the selection methods.

To keep in average around 7,400 EAN the outlets sample has been fixed at 14. In all the scenarios, also, in this case, the estimators of the indices are unbiased. However, when the outlets are selected through a PPS the estimates are more efficient, under both the allocation methods. Instead, between them, the optimal Neyman allocation seems to be less efficient for the outlets. Then the proportional allocation based on turnover of the previous year is preferable.

This advantage is more remarkable when the interest parameters are weighted price indices, such as Lowe and Fisher (Figure 16). In this case, there are no significant differences among consumption segments and among the level of RE% of the estimator of the indices.

As shown in the Figure 16, no significant differences arise using SRS or PPS, probably because most of the variability is at outlet level. This result points the attention to the importance on the allocation and selection method to be used at PSU level and on the size of the PSU sample. In some markets, this could bring an advantage in terms of RE% even if usually two stages sampling implies a higher design effect.

Figure 16 - Relative error (RE%) of Jevons, Lowe and Fisher estimates under stratified one-stage sampling with SRS and PPS selection of outlets. Torino, 2014

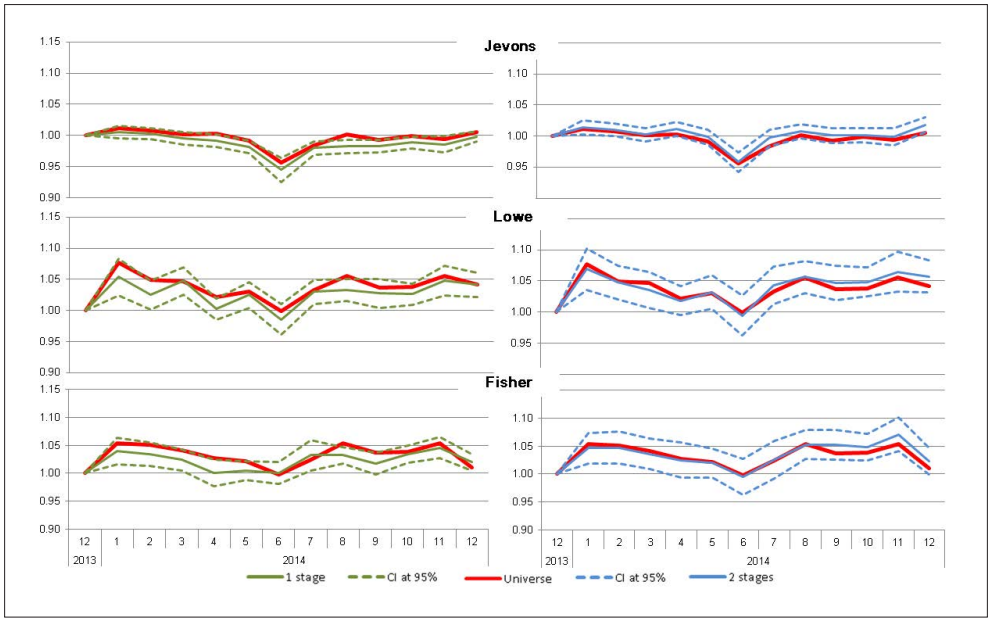


Looking at Figure 17 is possible to notice the difference among the indices estimated under the two different sampling strategies. The two sampling strategies compared are stratified one stage, proportional allocation of outlets and PPS selection versus stratified two stages, proportional allocation and PPS selection of outlets and re-proportionated Neyman allocation of EAN in the selected outlets in which EAN have been selected with SRS. The results of estimates on a single sample for the indices in the coffee segment in Torino obtained under the two sampling strategies has compared with the real value

(computed on the universe of SD of coffee segment in Torino). All of the estimates seems to catch properly the level and the trend of the related real index. The estimator of Jevons index has in both cases more narrow confidence intervals (CI) with respect to the other two, it means that its RE% is lower. In general, the length of CIs is wider under two stages sampling than under stratified one stage, even if the difference does not seem so large.

In terms of bias with respect to stratified sampling, in one stage and two stages sampling designs, the RB increases slightly, but the estimators can be still considered unbiased.

Figure 17 - Jevons, Lowe and Fisher indices for coffee segment estimated on one sample, confidence interval (CI) of estimates at 95% and real value (computed on the universe of SD). Torino, 2014



8. Fixed and dynamic population: second experimental phase and results

8.1 Objective and description of experiments

The second experimental study aimed to highlight the differences between static and dynamic population approach, using weighted and unweighted indices, and to measure the magnitude of sampling error and bias. The elementary price indices are computed considering both closed and open populations. When assuming a closed population, direct indices are built on a fixed basket of products defined at reference time, ignoring new products (fixed approach), when considering all series of an open population (dynamic approach), direct chain indices are built on matched series of two consecutive months (Ivancic *et al.*, 2011).

The two approaches refer to different sampling schemes. Under the static approach, the series are drawn through a two-stage sampling design, in which the Primary Stage-Units (PSU) are the outlets, the Secondary Stage Units (USS) are the EANs. The outlets, stratified by province, chain and type, are selected with probability proportional to their annual turnover. In each selected outlet, in each market, a sampling fraction of 20% of EANs is selected with probability proportional to annual turnover. Under the dynamic approach, only the selection of outlets is considered.

The experimental study is carried out on Rome province in 2015 and some consumption markets (ECR group). To represent the diversity of the population, three consumption markets different by their features are considered:

- Short Semolina Pasta, low dynamism with respect to products and low variability in prices;
- IGP-IGT Italian White Wine, medium dynamism with respect to products and high variability in prices;
- Laundry Bivalent Washing Machine Liquid + Gel, high dynamism with respect to products and medium variability in prices.

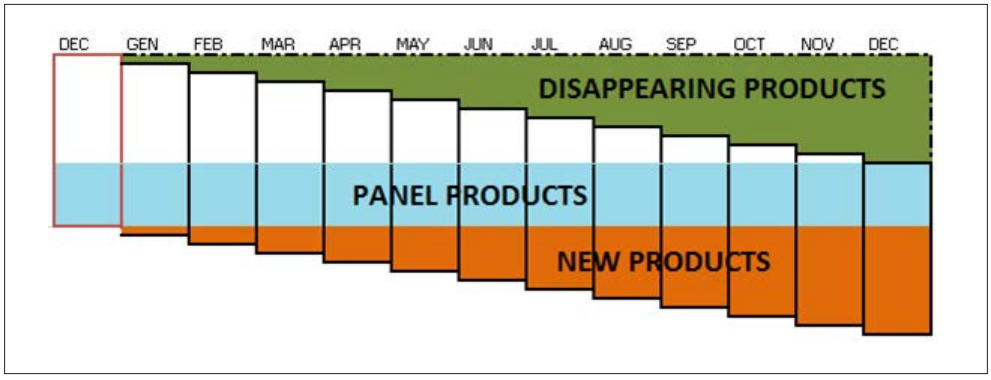
To take into account the variability due to the identification of disappearing, new and temporary missing products and probability sampling, a kind of super-population approach has been implemented.

The panel of products is the reference population. Starting from these data, alternative population have been generated where disappearing and new products, temporary missing, from time to time, are flagged applying survival functions; birth rate and "temporary-missing" rate are estimated on the observed complete data set (from which the panel is extracted as the 'always present series'). On these populations, the two sampling design is applied and elementary price indices are calculated[6]. The elementary indices considered are Jevons, Törnqvist and Fisher.

Under the dynamic approach, the monthly chained bilateral versions of these indices are used. In this case, the Jevons index is computed on all matched items for two months in a row and on a sub-sample identified by a threshold (matched-model with threshold, "Jevons wT"). The threshold is based on average expenditure shares across two adjacent months; items below the threshold are excluded from the computation. Therefore, an implicit weight is applied (Dutch method) (Van der Grient and de Haan, 2010, 2011).

The following schemes concern the panel series (Scheme 7) and the sub-population considered to decompose the overall survey error deriving from different sources.

Scheme 7 - Panel series, disappearing and new products



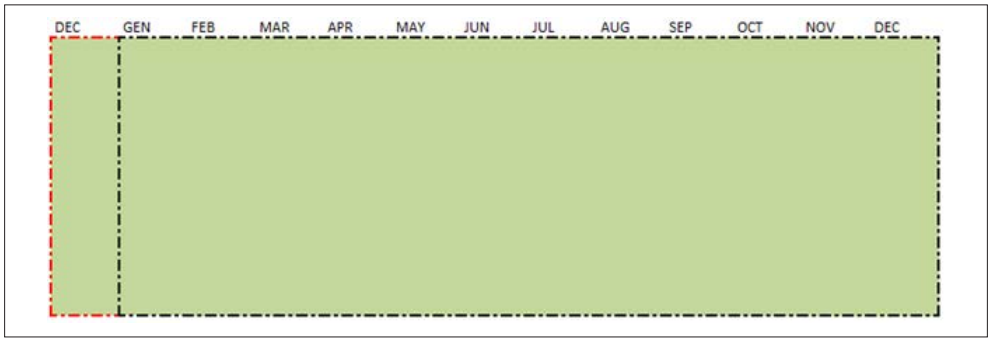
6 500 populations have been drawn and from each population 500 samples have been selected.

In the static approach three sub-populations are considered:

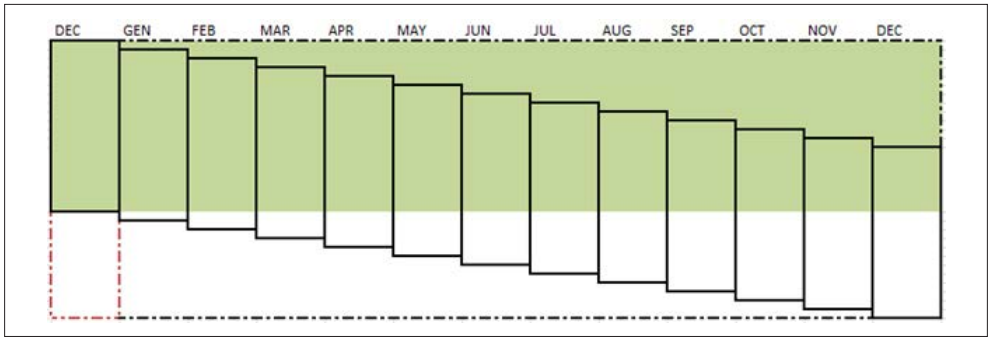
A. Panel series: all the products enter in the computation of indices (Scheme 8);
B. Static population without disappearing products: the products sold in December of the base year are followed during all the year. The new products are not considered by definition. However, disappearing products are assumed to be present (as if replaced) (Scheme 9).
C. "Pure" Static population: the disappearing products are not replaced and the new products are not considered by definition (Scheme 10).

The difference between C and B gives a measure of the error due to the shrinkage, while the difference between B and A quantifies the error due to ignoring the new products. The impact of temporary missing is derived comparing the values of C with and without temporary missing.

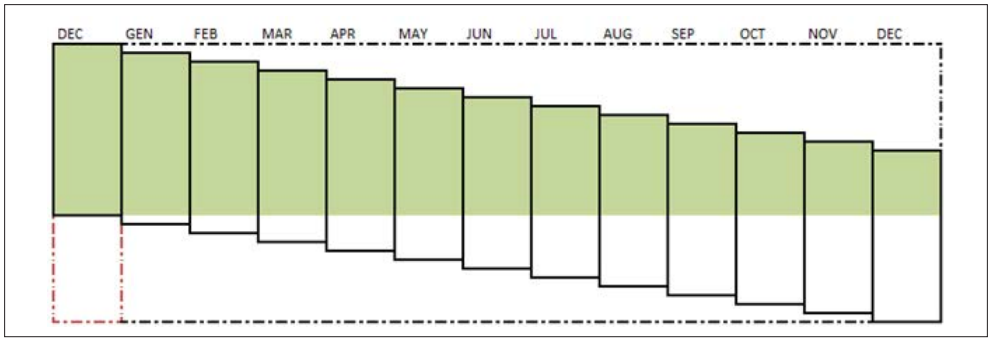
Scheme 8 - Panel series (A)



Scheme 9 - Static population without disappearing products (B)



Scheme 10 - "Pure" static population (C)



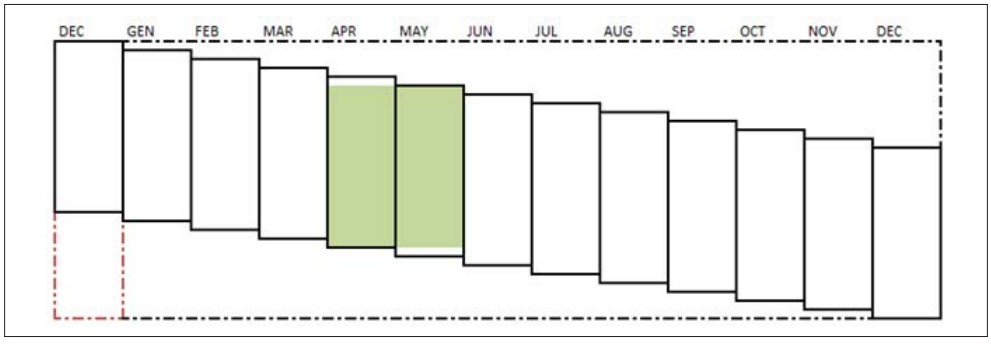
Under the dynamic approach, two sub-populations are considered:

A. Panel series. All the products enter in the computation of indices (Scheme 8).

D. "Pure" Dynamic population: Only the matched products in two months in a row enter in the computation of price indices (Scheme 11).

The difference between D and A can be seen as a difference between a static population and a dynamic one. In this case, the impact of temporary missing is derived by comparing the values of D with and without temporary missing.

Scheme 11 - "Pure" Dynamic population (D)



8.2 Main results

The most meaningful results of this experimental phase are shown to highlight the difference between the static and dynamic approach for the Jevons index.

The following Figure 18 shows the trend of the monthly Jevons index in the segments above listed and in scenarios C and A. The difference between C and A can be seen as a difference between a "pure" static population in which the disappearing products are not replaced and the new products are not considered and a static population (panel series).

Figure 18 shows that in the static approach the bias due to the shrinkage and to ignoring the entering products increases during the year, especially for more dynamic consumption markets.

The effects are small in the "Small semolina Pasta" market and do not affect the variability of estimates, instead, it seems to affect most "IGP-IGT Italian white wine" market that has medium dynamism but high variability in prices.

Figure 18 - Monthly Jevons indices in scenarios A and C for the consumption markets (static approach)

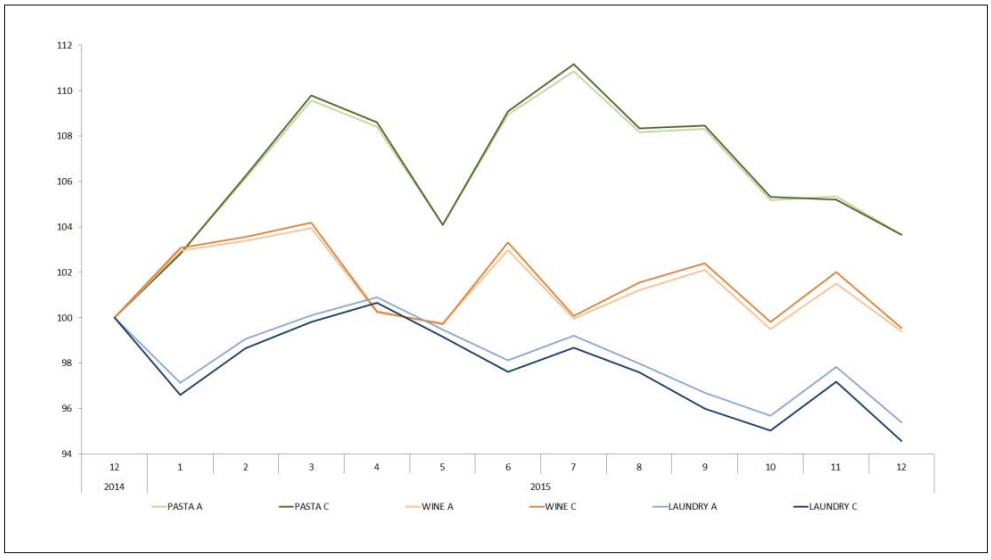
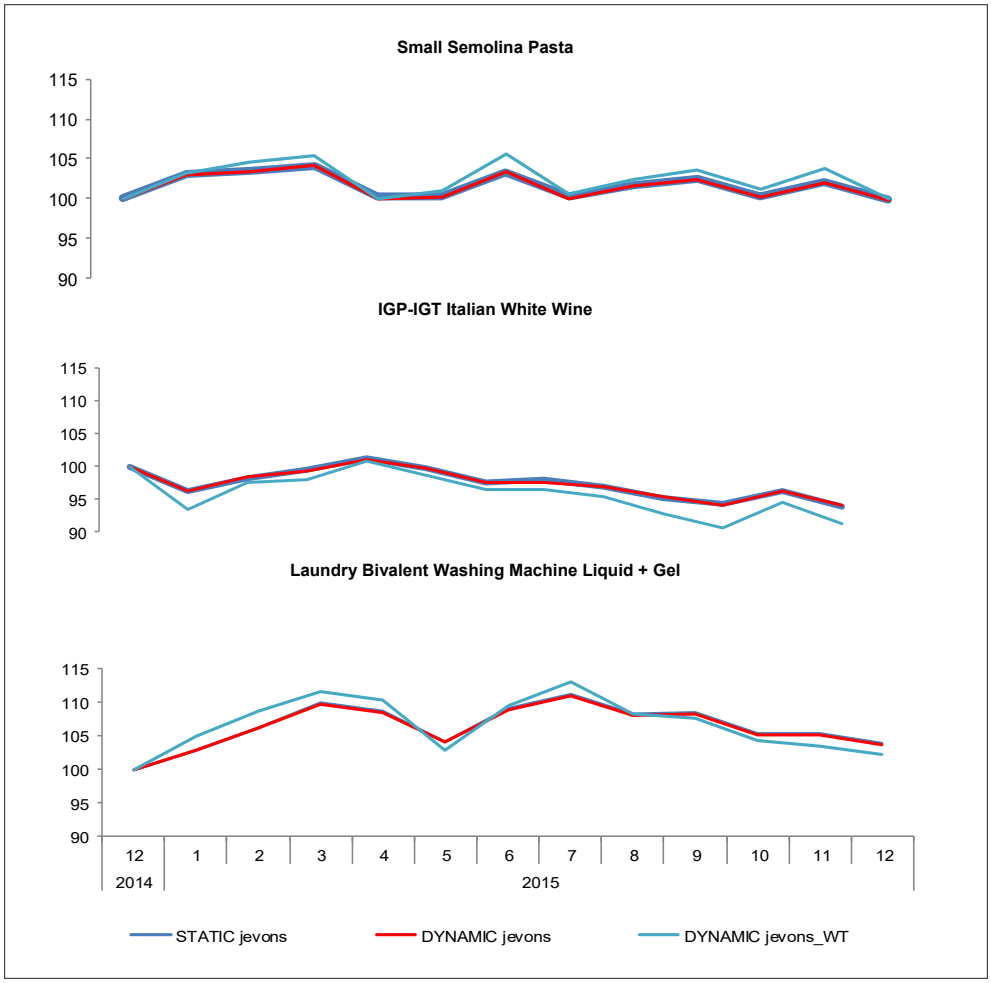


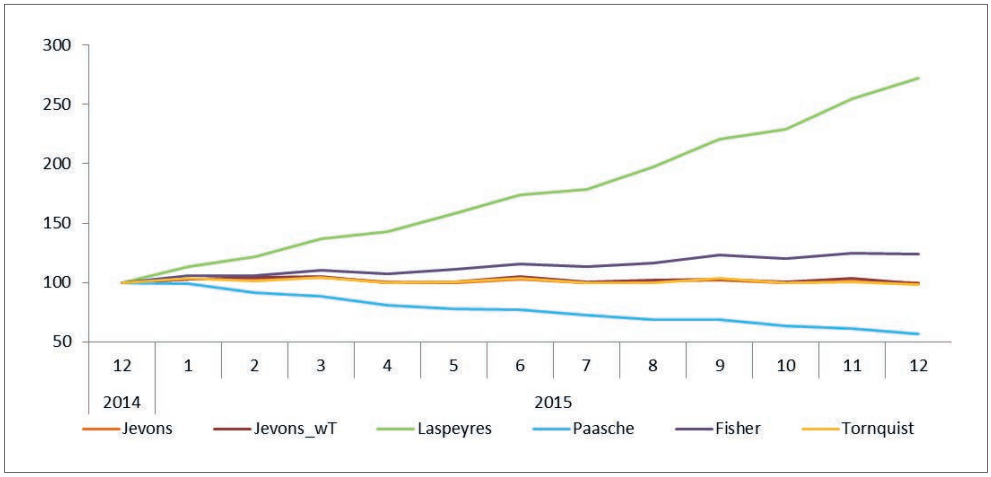
Figure 19 shows a tentative comparison among static and dynamic approach in a close population (scenario A) using the Jevons index with and without threshold. The threshold seems to have a not negligible impact on the levels of the index, mostly for the market with higher variability of prices (IGP-IGT Italian white wine).

Figure 19 - Monthly chained Jevons indices - with and without threshold - in scenarios A under the static and the dynamic approach



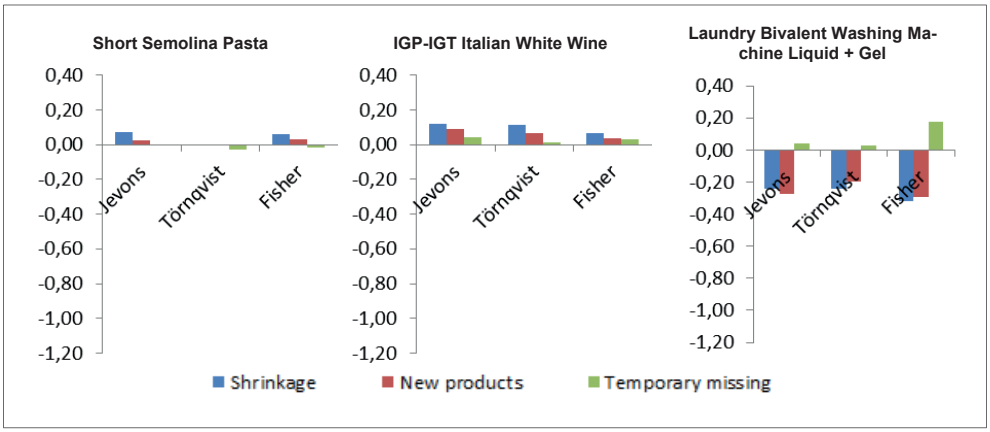
From Figure 20, where weighted and unweighted indices are compared for the "Short semolina pasta" market, it looks clear that the Jevons index does not suffer from chain-drift, while Laspeyres and Paasche index yes. While Jevons index – both with and without threshold – is stable during the year, Laspeyres and Paasche develop along divergent trends. Superlative Fisher and Törnqvist seem to be bounded between Laspeyres and Paasche values, even if Fisher index is not able to tone down the strong increase of the Laspeyres index and it affected by an upward chain-drift. Instead, the Törnqvist index seems to be nearer to Jevons.

Figure 20 - Chained weighted and unweighted price indices under the dynamic approach (Short semolina pasta market)



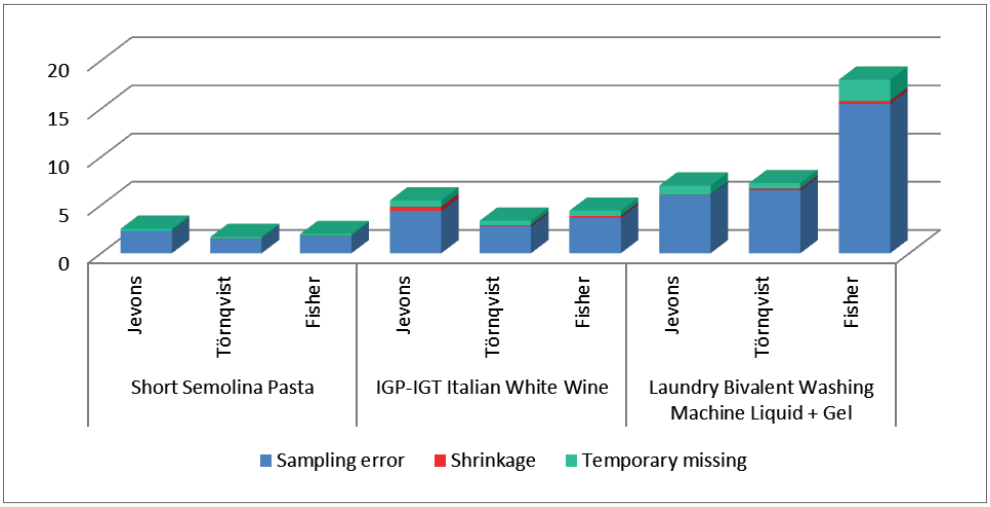
In the following figures, the different sources of bias and sampling variance are analysed in the considered scenarios and approaches.

Figure 21 - Relative percentage differences with respect to the «reference» value of price indices under the fixed approach



In the fixed approach, splitting the bias by sources, we can note in Figure 21 a different behaviour among indices and consumption markets. In general, the bias due to the shrinkage is higher than the bias due to excluding the new products and the presence of temporary missing. Furthermore, it increases with the dynamism of the consumption market.

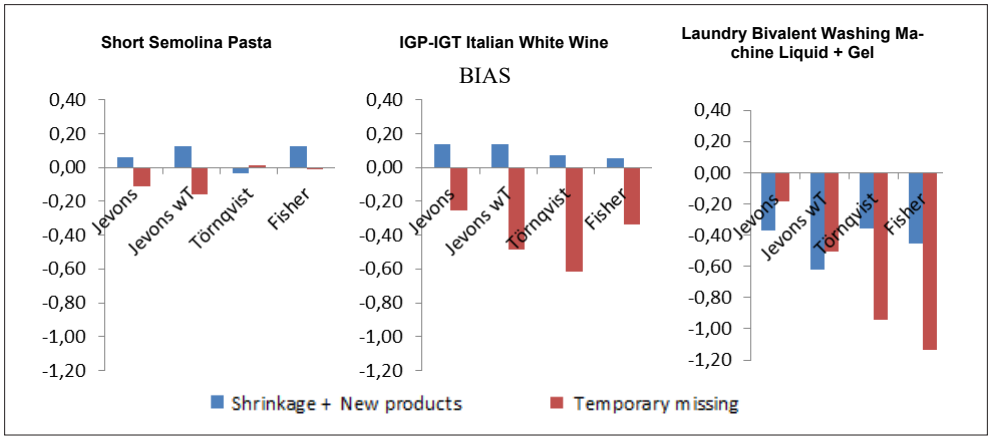
Figure 22 - The increase of percentage relative error due to the shrinkage of products during the year and to the temporary missing under the fixed approach



With respect to the overall variance, as shown in Figure 22, most part is due to the sampling. The shrinkage effect is small in the "Small semolina Pasta" market and does not affect the variability of estimates, instead it seems to affect most "IGP-IGT Italian white wine" market that has medium dynamism but high variability in prices. The bias due to the temporary missing affect more the market with higher dynamism ("Laundry Bivalent Washing Machine Liquid + Gel"). In general, also variance increases with the dynamism of the consumption market.

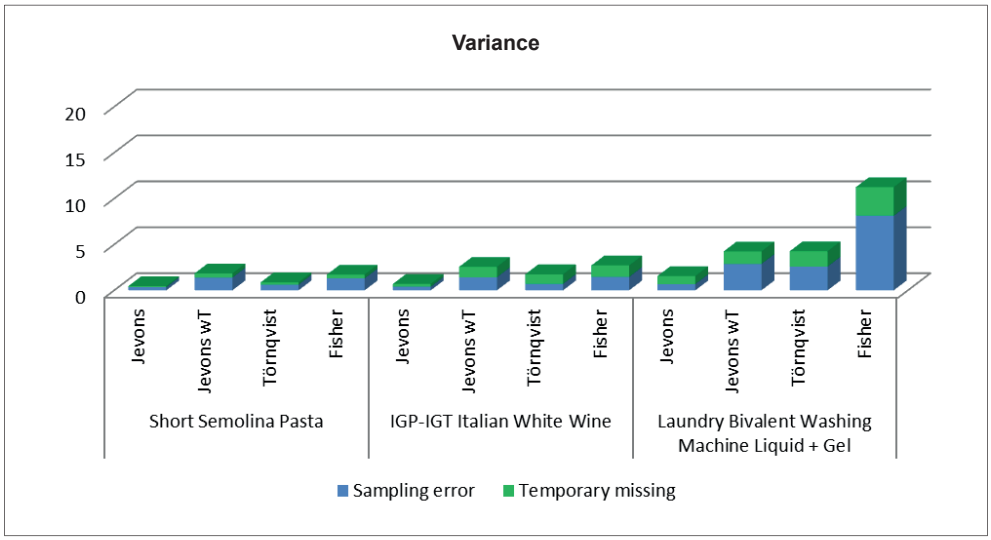
In the dynamic approach, the impact of shrinkage and new product on the bias (Figure 23) is similar to that one under the static approach, while that of temporary missing is higher. This is probably since the size of temporary missing included in the computation of the index under the dynamic approach is much larger than under the static approach.

Figure 23 - Relative percentage differences with respect to the «reference» value of price indices under the dynamic approach



The variance of the indices under the dynamic approach (Figure 24) is very low, due to the sampling design used (cluster instead of two-stage) and the large size of its sample. Also in this case, the variance increase when increasing the dynamism of the market, together with the variance due to temporary missing.

Figure 24 - The increase of percentage relative error due to the shrinkage of products during the year and to the temporary missing under the dynamic approach



In conclusion, also this study highlights the high heterogeneity of the market using scanner data. Sampling and non-sampling error depends on the aggregation formula used and the consumption market features. It seems that the estimators under the dynamic approach are affected by higher bias but much lower variability, especially sampling variability, with those computed under a static approach. This is due to the difference in sample size.

For the same reason, the impact in terms of bias of temporary missing is higher under dynamic approach than under static approach. Further studies can be addressed to derive the sampling and non-sampling error also of multilateral indices which could allow to overcome the chain drift and other issues of weighted indices in the dynamic approach.

9. Sample of outlets from the scanner data of grocery products

9.1 Scanner data in the process of estimation of inflation from 2018

Starting from January 2018 Istat introduces scanner data of grocery products (excluding fresh food) in the production process of estimation of inflation. This innovation concerns 79 indices of an aggregate of products belonging to 5 ECOICOP Divisions (01.02.05.09.12) (Istat, 2019).

In agreement with retail trade chains (RTCs) and with the collaboration of the Association of modern distribution and Nielsen, scanner data for 1,781 outlets (510 hypermarkets and 1,271 supermarkets) of the main 16 RTCs covering the entire national territory are monthly collected by Istat on a weekly basis at the item code level.

For 2018 the compilation of the CPI using scanner data is based on a fixed basket perspective: the use of these data has concerned some channels of the modern distribution, in particular hypermarkets and supermarkets of the main retail chains operating in Italy. Since 2020 Istat has extended the use of scanner data to other channels of the modern distribution (discounts, small sales areas and specialist drug) and has realised the transition to a flexible basket approach.

For the selection of the 2018 sample of outlets (hypermarket, supermarket) a probability design was implemented. Outlets were stratified according to provinces (107), chains (16) and outlet types (hypermarket, supermarket) in 888 strata. Probabilities of selection were assigned to each outlet based on the corresponding turnover value (potential). Concerning the selection of the sample of items, a static approach that mimics the traditional price collection method has been adopted. Specifically, a cut off sample of barcodes (GTINs) has been selected within each outlet/aggregate of products (covering 40% of turnover but selecting no more than the first 30 GTINs in terms of turnover). The products selected in December are kept fixed during the following year. A “tank” of potentially replacing outlets (258) and GTINs (until a coverage of 60% of turnover within each outlet/aggregate) has been detected in order to better manage the possible replacements during 2018.

For 2018, about 1,370,000 price quotes are collected each week to estimate inflation. For each GTIN, prices are calculated taking into account turnover and quantities (weekly price=weekly turnover/weekly quantities). Monthly prices are calculated with the arithmetic mean of weekly prices weighted with quantities. Scanner data (SD) indices of aggregate of products are calculated at outlet level as unweighted Jevons index (geometric mean) of GTINs elementary indices. Provincial SD indices of the aggregate of products are calculated with weighted arithmetic mean of outlet indices using sampling weights. Finally, for each aggregate of products, SD indices and indices referred to other channels of retail trade distribution are aggregated with weighted arithmetic mean using expenditure weights. The sampling design adopted for 2019 followed the same criteria adopted in the previous year for the definition of stratification and allocation but on an updated reference universe. The selection of the new sample was made by maximizing the overlap with the 2018 sample.

In 2020, the extension of the use of scanner data of grocery products to other channels of the modern distribution, including discounts, small sales areas and specialist drug, implied the definition of a new sampling design to take these new outlet types into account in calculating the CPI. In this case the same criteria used for the definition of sample of hypermarkets and supermarkets were adopted except for the stratification. Indeed, outlets were stratified taking into account only provinces and outlet types.

Finally, the overall sample for 2020 was obtained by selecting discounts, small sales areas and specialist drug, maintaining the same selection of hypermarkets and supermarkets used in the previous year. This choice was mainly determined by reasons related to the acquisition of scanner data. The final sample size is 4,073 outlets.

9.2 Sample of outlets from the Nielsen universe

9.2.1 Study of the variability of price indices

To allocate efficiently the sample of outlets (hypermarket, supermarket), the variability of the prices indices in the available 37 provinces and 6 chains has been studied.

Models relating provinces, chains and typology to variability of indices are studied to put more sample in the strata (province * chain * type) with higher variability of the parameter of interest.

The values of monthly price indices with Jevons, Laspeyres and Lowe aggregation formula have been computed for each available outlet of the 37 provinces and the distributions of the standard deviation of outlet monthly indices are analysed. The following figures show the standard deviation of three monthly price indices in different geographical areas.

At the regional level (Figure 25) the standard deviation of the monthly indices is small, except for Lombardia (due to the diversity among Milan and the other provinces) and Veneto (due to the diversity of Venice and the other provinces).

The standard deviation of monthly Jevons indices is more uniform compared to those of Laspeyres and Lowe.

Figure 25 - Standard deviation of monthly price indices by regions, 2015

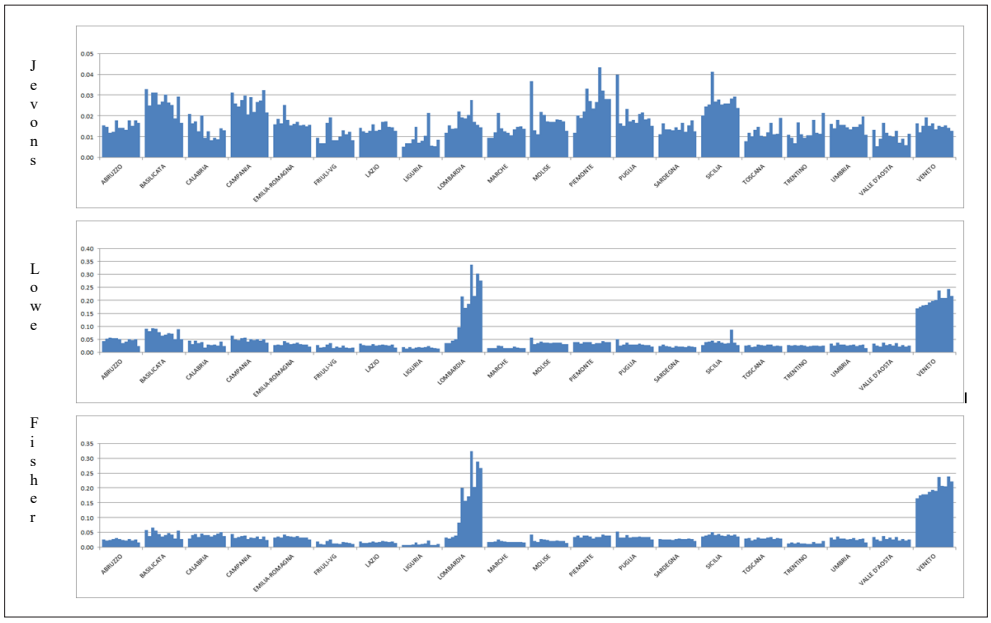


Figure 26 - Standard deviation of monthly indices by provinces, 2015

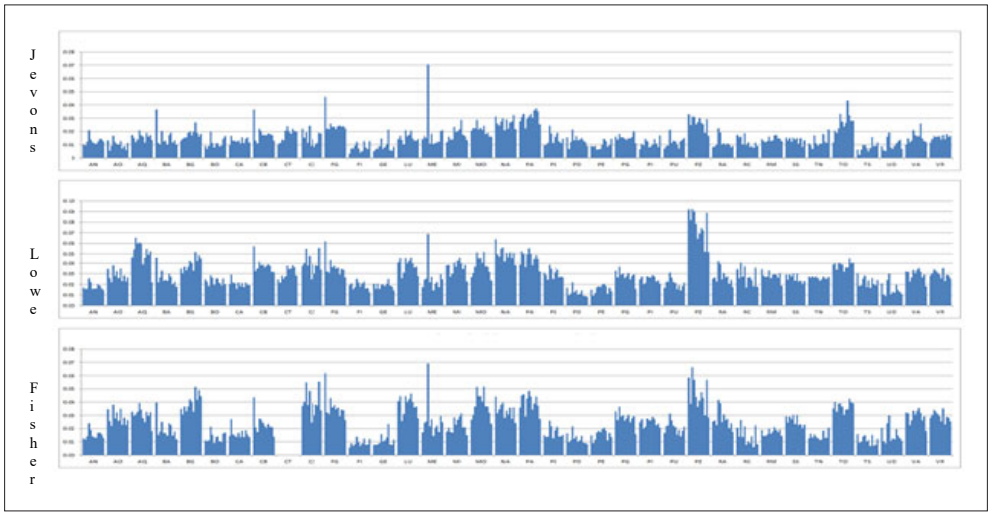
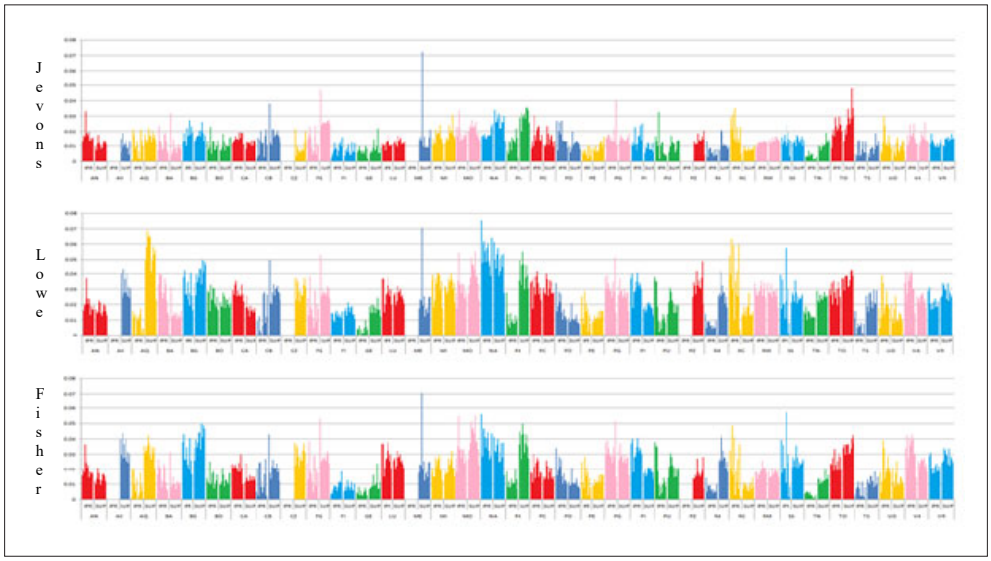


Figure 27 - Standard deviation of monthly indices by provinces and outlet type, 2015



At the provincial level, the standard deviation of the monthly indices increases a little (Figure 26). The three aggregation formulas provide the same results. There is a variability among provinces, but there is not a clear tendency we can use for improving the allocation.

Only in a few cases, the differences, with respect to the variability of monthly indices, between hyper and super-market are clear. But we still cannot assume this relation to address the allocation of outlets (Figure 27).

9.2.2 Sampling design for hypermarkets and supermarkets

The sample of outlets for 2018 was selected by the Nielsen universe represented by 16 chains (out of 25) that gave to Nielsen the authorisation to provide data to Istat, covering nearly 80% of total turnover of hypermarket and supermarket (but not all outlets are available) for in 2017.

Compared to the Nielsen Guide as a whole, considered as a theoretical universe, the sampling frame consists of only the outlets where Nielsen receives the elementary weekly data.

The outlets represent the Primary Stage-Units of the sample design for the selection of references. Outlets are stratified by province, chain and type (hypermarket, supermarket) and the selection is made using PPS, probability proportional to the potential of the outlet.

The objective followed in defining the sample design and the allocation criteria was to obtain a representative sample of the universe of the outlets available in a "neutral" way, *i.e.* not based on assumptions that if not verified could lead to a biased sample. In this way, it is possible to acquire information on the variability of the elementary indices for all the chains and all the provinces, to be used for the re-design of the sample with optimal allocation of the outlets in the strata.

Evaluations have been carried out on the exclusion of strata with very low weight, in terms of potential (turnover), within the province. The following Table 15 shows three hypotheses, with no threshold, with a threshold of 0.5% and 1%, the total number of strata and the number of strata with at least one outlet available.

Table 15 - Distribution of available outlets, strata and strata with available outlets by threshold

Threshold	Available outlet	N. strata	N. strata with available outlets
0	4,733	1,158	913
0.005	4,523	1,094	888
0.01	4,653	1,039	861

Based on the analysis the intermediate option was chosen, the threshold equal to 0.005 (0.5%), which allows at the same time to limit the number of excluded outlets and limit the number of strata with only one outlet.

With a threshold of 0.01, 80 available outlets would be excluded, of which 7 hypermarkets. Their potential is equal to 0.86% of the total, while with a threshold of 0.005, 35 available outlets would be excluded, of which 1 hypermarket (in Lombardy). Their potential is 0.16% of the total.

Allocation between provinces and strata

From the analyses carried out on the data available for the 37 provinces and 6 chains, no clear shreds of evidence emerged on the variability of the price indices that allowed identifying a relationship between variability and potential (the only variable available on all the outlets, including those belonging to the additional chains) from use as an allocation criterion also for the remaining provinces and chains.

For 2018, an allocation based on a compromise between proportionality criteria with respect to potential and with respect to the number of outlets in the strata is defined. The attribution of an excessive weight to the potential would have led to include almost all hypermarkets in the sample, limiting the number of supermarkets.

The allocation of 2,100 outlets, defined ex-ante, was obtained in two steps and is based on the total potential and number of outlets referring to the universe of outlets available and not (theoretical universe).

The number of outlets by the province has been defined on a compromise between three allocations:

- a uniform allocation among the 107 provinces (with a weight of 0.6);
- an allocation proportional to the number of outlets in the provinces (with a weight of 0.3);
- an allocation proportional to the total potential in the provinces (with a weight of 0.1).

The number of outlets by type (hypermarket, supermarket) and chain within each province have been defined based on a compromise between two allocations:

- an allocation proportional to the total number of outlet (with a weight of 0.65);
- an allocation proportional to the total potential of the PV (with a weight of 0.35).

The sample sizes per stratum were therefore defined on the theoretical universe and then adjusted to take into account the number of outlets available. Figure 28 shows the distribution of the sample of outlets in the Italian regions.

Figure 28 - Distribution of the sample of outlets (hypermarkets, supermarkets) by region. Year 2018

For 2019 the same sampling design was used updating only the reference universe (Nielsen 2018).

The selection of the sample of outlets (hypermarket, supermarket) was made by maximizing the overlap of the new sample with the sample of the previous year. The overlap achieved amounts to almost 85%.

The following table shows the distributions of the sample of outlets selected at the regional level.

Table 16 - Distribution of the sample of outlets (hypermarkets, supermarkets) by region. Year 2019

Region	N. outlet
Piemonte	169
Valle D'Aosta/*Vallée d'Aoste*	7
Lombardia	327
Trentino-Alto Adige/*Südtirol*	42
Veneto	177
Friuli-Venezia Giulia	75
Liguria	76
Emilia-Romagna	185
Toscana	168
Umbria	40
Marche	89
Lazio	137
Abruzzo	70
Molise	26
Campania	108
Puglia	105
Basilicata	16
Calabria	76
Sicilia	165
Sardegna	93
Totale	**2,151**

9.2.4 Sampling design for discounts, small sales areas and specialist drug

In 2019, Nielsen's updated the outlet list, which now contained also discounts, small sales areas and specialist drug, as well as hypermarkets and supermarkets, already present in previous years. The sample allocation of outlets, also for the 2020 sample, has been studied according to the potential.

The sample of new outlet types for 2020 was selected by the Nielsen universe (year 2019). Compared to the Nielsen Guide as a whole, considered as a theoretical universe, the sampling frame consists of only the outlets from which Nielsen receives elementary weekly data.

The outlets are the Primary Stage-Units of the sample design for the selection of references. Outlets are stratified by province and type (discounts, small sales areas and specialist drug) and the selection is performed using PPS, probability proportional to the potential of the outlet.

The sample size, defined ex-ante, of 4,000 outlets was obtained (analogously to what happened in the last survey) in two steps and is based on the total potential and number of outlets referring to the universe of outlets available and not (theoretical universe).

The number of sample outlets by province has been defined on the basis of a compromise between three allocations:

- a uniform allocation among the 107 provinces (with a weight of 0.6);
- an allocation proportional to the number of outlets in the provinces (with a weight of 0.2);
- an allocation proportional to the total potential in the provinces (with a weight of 0.2).

The number of outlets by type (discounts, small sales areas and specialist drug) within each province has been defined based on a compromise between two allocations:

- an allocation proportional to the total number of outlet (with a weight of 0.42);
- an allocation proportional to the total potential of the PV (with a weight of 0.58).

In the strata in which there are fewer outlets than those allocated, all units in the stratum have been selected.

The final sample consists of 1,951 outlets belonging to discounts, small sales areas and specialist drug. The following table shows the distributions of the sample of outlets selected at the regional level.

Table 17 - Distribution of the sample of outlets (discounts, small sales areas and specialist drug) by region. Year 2020

Region	N. outlet
Piemonte	136
Valle D'Aosta/*Vallée d'Aoste*	12
Lombardia	220
Trentino-Alto Adige/*Südtirol*	46
Veneto	126
Friuli-Venezia Giulia	60
Liguria	80
Emilia-Romagna	138
Toscana	154
Umbria	36
Marche	80
Lazio	134
Abruzzo	68
Molise	32
Campania	131
Puglia	141
Basilicata	39
Calabria	69
Sicilia	168
Sardegna	81
Total	**1,951**

10. Conclusions

The results of the experimental phases carried out using the first scanner data sets provided to Istat allowed us to individuate the most efficient sampling scheme for the selection of outlets and references in the static approach and to obtain a measure of sampling error and bias in the dynamic approach. The first experiments produced interesting results regarding the performance of sampling schemes and index formulas in a closed population context and fixed approach. They lead to the conclusion that probability sampling is the better choice in this context.

The possibility of switching to a dynamic approach requires, from an economic perspective, to deal with some complex issues. In fact, weighted indices would enable us to exploit better the potential of scanner data for the estimate of an elementary index, but they are affected by different drawbacks, first of all, the chain drift.

The international debate on the use of scanner data and therefore of a dynamic population approach is currently centred on the issue of how to resolve the chain drift problem related to the chaining of weighted price indices (also Fisher and Törnqvist), while at the same time maximizing the number of matches in the data. A solution to the chain-drift issue might be the construction of weighted transitive multilateral price indices, which are free from chain drift by definition, but other issues must be solved to ensure that previously published index numbers will not be revised. Chaining matched-model superlative indices are recommended in the ILO Manual (ILO, 2004) for the satisfactory properties that characterize them.

The outlined second experimental phase will provide evidences on the pros and cons of the two approaches, highlighting in particular empirical and theoretical drawbacks of the dynamic approach which is the one that Istat introduced in 2020. Therefore, the Institute has made a gradual transition from an approach based on a fixed basket to an approach based on a flexible basket. From this perspective, Istat is currently participating in the international debate focussed on the search for solutions that aim to the full use of the information contained in the SD and consequently to the use of weighted elementary price indices for the calculation of the CPI.

References

Anderberg, M.R. 1973. *Cluster Analysis for Applications*. Cambridge, MA, U.S.: Academic Press.

Chessa, A.G., J. Verburg, and L. A. Willenborg. 2017. "Comparison of Price Index Methods for Scanner Data". Paper presented at the *15th Meeting of the Ottawa Group*. Eltville, Germany, 10th - 12th May 2017.

de Haan, J., E. Opperdoes, and C.M. Schut. 1999. "Item selection in the Consumer Price Index: Cut-off versus probability sampling". *Survey Methodology*, 25(1): 31-41.

de Haan, J., and H.A. van der Grient. 2011. "Eliminating chain drift in price indexes based on scanner data". *Journal of Econometrics*, 161: 36-46.

de Haan, J., L. Willemborg, and A.G. Chessa. 2016. "An overview of price index methods for scanner data" (preliminary draft).

Feldmann, B. 2015. "Scanner data: current practice". Presentation at the *Workshop Scanner Data*. Roma, Italy: Italian National Institute of Statistcs – Istat, 1st - 2nd October 2015.

Gábor, E., and P. Vermeulen. 2014. "New evidence on elementary index bias". European Central Bank, *Working paper series*, N. 1754/ December 2014.

International Labour Office - ILO, International Monetary Fund - IMF, Organisation for Economic Co-operation and Development - OECD, Statistical Office of the European Communities - Eurostat, United Nations, and The World Bank. 2004. *Consumer price index manual: Theory and Practice*. Geneva, Switzerland: ILO Publications.

Ivancic, L., W.E. Diewert, and K.J. Fox. 2011. "Time Aggregation and the Construction of Price Indexes". *Journal of Econometrics*, 161(1): 24-35.

Nygaard, R. 2010. "Chain drift in a monthly chained superlative price index". *Workshop on scanner data*. Geneva, Switzerland, 10th May 2010.

Norberg, A. 2014. *Sampling of scanner data products offers in the Swedish CPI. Draft version 8*. Solna and Örebro, Sweden: Statistics Sweden.

Italian National Institute of Statistics - Istat. 2019. "Prezzi al Consumo. Dati defnitivi". *Statistiche flash*. Roma, Italy: Istat. https://www.istat.it/it/archivio/226109.

Rais, S. 2008. "Outlier detection for the consumer price index". In *Proceedings of the Survey Methods Section of the SSC Annual Meeting*. Ottawa, Ontario, Canada, 25th - 27th May 2008.

Rosén, B. 1997a. "Asymptotic theory for order sampling". J*ournal of Statistical Planning and Inference*, 62(2): 135–158.

Rosén, B. 1997b. "On sampling with probability proportional to size". *Journal of Statistical Planning and Inference*, 62(2): 159–191.

Saidi, A., and S. Rubin Bleuer. 2010. "Detection of outliers in the Canadian consumer price index". Paper presented at the *Conference of European Statisticians*. Ottawa, Ontario, Canada, 16th - 18th May 2015.

Sampford, M.R. 1967. "On sampling without replacement with unequal probabilities of selection". *Biometrika*, 54(3-4): 499–513.

Van der Grient, H., and J. de Haan. 2010. "The Use of Supermarket Scanner Data in the Dutch CPI". Paper presented at the *Joint ECE/ILO Workshop on Scanner Data*.

Van der Grient, H., and J. de Haan. 2011. "Scanner Data Price Indexes: The "Dutch" Method versus Rolling Year GEKS". Paper presented at the *12th Meeting of the Ottawa Grou*p. Wellington, New Zealand, 4th - 6th May 2011.

Vermeulen, B.C., and H.M. Herren. 2006. "Rents in Switzerland: sampling and quality adjustment". Paper presented at the *11th Meeting of the Ottawa Group*. Neuchâtel, Switzerland, 27th – 29th May 2006.

An imputation procedure for the Italian attained level of education in the register of individuals based on administrative and survey data

Marco Di Zio, Romina Filippini, Gaia Rocchetti [1]

Abstract

The paper describes the mass imputation procedure of the level of education in the Base Register of Individuals of the Italian National Institute of Statistics – Istat. The procedure integrates data of different nature: information deriving from administrative sources, from the 2011 population census and from the 2018 permanent census survey. The procedure is complex and is composed of different steps depending on the information of the sources. The imputation is based on log-linear models which, compared to classical methods such as the hot-deck imputation, allow greater flexibility in modelling associations. The work also illustrates the comparisons between the register estimates obtained with imputation with those of the census sample survey in order to highlight the advantages and limitations of the proposed procedure.

Keywords: statistical register, data integration, mass imputation.

1 Marco Di Zio (dizio@istat.it); Romina Filippini (filippini@istat.it); Gaia Rocchetti (grocchetti@istat.it), Italian National Institute of Statistics – Istat.

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.

1. Introduction[2]

The Italian Base Register of Individuals (BRI) is a comprehensive statistical register storing data gathered from various data sources. In BRI, core variables as place and date of birth, gender, citizenship are associated to each unit. Moreover, a classification variable denoting people resident in Italy is introduced. The subset of resident people is the basis of the next Italian census that will be as much as possible register-based. According to this idea, given the high amount of available administrative information, a prediction of the attained level of education for the resident people in BRI is proposed.

The main sources containing administrative information originate from the Ministry of Education, Universities and Research (MIUR). MIUR provides information about the attained level of education and student's attendance to a course (*e.g.* attending in the first year of primary education). Administrative data refer to students from 2011 onwards. For the rest of the people not included in this period, we may resort to the 2011 Census information. Unfortunately, not all the people classified as resident after 2011 belong to these two data sets, as for instance immigrated people entered Italy after the Census that have not attended any educational course. Another important source of information is the sample survey collected for the permanent Italian census starting from 2018. These data are particularly important to fill the informative gaps of MIUR and 2011 Census data. We remind that the so-called permanent census is a system of yearly surveys and administrative data organised in registers that once combined are supposed to provide each year the main Census figures.

The focus of the work is on the prediction or mass imputation (in this application, we adopt the two terms as synonymous) of the attained level of education in the Base Register of Individuals. A mass imputation procedure is justified by the high amount of detailed available information. Similar studies are available in other NSIs, see for instance Scholtus and Pannekoek (2015), Daalmans (2017).

2 Although the article is the result of a joint work, the single parts are authored as follows: Sections 1 and 5, Paragraphs 3.1 and 3.3 by Marco Di Zio; Section 2 and Paragraphs 2.1, 2.2 and 2.3 by Gaia Rocchetti; Paragraph 3.2 and Section 4 by Romina Filippini.

The procedure discussed in the paper follows the study by Di Cecco *et al.* (2018) where different methods were evaluated on preliminary data. The procedure chosen for the prediction of level of education is applied to the 2018 BRI and, although the 2018 sample survey is not yet completely cleaned, we expect survey data to be close enough to the final ones that will be used for producing official estimates.

The paper is structured as follows. Section 2 depicts the informative context describing the data sources used for the prediction. The imputation procedure is presented in Section 3, and some results of the analysis carried out in order to assess the outcomes of the procedure are reported in Section 4. Section 5 presents some final remarks and future developments.

2. Informative context

2.1 Data sources description

In carrying out the prediction procedure, data of different nature are jointly used. In fact, the procedure combines administrative, traditional Census and sample survey data. For this exercise, the procedure is applied to a preliminary version of data for which the population in BRI considered as usually resident in Italy at 31st of December 2018 amounts to 60,433,360 units. BRI also includes the main personal information, *i.e.* the core variables – place and date of birth, gender, citizenship – used in the present study. Those core variables are obtained through an extensive utilisation of administrative data, reconciled and stored yearly in the BRI.

Administrative information on the attained level of education (ALE hereinafter) is gathered making use of the information collected by the Ministry of Education, University and Research and processed in Istat with the purpose of creating a database on Education and Qualification, named BIT (see Runci *et al.*, 2017). BIT collects, checks and integrates data from different sources provided by MIUR, on a yearly basis, about the ALE and the attendance to a course (*e.g.* attending in the first year of primary education) of students. Data on the ALE is available at the reference time, set on 31st of December 2017; meanwhile, data on the attendance to a course refer to the academic year 2017/2018 (BIT 2017 hereinafter). Summarizing, BIT 2017 collects information on the ALE achieved between 2012 and 2017 for 13,966,581 units.

People, that have not attended any course since 2011, are not in BIT. For our purposes, we turn to data from 2011 Census to fill the gap. The 2011 Census operations, whose reference date was 31st of October 2011 (CENS 2011 hereinafter), surveyed 59,433,744 individuals. For our needs, data on educational attainment was collected for persons aged 9 or older, who were still living in Italy on the 31st of December 2018, for a total of 53,745,821 units.

Another important source of information is given by the 2018 sample survey conducted for the permanent Italian census. In this sample, units are

asked about their educational level. More precisely, survey data used for the prediction are obtained by the integration of the list and the area samples. They approximately amount to 5% of the total population.

In addition, auxiliary administrative data on ALE can be taken from the registration and cancellation forms for transfer of residence (APR4) gathered for the period 2012-2017. ALE on APR4 is self-declared by individuals that fill the form in order to apply for a new registration in Italy coming from abroad and/or when they change usual residence. In APR4, ALE comes with 4 levels of classification:

1. Up to the elementary license corresponding to ISCED[3] 0, 1;
2. Lower secondary education corresponding to ISCED 2;
3. Secondary and short cycle tertiary education corresponding to ISCED 3, 4, 5;
4. Tertiary and post tertiary education - ISCED 6, 7, 8.

2.2 Reconciling classifications and computing ALE from administrative sources

Both CENS 2011 and BIT 2017 use detailed and reciprocally consistent classifications of educational level; consequently, data were univocally reclassified according to 8-items dissemination classification (named CDIFF) adopted by Istat for the purpose of disseminating permanent census data on the ALE. In particular, mapping operations are carried out such that items in the classifications adopted by CENS 2011 (12 items and a separate question for those having obtained a doctoral or equivalent level) and BIT 2017 (16 items) could be homogenously reclassified into the new one (17 items; Istat 2017 hereinafter). Furthermore, we univocally recode data into the CDIFF classification (Table 2.1).

3 ISCED (International Standard Classification of Education) is a statistical framework created by UNESCO for organising information on education (http://uis.unesco.org/).

Table 2.1 - Correspondence table between Istat 2017 and CDIFF 2018 classifications on ALE

CDIFF 2018 classification	BRI and Survey Sample 2018 for the permanent census 2020 classification
1 - Illiterate	01) Illiterate
2 - Literate but no formal educational attainment	02) Literate but no formal educational attainment
3 - Primary education	03) Final assessment (Primary school)
4 - Lower secondary education	04) Diploma of lower secondary education
5 - Upper secondary education	05) Diploma of upper secondary education (2-3 years) 06) IFP - Vocational training qualification (three-year courses)/ Professional diploma (fourth year) 07) Diploma of upper secondary education (4-5 years) 08) Certification of higher technical specialisation (IFTS)
6 - Bachelor's degree or equivalent level	09) Diploma of Higher Technical (ITS) 11) University diploma 12) Fine Arts, Drama, Dance and Music First level academic diploma (Bachelor's degree) 13) *Laurea triennale* (I level, Bachelor's degree)
7 - Master's degree or equivalent level	10) Fine Arts, drama, Dance and Music Diploma (2-3 years) 14) Fine Arts, Drama, Dance and Music Second level academic diploma (Master's degree) 15) *Laurea (4-6 years*, Master's degree) 16) *Laurea biennale specialistica* (II level, Master's degree)
8 - PhD level	17) Research Doctorate (PhD)/ Advanced research academic diploma

Source: Istat

It is worth noticing that the choice of using both CENS 2011 and BIT 2017 comes from a comprehensive data quality analysis (see Di Cecco *et al.*, 2018). Here, for our purpose, we shortly present the principal results on data consistency, based on a cross-comparison at a micro level.

Table 2.2 - Consistency of data on ALE in CENS 2011 and BIT 2017

	a.v.	%
BIT 2017 > CENS 2011	7,705,099	80.2
BIT 2017 = CENS 2011	1,763,050	18.3
BIT 2017 < CENS 2011	144,332	1.5
Total RBI 2018 population aged >8 years	9,612,481	100.0

Source: Istat

Table 2.2 shows that out of about 9.6 millions of individuals co-present in the two datasets, 18.3% shows the same level of education in both sources. Moreover, 7.7 million (80.2%) gained a higher degree than observed in CENS 2011. The remaining 1.5% - almost 144 thousands population units – instead, reports inconsistent data, being the most recent level of education lower than the one assigned in CENS 2011.

The reasons of such inconsistencies are not easily identifiable. They are probably due to response errors. For instance, as far as cases in which BIT 2017 data are lower than data registered in CENS 2011 operations, the majority of cases concerns units reporting a "Diploma of upper secondary instead" of a "Diploma of lower secondary education", or a "Laurea triennale (I level, Bachelor's degree)" instead of a "Laurea (4-6 years, Master's degree)". To some extent, they may also be caused by linkage errors.

In order to reconcile the information, in the case of two different information on ALE coming from the two different sources, we replaced CENS 2011 data with BIT 2017 data. In fact, not only data that MIUR provided on yearly basis are usually reliable but also the process leading to the construction of BIT is a well-established one and is characterised by high quality standards (see Runci *et al.*, 2017).

2.3 Coverage and characteristics of subpopulation segments

An important aspect to analyse when using administrative data is the coverage of the sources, in fact they generally focus on specific populations. Table 2.3 classifies target population in main subgroups categorised by presence or absence of information on educational attainment. Data from BIT 2017 covers 22.1% of the overall BRI 2018 population; instead, people observed not in BIT but in CENS 2011 provides the most consistent part of coverage (67.7%). As far as we consider information available for people aged at least 9 years, the total coverage of administrative data is about 95%.

Table 2.3 - Reference population by presence in CENSUS 2011 and BIT 2017

	Total population		Aged 9 years and over	
	a.v.	%	a.v.	%
Present in BIT 2017	13,388,736	22.1	12,292,304	22.0
Present in CENS 2011 only	40,931,241	67.7	40,931,231	73.2
Records without information on ALE	6,113,383	10.1	2,685,623	4.8
Total BRI 2018 Population	60,433,360	100.0	55,909,158	100.0

Source: Istat

Despite the high coverage rate of administrative and Census data, there are still about 2.7 million of eligible units older than 9 years without data on ALE. These are either people entered Italy after 2011 that have not attended any course covered by MIUR, or people not caught by the 2011 Census. Concerning the latter, during post-Census operations, the collaboration with municipalities (named SIREA operation) allowed to identify 1,403,991 individuals who could not be found in CENS 2011 but that were resident: they were "detected" for the purpose of counting resident population but they did not answered the questionnaire.

An in-depth analysis shows that these three groups of population - namely CENS 2011, BIT and people without any official administrative information on ALE – have slightly different distribution for what concerns principal core variables. Table 2.4 shows, in fact, that people without administrative and Census data on ALE are, on average, older. In more detail, individuals without information on ALE show higher percentages in the age classes 29 -39 years (29.9% as against 13.8% of total BRI population) and 40-49 years (21.6% as against 16.6% of total BRI population).

Table 2.4 - Age distribution in BRI by data sources: BIT 2017, CENSUS 2011, and records without information on ALE

	BIT 2017	CENS 2011	Records without information on ALE	BRI 2018 Population aged at least 9 years	
Age	%	%	%	%	a.v.
9-10	9.0	0.0	1.6	2.0	1,145,028
10-11	13.6	0.0	1.7	3.1	1,720,250
14-18	23.0	0.1	1.3	5.2	2,889,161
19-22	18.1	0.1	4.6	4.3	2,385,385
23-25	11.8	0.6	4.9	3.3	1,820,519
26-28	8.8	1.7	6.6	3.5	1,943,501
29-39	12.0	13.3	29.9	13.8	7,732,149
40-49	2.4	20.5	21.6	16.6	9,264,211
50-69	1.4	39.0	22.6	29.9	16,725,020
70+	0.0	24.8	5.2	18.4	10,283,934
Total	100.0	100.0	100.0	100.0	55,909,158

Source: Istat

Moreover, in the subpopulation of individuals without any information on ALE there are more male than female (52.5% of male vs. 48.5% in BRI population) and mostly a dramatic larger percentage of Not Italian: 67.7% against 8.3% in the total BRI 2018 population (see respectively Table 2.5 and 2.6).

Table 2.5 - Gender distribution in BRI by data sources: CENSUS 2011, BIT 2017 and records without information on ALE

	BIT 2017	CENS 2011	Records without information on ALE	BRI 2018 Population aged at least 9 years	
Gender	%	%	%	%	a.v.
Male	50.2	47.7	52.5	48.5	27,104,126
Female	49.8	52.3	47.5	51.5	28,805,032
Total	100.0	100.0	100.0	100.0	55,909,158

Source: Istat

Table 2.6 - Italian/Not Italian distribution in BRI by data sources: CENSUS 2011 and BIT 2017 and records without information on ALE

	BIT 2017	CENS 2011	Records without information on ALE	BRI 2018 Population aged at least 9 years	
Citizenship	%	%	%	%	a.v.
Italian	94.1	94.8	32.6	91.7	51,252,687
Not Italian	5.9	5.2	67.4	8.3	4,656,471
Total	100.0	100.0	100.0	100.0	55,909,158

Source: Istat

2.4 Informative gaps and the use of auxiliary data

Data sources have some informative gaps. BIT, having MIUR data as exclusive source, reports only information for students that during the period 2012-2017 have enrolled a course that MIUR formally recognises. In particular, MIUR takes into account only courses supplied by an Italian qualified Institution on the Italian territory (*i.e.* International Institutions operating in the Country are not included). Furthermore, it is worthwhile to remark that BIT does not include qualification courses like Fine Arts, Drama, Dance and Music academic diplomas and more relevantly training and vocational careers managed by Italian Regions that are not required to provide data to MIUR. The main consequence is an underestimation of the level of education, also for the units in the subset reporting 2011 Census ALE. In fact, the imputation procedure associates the potential lower ALE registered in CENS 2011 to those units that, during the period 2012-2017, had been enrolled in a couse not registered in BIT. As a consequence, for all population units for which schooling or training is over, it has to be experimented the use of auxiliary information.

Another critical issue concerning BIT has to do with timeliness. The lag between the moment in which BIT data are available and the BRI reference time makes it necessary to implement procedures for the prediction of the variable. BIT data are available with a delay of 12 to 24 months respect to the reference time. As shown afterwards, the attained level of education at time t should be predicted by having available one-year lagged data; information of attendance of educational courses has instead a lower delay, being related to the academic year [t-*12* months, t].

We need to resort to additional data to both fill the informational and temporal gap. The core information comes from the survey data collected during the permanent census operation in October 2018. The sample survey gathers information for about 2.6 million of units (Table 2.7). The sample survey ALE has been originally classified according to the Istat 2017 classification and has been recoded to the 8-items of dissemination for the purpose of prediction ALE for 2018. As Table 2.7 shows, information gathered by the sampling survey operation mostly overlap with ALE data coming from CENS 2011 (74.3%), though a 3.6% could help fillling the information gap for records lacking ALE.

Table 2.7 - Sample 2018 population by presence in CENSUS 2011 and BIT 2017

	Sample 2018		BRI 2018 Population aged at least 9 years	
	a. v.	%	a.v.	%
Present in BIT 2018	12,258,146	22.1	12,292,304	22.0
Present in CENS 2011 only	41,132,566	74.3	40,931,231	73.2
Records without information on ALE	1,977,346	3.6	2,685,623	4.8
Total	55,368,058	100.0	55,909,158	100.0

Source: Istat

The additional auxiliary administrative information on ALE from APR4 forms allows collecting data on about 5.2 million of units (Table 2.8). It is worth noticing that though APR4 data mostly overlap data from CENS 2011, 19.8% of observations covers the segment of population without any administrative information on ALE. This subpopulation is composed by either people more inclined to move across the Country, or entered Italy from abroad during the period 2012-2017. Thus, it is likely that the subpopulation without ALE is less “detectable” than the rest on the BRI 2018 units, and this can be the reason of the underestimation of the sample survey reported in Table 2.7.

Table 2.8 - APR4 form data by data sources in BRI: BIT 2017, CENSUS 2011 and records without information on ALE (row percentages and total absolute values)

	BIT 2017	CENS 2011	Records without information on ALE	BRI 2018 Population aged at least 9 years
No APR4 data	22.2	74.6	3.3	50,702,985
With APR4 data	20.2	60.0	19.8	5,206,173
Total	22.0	73.2	4.8	55,909,158

To conclude, it is worth noticing that the nature of APR4 data is different both qualitatively and in substance, since it is self-declared information and never submitted to the standard editing/quality control procedures. However, a preliminary consistency analysis of information for individuals presenting data on attained ALE computed on administrative data (that is CENS 2011 updated with BIT 2017 data) and APR4 (4,175,256 observations) shows that APR4 presents a sufficient degree of consistency on the level of education 1 - Up to elementary license (83.9%) though it is decidedly lower for the other levels, with slightly higher percentages (67.7%) in 4 - Tertiary and Post Tertiary Education (see Table 2.9).

Table 2.9 - ALE in 2017 (administrative data) and in APR4 (row percentages and total absolute values)

	ALE in APR4 (2012-2017)				
	1 - Up to Primary education	2 - Lower secondary education	3 - Secondary and short cycle tertiary education	4 - Tertiary and post tertiary education	Total a.v.
1 - Up to Primary education	83.9	11.6	3.6	0.9	599,765
2 - Lower secondary education	32.8	52.8	12.5	1.8	1,281,587
3 - Secondary and short cycle tertiary education	20.2	15.3	57.2	7.3	1,536,681
4 - Tertiary and post tertiary education	16.4	3.9	12.0	67.7	757,223
Total	32.5	24.2	27.6	15.7	4,175,256

Source: Istat

3. Imputation of the attained level of education

3.1 The imputation procedure

In this section, we illustrate a procedure for the prediction of the attained level of education at the reference year *t* of the resident population in BRI. At time *t*, the BRI contains the following structural information:

- The resident population at 31/12/t;
- Gender - (G);
- Date of birth - (D);
- Place of birth - (P);
- Country of citizenship at 31/12/t.

From the MIUR administrative data, we have used:

- the attained level of education at 31/12/t-12 months - (I^{t-12});
- the year attendance of educational courses in the time period [t-12, t], *e.g.* 1st year, 2nd year,.. - (F^t);
- the type of school (liceo, other) – (L).

We remind that the year of attendance of previous years as [*t*-24, *t*-12] and so on are available as well.

From the APR4 administrative data, we have exploited the self-declared ALE (I^{apr}) with 4 levels of classification as detailed in Section 2.

For the application of the procedure, the following transformed variables are also computed:

- Italian, not Italian citizenship – (C^t);
- Age in 8 classes - (E8).

As aforementioned, in addition to administrative data, we may resort to information on the ALE from the 2011 Italian Census and from a sample survey referring to the target time *t*. We notice that for units not in MIUR but in the Census 2011, the ALE at time *t*-12 (I^{t-12}) is the one reported in the 2011 Census. Finally, we denote with I^tS the ALE at time t observed in the sample survey.

Let I^t be the target variable, *i.e.* the ALE at time t that we would like to predict. We denote with A the subset of data for which information from MIUR is available, with B the set of units for which only information from Census 2011 is available and with C the subset of data observed neither in the Census nor in MIUR. Table 3.1 depicts the data/information scenario that we need to take into account when making predictions for I^t. Grey cells represent missing data and the last column shows the relative frequencies of groups with respect to the population with at least 9 years.

Table 3.1 - Tabular representation of the informative context for mass imputation of the attained level of education at time *t*

X_{BRI}				X_{miur}				Sample	Prediction	Group
G	E	P	C^t	$L^{(t)}$	$I^{(t-12)}$	$F^{(t)}$	I^{apr}	I^t_S	I^t	
										A 22%
										B 73%
										C 5%

Source: Istat

3.2 Mass imputation process flow

Groups A, B and C are characterised by different patterns of available information which determine different models for the estimation of ALE in 2018. In particular, ALE estimation may be either deterministic or probabilistic. The overall process of data treatment, model estimation and ALE imputation in the three groups A, B and C is summarised in Figure 3.1.

The main difference is between group A and the others. Group A is composed by "Active" people, which are attending a course in academic year t-12/t, while groups B and C are "Inactive" people, not attending any course in the same period, which is the last available from administrative sources.

In group A, administrative data provide longitudinal information on school enrollment. Thanks to the great information capacity of these administrative data, it is not necessary to resort to ALE observed in the 2018 sample. Information on ALE in the year t-12 (I^{t-12}) and information on year attendance of educational courses in academic year t-12/t (F^t) are available for all the individuals in group A. This allows identifying the probability of obtaining a new qualification based on schooling charachteristics of each individual.

Out of them, a subset of individuals with a zero probability of changing the educational level, from t-12 to t, is identified. Therefore, for this subset of "No-Change" people, it is not necessary to estimate a model for the imputation of ALE, since ALE in 2018 (I^t) is equal to ALE in 2017 (I^{t-12}).

The subset of "No-Change" is identified by one of the following conditions:

1. attending year 1, 2, 3 or 4 of primary school (Primary education is acquired at the end of year 5);
2. attending year 1 or 2 of lower secondary school (Lower secondary education is acquired at the end of year 3);
3. attending year 1, 2 or 3 of upper secondary school (Upper secondary education is acquired at the end of year 5; in high school you can attend two years in one);
4. attending upper secondary school and still having an Upper secondary education;
5. enrolled in a first level university course and still having a Bachelor's or a Master's degree;
6. enrolled in a university course and still having a Master's degree;
7. attending year 1 of a PhD course or still having a PhD.

People of group A, who do not meet any of the above conditions, have a non-zero probability of obtaining a higher qualification than that held in year t-12. For each individual of this "Change" subset the estimate of the probability distribution of achieving a new qualification in time t is based on individual characteristics and school attendance in academic year t-12/t (F^t). The model is estimated using only administrative sources. The underlying hypothesis is that the probability of obtaining a higher qualification between the years t-12 and t is equal to that between the years t-24 and t-12.

On the other side, group B and C are composed by "Inactive" people, this means people not enrolled in any course covered by MIUR in academic year t-12/t. Due to some informative gaps in administrative sources (see Section 2.3), there is a non-zero probability that an individual belonging to these groups is either enrolled in academic year t-12/t or has been enrolled in previous academic years in a school course not covered by MIUR.

For people in group B, information on previous educational level is available from administrative sources or from data collected in the 2011 Census.

For people interviewed in the 2011 Census who was enrolled in a school course covered by MIUR between 2011 and 2016 (but not in 2017/2018), the most updated information on ALE comes from MIUR. This subgroup is composed of individuals on average younger, who have recently dropped out of a school course covered by MIUR.

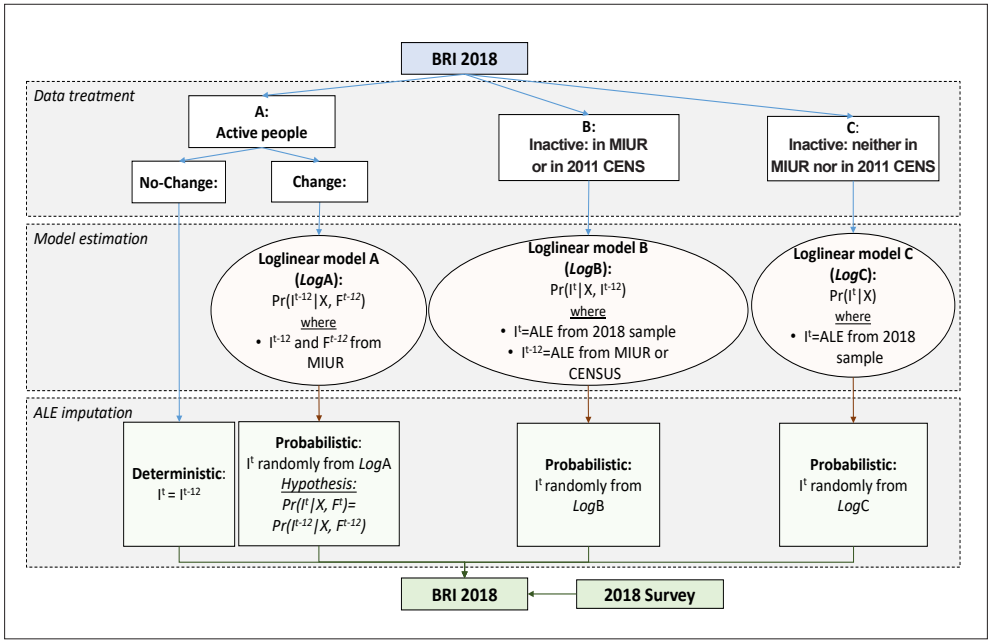
On the other hand, for people not enrolled in any school course after 2011, the only available information on ALE refers to 2011. They are mainly adults, long since out of school and probably less likely to change their educational level.

In both cases, the available information on ALE may not be error free due to coverage error (MIUR) or response error (2011 Census). For this reason, the model is estimated on units interviewed in the 2018 survey using the observed ALE as target variable. However, due to their different characteristics, individuals with information on ALE from MIUR or from 2011 Census are treated separately in the estimation process.

For people in group C, no information on ALE is available neither from MIUR nor from 2011 Census, so it is necessary to estimate a probability distribution of ALE for each pattern of available information on individual characteristics. ALE observed in 2018 survey is considered as target variable.

As a last step of imputation, for all the individuals observed in the 2018 sample, the observed ALE is directly used as prediction in BRI.

Figure 3.1 - Imputation process flowchart



Source: Istat

3.3 Model estimation and imputation

The general idea is to estimate a model for the prediction of I^t given the values of known covariates X. In particular, we estimate the conditional probabilities $h(I^t|X)$ and then impute I^t by randomly taking a value from this distribution. The conditional probabilities $h(I^t|X)$ are estimated by means of hierarchical log-linear models as follows. First, a log-linear model is applied to the contingency table obtained by cross-classifying the variables (I^t, X) to estimate their expected counts $\hat{N}(I^t,X)$ from which we can compute the counts $\hat{N}(X)$. The estimated conditional probability distribution $\hat{h}(I^t \mid X)$ is easily obtained by computing $\hat{N}(I^t, X)/(X)$. This approach includes as a special case the random hot-deck when a saturated log-linear model is assumed, but it has the advantage of allowing the use of more parsimonious model as well. This is an important characteristic especially when the number of variables and of the contingency table cells increase.

In order to take into account the missing data mechanism, sampling weights adjusted for non-response (that is indeed low in this survey) are used. It is adopted a pseudo-maximum likelihood approach that consists in estimating log-linear models on weighted count data (Thibaudeau *et al.*, 2017, Skinner *et al.*, 2010).

Similarly to hot-deck, it may happen that a missing observation is not imputed because its covariates have a pattern not observed in the sample. In order to overcome this problem, a sequence of log-linear models with increasing levels of aggregation of covariates are used to impute values. Models are chosen by means of cross-validation, in fact different covariates may induce the selection of different models.

For the units observed in the sample, the observed values I^t_S are used as prediction in BRI. This choice has the advantage of preserving the consistency of predicted ALE in BRI with the variables observed in the sample survey, and consequently statistical models on those variables may use micro-data in BRI without any problems concerning micro-consistency.

Different log-linear models are used within groups A, B and C, mainly because of the different available information. As already remarked, in group A a log-linear model is estimated by using only administrative data, while for the other groups log-linear models are estimated by using survey data as well. In the following, some details are given for each group.

Imputation in Group A - Change.

This group is characterised by active people, which means people that are currently attending a course in the reference year. Administrative information is particularly important in this group, in fact the aim is essentially to predict the attainment of educational level given that is known which is the year of the course they are attending during the year [t-12, t].

We have decided to estimate the conditional probabilities of obtaining a new educational level by using only administrative data. The imputation method consists in the estimation of a model applied to data referring to 1 year before the time reference, and then by applying the estimated model to the year of reference. In the specific application, firstly we have estimated the

conditional probabilities on data to predict the ALE at 2017 (known in the administrative data) by using data available in the interval time [2016, 2017]. Then, we have applied the model to predict the ALE at the reference time t=2018. The underlying idea is that there is no variation into the conditional probabilities in one year, and that the error introduced by this assumption is lower than the sampling error introduced by using instead sample survey data.

In order to ease referring to models, we adopt the classic notation for hierarchical log-linear models where only the highest-order interaction term for each variable is reported (see Agresti, 2002, pp. 320).

The log-linear model used in the first step of the sequence of imputations is the saturated model:

$$[C^t, E8^t, F^t, L^t, I^t] \qquad (1)$$

that is first estimated with t=2017 by region and then applied to t=2018. Although, as previously declared, a sequence of models is used to impute data, most of the non-responses are imputed by using (1).

Imputation in Group B.

People not attending any course covered by MIUR in *t-1/t* characterise this group. These are people that either have decided to stop their studies or that are attending some courses unfortunately not covered by MIUR. Because of the MIUR under-coverage, it is necessary to resort to sample survey data. The conditional probabilities $h(I^t|X)$ are estimated by region through the log-linear model

$$[Prov, I^t]\,[G, C^t, E8^t, I^{t-12}, I^t] \qquad (2)$$

where Prov is the province of residence. The model is estimated on t=2018 by considering the observed values in the sample, *i.e.* $I^t = I^t_s$.

Also in this group, a sequence of imputation models are used, however almost all the units are imputed by using model (2).

Imputation in Group C.

This group is characterised by two types of units (denoted by the variable D):

- individuals resident on the Italian territory but not detected by the 2011 Census (D=1);
- individuals entered Italy after 2011 and that have not attended any training courses released by MIUR since 2011 (D=2).

These are two populations with distinct socio-demographic characteristics (see Di Cecco *et al.*, 2018) and it is important to include the variable D in the model to distinguish them. Although affected by missing values, another important information is the self-declared ALE I^{apr} reported in APR4. This cannot be used directly as a value to assign to the I^t both because of its level of classification that is too much aggregated, and because of its level of quality being a self-declared variable. However, it results strongly correlated to I^t, therefore it is used as a covariate in the model. This is certainly the most critical population because of their peculiarity and the limited amount of administrative information. In order to fill the lack of knowledge, it is important to use survey data that report the ALE at time *t*.

The model selected for the first step through cross-validation is

$$[\text{Prov}, I^t]\,[E8^t, I^t]\,[G]\,[C^t, I^t]\,[I^{apr}, I^t]\,[D, I^t] \qquad (3)$$

The model is estimated on *t*=2018 by considering the observed values in the sample, *i.e.* $I^t = I^t_s$. Also in this group, almost all the units are imputed according to this model.

A general remark is concerned with the use of other potential covariates like income and type of occupation. Unfortunately, the type of occupation is not available in the due time to be introduced in the modelling. As far as income is concerned, we notice that it is not available for the whole population and it refers to time t-2. Nevertheless in Di Cecco *et al.*, 2018, where a model for predicions in two years (from *t*-2 to *t*) was studied, income resuled as an explicative covariate. However, with the data at hand and by considering the procedure so far illustrated, the introduction of income in the model resulted in strange results in the aggregates and, also by considering the difficulty in the construction of this information in the current Istat production system,

we have opted for excluding it from the current procedure. It is indeed true that, once information on income will be more timely and stable in the Istat production system, additional analysis should be performed in order to take into account the possibility of using such information.

4. Analysis for the assessment of the predictions

In this section, we illustrate the results of some analysis carried out in order to assess the quality of the procedure. Analysis on micro-data and aggregates are performed. Results computed on BRI are analysed and compared with the ones computed on data collected in the sample of the 2018 permanent census, and with the ones computed on data from administrative sources and 2011 Census where available. As far as micro level analysis is concerned, the transitions from 2017 observed ALE to 2018 estimated ALE are studied. In the macro level validation, comparisons between distributions of observed and estimated 2018 ALE are analysed.

Table 4.1 shows the number of imputed values for each imputation step. For the individuals interviewed in the 2018 sample the imputation is deterministic since the predition coincides with the observed value. Excluding the "A-No change" group, which represents the 6.7% of the population, almost all the imputations (88.4%) occur in step 1.

Table 4.1 - Distribution of imputation steps - absolute value (a.v.) **in thousands and percentage values** (%)

Imputation step	a.v.	%
Group A		
A - No change	3,762	6.7
A - Change - step 1	3,581	6.4
A - Change - step 2	4	0.0
A - Change - deterministic: estimated = admin.	9	0.0
Group B		
B - step 1	43,275	77.4
B - step 2	53	0.1
B - step 3	7	0.0
B - step 4	5	0.0
B - step 5	1	0.0
Group C		
C - step 1	2,574	4.6
C - step 2	37	0.1
C - step 3	4	0.0
2018 Sample	2,597	4.6
Total	55,909	100.0

Source: Istat

For groups A and B, transitions between observed and estimated ALE provide a first evaluation of the procedure. In group A, the estimated 2018 ALE is in most cases consistent with the 2017 information from administrative sources (Table 4.2). This happens when the estimated 2018 ALE confirms the 2017 ALE or increases it by one degree. On the other side, inconsistencies arise when the 2018 estimated ALE is lower than the observed 2017 ALE or when the estimated 2018 ALE is more than one degree higher than the 2017 ALE. The inconsistencies regard the subset of people interviewed at the 2018 sample, for which the collected information is used as prediction (see Section 3).

It is worthwhile to report that data editing of sample data was performed mainly looking for the consistency within the sample. Administrative data were used in the sample data editing and imputation process for two main purposes: a macro level validation of the sample data on ALE and a micro level comparison when the sample data on ALE was inconsistent within the sample. Only in this case the administrative ALE was considered in substitution of the ALE declared by respondents.

Table 4.2 - Group A: transition from ALE 2017 (administrative data) to ALE 2018 (estimated data) – row percentage and total absolute value in thousands

		ALE 2018 (estimate)								
	ALE 2017 (administrative)	1	2	3	4	5	6	7	8	Total (*a.v.*)
1	Illiterate	-	-	-	-	-	-	-	-	-
2	Literate but no ed. Attainment	0.0	65.8	34.2	0.0	0.0	0.0	0.0	-	1,628
3	Primary education	0.0	0.0	65.8	33.5	0.6	0.0	0.0	-	1,791
4	Lower secondary education	0.0	0.0	0.0	77.1	22.6	0.1	0.1	0.0	3,565
5	Upper secondary education	0.0	0.0	0.0	0.0	90.4	7.1	2.4	0.0	3,449
6	Bachelor's degree	0.0	-	0.0	0.0	0.2	81.4	18.2	0.2	923
7	Master's degree	0.0	0.0	0.0	0.0	0.1	0.1	97.0	2.8	892
8	PhD	-	-	-	0.0	0.0	0.0	0.8	99.1	45
	Total	0.0	8.7	14.1	27.3	32.0	8.2	9.1	0.6	12,292

Source: Istat

In group B, the estimated 2018 ALE shows some inconsistencies with ALE in 2017 (Table 4.3). The information on ALE in 2017 derives from the 2011 Census and regards individuals who have not enrolled in any standard training course from 2011 to 2017 so the educational level is not changed

until 2017. The basic hypothesis is that the information collected in 2011 is not error-free and that the administrative data on school attendance may be under-covered, therefore, for this sub-population, the information on the educational level in 2017 can be corrected based on the information from the 2018 sample. There is no restriction on the fact that the estimated ALE in 2018 should be higher than that of 2017.

Table 4.3 - Group B: transition from ALE 2017 (CENS 2011) to ALE 2018 (estimated data) - row percentage and total absolute value in thousands

		ALE 2018 (estimate)								
	ALE 2017 (CENS 2011)	1	2	3	4	5	6	7	8	Total (*a.v.*)
1	Illiterate	46.2	22.1	19.7	8.8	2.5	0.2	0.5	0.0	371
2	Literate but no ed. Attainment	5.3	38.4	42.2	10.1	3.3	0.1	0.6	0.0	1,182
3	Primary education	0.6	5.6	78.7	12.5	2.3	0.1	0.3	0.0	7,298
4	Lower secondary education	0.1	0.6	5.8	78.9	13.9	0.2	0.5	0.0	12,937
5	Upper secondary education	0.1	0.2	1.0	6.5	89.3	1.2	1.6	0.0	14,05
6	Bachelor's degree	0.0	0.2	0.6	3.1	16.8	61.9	17.1	0.2	943
7	Master's degree	0.0	0.1	0.7	2.0	3.9	2.3	90.0	1.0	4,016
8	PhD	0.0	0.1	0.5	1.0	2.1	0.7	27.3	68.4	136
	Total	0.7	2.6	17.7	30.0	36.4	2.1	10.1	0.4	40,931

Source: Istat

In order to evaluate the imputation procedure in a macro level approach, the estimated ALE in 2018 ($\widehat{I^t}$), obtained on the Italian resident population is compared with the data collected in the 2018 census sample, appropriately weighted (I_s^t). In particular, we focus on the differences between the frequency distributions of estimated 2018 ALE in BRI and the distribution computed on weighted sample data. A synthetic measure of the difference between distributions is given by the average of the absolute values of the differences between percentage of each item, in absolute (*AD*) and relative (*RD*) terms. Specifically:

$$AD = \frac{\sum_{i=1}^{8}|D_i|}{8} = \frac{1}{8}\sum_{i=1}^{8}\left|fr\left(\widehat{I^t}\right)_i - fr\left(\widehat{I_s^t}\right)_i\right|$$

$$RD = \frac{\sum_{i=1}^{8}|Dr_i|}{8} = \frac{1}{8}\sum_{i=1}^{8}\frac{|fr(\widehat{I^t})_i - fr(\widehat{I_s^t})_i|}{fr(\widehat{I_s^t})} * 100$$

where $fr(\widehat{I}^t)_i$ is the relative frequency of ALE item *i* estimated in 2018 through the model and $fr(\widehat{I}_s^t)_i$ is the relative frequency of ALE item *i* estimated with the 2018 weighted sample.

The macro level comparison between BRI and sample estimates shows that the two distributions are very similar (Table 4.4). The distribution of the estimated ALE differs from the weighted sample data by 0.21% points on average on each item; the differences are concentrated in level 5 "Upper secondary education", which is the most frequent one. In relative terms (*Dr*), differences are concentrated in the extreme and less frequent levels. In particular level 1 "Illiterate" and level 2 "Literate but no formal educational attainment" are confused and difficult to be predicted.

Table 4.4 - Model and sample estimates of 2018 ALE (absolute values in thousands) and absolute (*D*) and relative (*Dr*) differences between model and sample percentages

		Model		Sample		Model – Sample (a)	
ALE 2018		a.v.	%	a.v.	%	D_i	Dr_i
1	Illiterate	353	0.6	330	0.6	0.03	5.83
2	Literate but no ed. Attainment	2,295	4.1	2,073	3.7	0.36	9.65
3	Primary education	9,293	16.6	9,137	16.5	0.12	0.72
4	Lower secondary education	16,509	29.5	16,168	29.2	0.33	1.12
5	Upper secondary education	19,718	35.3	19,873	35.9	-0.63	-1.74
6	Bachelor's degree	1,977	3.5	1,962	3.5	-0.01	-0.16
7	Master's degree	5,531	9.9	5,598	10.1	-0.22	-2.15
8	PhD	233	0.4	227	0.4	0.01	1.67
Total		55,909	100.0	55,368	100.0	*AD*=0.21	*RD*=2.88

Source: Istat
(a) The calculations from the table may give different numbers due to the approximation.

The comparison between target and estimated distributions for groups A, B and C shows different behaviours. In particular, in group C the estimated distribution differs from that of the sample, more than it differs in the A and B groups. This is due to the lower quantity and quality of available information (Table 4.5). On the contrary, in group B the estimated and sampled ALE distributions are almost perfectly equivalent. This is mainly related to the greater number of the individuals in group B, in addition to the fact that the information on ALE from the sample is used as response variables in the

model (this also applies for group C, but not for group A). It is worthwhile to notice that when class 1 and 2 are jointly considered (see Table 4.4) the difference is not high, in fact these two modalities are generally hardly to discriminate and most of the times their counts are jointly provided in the published tables. A remarkable difference is also in class 3 (Upper secondary education), we notice that this is observed both in A and C. Further analysis, jointly performed with subject matter experts, are still in progress to understand the reasons behind those differences.

Table 4.5 - Model and sample estimates of 2018 ALE (percentage values) and absolute differences (D) between model and sample percentages in the three groups A, B and C

		Model			Sample			Model – Sample		
		A	B	C	A	B	C	A	B	C
ALE 2018		%	%	%	%	%	%	$D_i^{(*)}$	$D_i^{(*)}$	$D_i^{(*)}$
1	Illiterate	0.0	0.7	1.9	0.0	0.7	1.7	-0.0	0.0	0.2
2	Literate but no ed. Attainment	8.7	2.6	6.4	7.5	2.5	5.4	1.2	0.0	1.0
3	Primary education	14.1	17.7	11.7	14.3	17.4	10.7	-0.2	0.3	1.0
4	Lower secondary education	27.3	30.0	32.0	26.8	29.7	33.2	0.4	0.3	-1.2
5	Upper secondary education	32.0	36.1	33.1	33.5	36.4	34.4	-1.5	-0.3	-1.3
6	Bachelor's degree	8.1	2.2	3.5	8.1	2.2	3.4	0.1	-0.0	0.2
7	Master's degree	9.2	10.4	10.6	9.2	10.7	10.6	-0.0	-0.3	0.0
8	PhD	0.6	0.4	0.6	0.5	0.4	0.5	0.0	-0.0	0.1
Total		100.0	100.0	100.0	100.0	100.0	100.0	AD=0.43	*AD*=0.16	*AD*=0.62

Source: Istat

The distribution of ALE will be published yearly by Istat taking into account for some other variables such as gender, age classes and citizenship so it is important to take into account the distributional accuracy in these specific subpopulations. Looking at the distribution of ALE 2018 by citizenship, differences between estimated and weighted sample data are evident especially on the sub-population of Not Italian people (Table 4.6) in which we observe an average difference of 0.39 points on each estimated item with respect to the weighted sample. The Not Italian subpopulation is small with respect to the total (9%) and is characterised by particular features and less information available determining a different fit of the model.

Table 4.6 - Model and sample estimates of 2018 ALE (absolute values in thousands) and absolute (*D*) differences between model and sample percentages, by citizenship

		Model				Sample				Model – Sample (a)	
		Italian		Not Italian		Italian		Not Italian		Italian	Not Italian
ALE 2018		a.v.	%	a.v.	%	a.v.	%	a.v.	%	D_i	D_i
1	Illiterate	270	0.5	83	1.8	258	0.5	73	1.6	0.02	0.14
2	Literate but no ed. attainment	1,976	3.9	319	6.9	1,8	3.5	273	6.2	0.32	0.69
3	Primary education	8,787	17.1	506	10.9	8,658	17.0	479	10.8	0.15	0.06
4	Lower secondary education	14,968	29.2	1,541	33.1	14,686	28.8	1,482	33.4	0.37	-0.35
5	Upper secondary education	18,049	35.2	1,668	35.8	18,232	35.8	1,641	37.0	-0.58	-1.20
6	Bachelor's degree	1,812	3.5	166	3.6	1,812	3.6	149	3.4	-0.02	0.19
7	Master's degree	5,175	10.1	356	7.6	5,278	10.4	320	7.2	-0.26	0.42
8	PhD	215	0.4	18	0.4	212	0.4	15	0.3	0.00	0.05
Total		51,252	100.0	4,656	100.0	50,936	100.0	4,431	100.0	*AD*=0.22	*AD*=0.39

Source: Istat
(a) Warning: the calculations from the table may give different numbers due to the approximation.

Looking at the territorial level, Table 4.7 shows the differences between estimated and observed percentages of each item (D) and the mean of absolute differences of each item (AD) by region. Even if in general there is a small variability between regions, it can be seen that northern regions have lower differences between estimated and observed distributions (see Piemonte, Valle d'Aosta/Vallée d'Aoste, Lombardia, Friuli-Venezia Giulia and Emilia-Romagna), vice versa in the southern regions and islands (Puglia, Calabria, Sicilia and Sardegna). It is worth noting that item 1 ("Illiterate") and 2 ("Literate but no educational attainment") are over-estimated in all regions while item 5 ("Upper secondary education") is always under-estimated. Further analyses are needed to understand the reasons.

Table 4.7 - Item absolute differences (Di) and mean of absolute differences (AD) between model and sample percentages by region

Regions	Illiterate	Literate but no att.	Primary ed.	Lower sec. ed.	Upper sec. ed.	Bachelor's degree	Master's degree	PhD	Mean
	(D_1)	(D_2)	(D_3)	(D_4)	(D_5)	(D_6)	(D_7)	(D_8)	(AD)
Piemonte	0.04	0.35	0.15	0.22	-0.54	-0.09	-0.13	0.01	0.19
Valle d'Aosta/*Vallée d'Aoste*	0.03	0.18	-0.08	0.43	-0.33	-0.13	-0.14	0.03	0.17
Lombardia	0.03	0.37	-0.08	0.06	-0.4	0.02	-0.01	0.01	0.12
Trentino-Alto Adige/*Südtirol*	0.02	0.22	-0.3	-0.3	-0.31	0.04	0.58	0.04	0.23
Bolzano/Bozen	*0.01*	*0.11*	*-0.58*	*0.39*	*-0.27*	*-0.06*	*0.37*	*0.01*	*0.23*
Trento	*0.04*	*0.33*	*-0.02*	*-0.98*	*-0.34*	*0.13*	*0.78*	*0.07*	*0.33*
Veneto	0.03	0.34	0.11	0.34	-0.7	0.01	-0.13	0	0.21
Friuli-Venezia Giulia	0.05	0.14	0.09	0.18	-0.6	0.08	0.07	-0.01	0.15
Liguria	0.01	0.29	0.32	0.77	-0.73	-0.07	-0.56	-0.02	0.35
Emilia-Romagna	0.02	0.33	-0.1	0.15	-0.41	-0.01	-0.03	0.03	0.14
Toscana	0.03	0.35	0.16	0.47	-0.65	0.01	-0.37	0	0.26
Umbria	0.05	0.43	0.22	0.49	-0.77	0.01	-0.41	-0.02	0.3
Marche	0.02	0.25	0.15	0.58	-0.83	0	-0.18	0.01	0.25
Lazio	0.03	0.44	0.14	0.68	-0.73	-0.07	-0.44	-0.03	0.32
Abruzzo	0.03	0.44	0.19	0.54	-0.66	-0.1	-0.45	0	0.3
Molise	0.04	0.19	-0.35	0.42	-0.13	0.08	-0.23	-0.01	0.18
Campania	0.03	0.45	-0.02	0.05	-0.54	0.07	-0.06	0.03	0.16
Puglia	0.08	0.38	0.3	0.67	-0.91	0	-0.52	0.01	0.36
Basilicata	0.1	0.4	0.05	-0.41	-0.64	0.05	0.43	0.02	0.26
Calabria	0.03	0.32	0.69	0.56	-1.37	-0.02	-0.24	0.03	0.41
Sicilia	0.05	0.37	0.43	0.46	-0.76	-0.04	-0.53	0.03	0.33
Sardegna	0.06	0.3	0.25	0.58	-0.6	0.04	-0.61	-0.03	0.31
Italy	0.04	0.37	0.12	0.33	-0.63	-0.01	-0.22	0.01	0.21

Source: Istat

Most of the analyses illustrated so far are concerned with aggregates, that are the first main goal of the procedure. In fact, the decision to impute with a random draw from the estimated conditional distribution is aimed at increasing the preservation of distributions, while unfortunately decreasing the predictive accuracy (at micro-level) of the model.

Nevertheless, it is interesting to look at the predictive accuracy of the model, since data are predicted at micro level in BRI. In Table 4.8, we report the differences at micro-level computed comparing the imputed ALE vs the ALE observed in the sample survey. Out of the whole sample, 74% of units are exactly predicted. As expected, the best predictive accuracy is in set A (88%) that is in fact the subset of data with the highest amount of administrative

information. It is worthwhile to remind that in this subset, the model is estimated by using only administrative data, and this makes the result even more interesting. On the other side, we notice the poor performance in terms of micro-predictions of the model in the set C. This was expected as well, since C is characterised by a very low level of auxiliary information, but it fortunately refers to a small part of the total population (2.8% of the data used for the comparison).

Table 4.8 - Differences at micro level between estimated and observed ALE in the sample survey. DIF is equal to 1 when values are different

	Group			
	A	B	C	Total
DIF	%	%	%	%
0	87.9	71.9	34.0	74.3
1	12.1	28.1	66.0	25.7
Total	100.0	100.0	100.0	100.0

Source: Istat

5. Final remarks and future developments

In this paper a mass imputation procedure for the attained level of education is described. The procedure combines different data sources: Administrative data, sample survey data and Census data.

The imputation models are based on log-linear models, which have the advantage over the traditional hot-deck procedures to be more parsimonious. This flexibility is an important issue since as noted in De Waal (2016) "mass imputation relies on the ability to capture all relevant variables and relevant relations between them in the imputation model, and to estimate the model parameters sufficiently accurately".

Methods to estimate the variance of register-based statistics built by using administrative and sampling data are being tested. They are based on resampling techniques for finite population, see Chen *et al.* (2019) for a general discussion and Di Consiglio *et al.* (2019) and Scholtus (2018) for the cases of integrated administrative data.

Istat has planned to produce BRI on a yearly basis, hence the imputation model proposed in the paper should be modified in order to include sampling information referring to each year, that in the illustrated case means it should be designed a model based on sample data related to 2018 and 2019 to predict the ALE 2019.

Further analysis will be dedicated to the use of additional information to improve the predictions, for instance, the inclusion of family composition can be important to this aim.

An important issue is related to the production of 2021 Census figures. In this paper, ALE is predicted with a classification based on 8 categories, while for the 2021 Census a more detailed classification is required. Further studies are needed to produce predictions for the attained level of education at a finer classification.

References

Agresti, A. 2002. *Categorical Data Analysis*. Hoboken, NJ, U.S.: John Wiley & Sons.

Chen, S., D. Haziza, C. Léger, and Z. Mashreghi. 2019. "Pseudo-population bootstrap methods for imputed survey data". *Biometrika*, Volume 106, Issue 2: 369-384.

Daalmans, J. 2017. "Mass imputation for Census estimation". In United Nations Economic Commission for Europe – UNECE, *Conference of European Statisticians, Group of Experts on Population and Housing Censuses*. 19th Meeting, Geneva, Switzerland, 4th - 6th October 2017.

de Waal, T. 2016. "Obtaining numerically consistent estimates from a mix of administrative data and surveys". *Statistical Journal of the IAOS*, Volume 32, N. 2: 231-243.

Di Consiglio, L., M. Di Zio, and D. Filipponi. 2019. "An empirical evaluation of latent class models for multisource statistics". *Presentation at ITACOSM 2019*. Firenze, Italy, 5th -7th June 2019.

Di Cecco, D., D. Di Laurea, M. Di Zio, R. Filippini, P. Massoli, and G. Rocchetti. 2018. "Mass imputation of the attained level of education in the Italian System of Registers". In United Nations Economic Commission for Europe – UNECE, *Workshop on Statistical Data Editing*. Neuchâtel, Switzerland, 18th -20th September 2018.

Runci, M.C., G. Di Bella, and F. Cuppone. 2017. "Integrated Education Microdata to Support Statistics Production". In Lauro, N.C., E. Amaturo, M.G. Grassia, B. Aragona, and M. Marino (*eds*.). *Data Science and Social Research. Epistemology, Methods, Technology and Applications*. Heidelberg, Germany: Springer International Publishing, *Studies in Classification, Data Analysis, and Knowledge Organization*.

Scholtus, S., and J. Pannekoek. 2015. "Mass-imputation of educational levels". *Internal report* (available in Dutch). The Hague and Heerlen, The Netherlands: Statistics Netherlands – CBS.

Scholtus, S. 2018. "Variances of Census Tables after Mass Imputation". *Discussion paper*. The Hague and Heerlen, The Netherlands: Statistics Netherlands – CBS.

Skinner, C., and L.-A. Vallet. 2010. "Fitting log-linear models to contingency tables from surveys with complex sampling designs: an investigation of the Clogg-Eliason approach". *Sociological Methods & Research*, Volume 39, Issue 1: 83-108.

Thibaudeau, Y., E. Slud, and A. Gottschalck. 2017. "Modeling log-linear conditional probabilities for estimation in surveys". *The Annals of Applied Statistics*. Volume 11, Issue 2: 680-697.

La Rivista di Statistica Ufficiale accoglie lavori che hanno come oggetto la misurazione e la comprensione dei fenomeni sociali, demografici, economici e ambientali, la costruzione di sistemi informativi e di indicatori come supporto per le decisioni pubbliche e private, nonché le questioni di natura metodologica, tecnologica e istituzionale connesse ai processi di produzione delle informazioni statistiche e rilevanti per il perseguimento degli obiettivi della statistica ufficiale.

La Rivista di Statistica Ufficiale si propone di promuovere la collaborazione tra il mondo della ricerca scientifica, gli utilizzatori dell'informazione statistica e la statistica ufficiale, al fine di migliorare la qualità e l'analisi dei dati.

La pubblicazione nasce nel 1992 come collana di monografie "Quaderni di Ricerca Istat". Nel 1999 la collana viene affidata a un editore esterno e diviene quadrimestrale con la denominazione "Quaderni di Ricerca - Rivista di Statistica Ufficiale". L'attuale denominazione, "Rivista di Statistica Ufficiale", viene assunta a partire dal n. 1/2006 e l'Istat torna a essere editore in proprio della pubblicazione.

1B012016003000000

ISSN 1828-1982

€ 15,00