VOLUME LXX – N. 4 OTTOBRE-DICEMBRE 2016

# RIVISTA ITALIANA DI ECONOMIA DEMOGRAFIA E STATISTICA

*Sede Legale:* C/O Studio Associato Cadoni, Via Ravenna n.34 – 00161 ROMA.
sieds.new@gmail.com, rivista.sieds@gmail.com

## INDICE

# STATISTICS IN THE BIG DATA ERA[1]

Agostino Di Ciaccio, Giovanni Maria Giorgi

## 1. Introduction

It is estimated that about 90% of the currently available data have been produced over the last two years. Of these, only 0.5% is effectively analysed and used. However, this data can be a great wealth, the oil of 21st century (Sondergaard, 2011), when analysed with the right approach. In this article, we illustrate some specificities of these data and the great interest that they can represent in many fields.

In section 2 we analyse some common data sources and the big-data characteristics in various application contexts. In section 3 the relevance of the cloud computing is considered and, in section 4, some problematic aspects of big-data are reported. New challenges for the statistical analysis are considered in the section 5 and 6, suggesting some strategies.

## 2. The data deluge

Doug Laney (2001) defined the concept of big-data as related to three main keywords: *volume*, *velocity* and *variety*.

*Volume* indicates that data have a large number of units and variables. *Velocity* is needed by real time analysis of streaming data, generated for example by sensors. *Variety* indicates that data comes in all types of formats – from structured, numeric data to unstructured text documents, email, video.

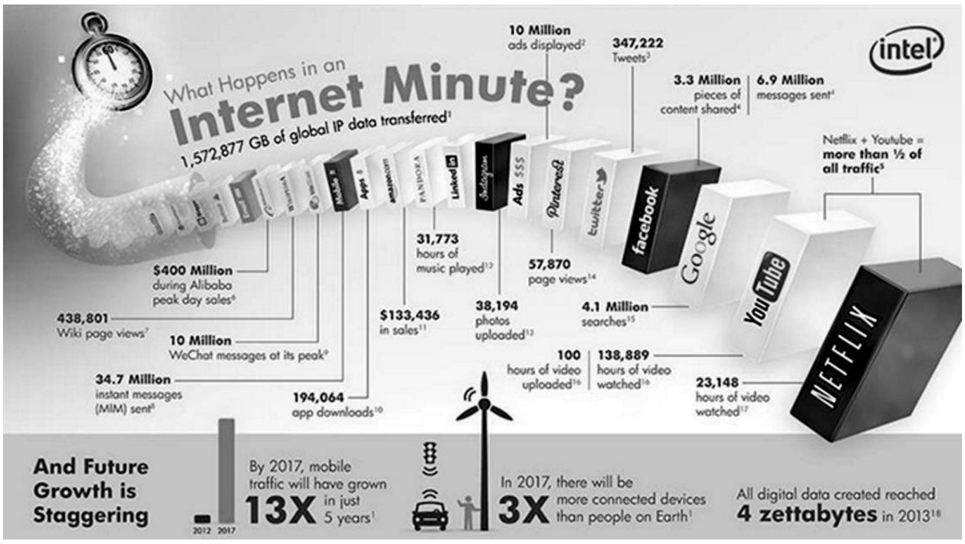
Often, big-data have a different structure than the traditional data-base and are sometimes called *date-lake*. In traditional database, the data set must be carefully designed before you can enter data. Conversely, a *data-lake* indicates a data storage repository that contains a large amount of raw data in native format, with a high level definition of what exists in the lake of data.

[1] Invited paper to the 53rd SIDES Scientific Meeting – Rome 2016.

The United Nations Economic Commission for Europe (UNECE, 2013) classified the big-data sources as:

- *Human-sourced information*: social networks, blogs, pictures, videos, search engine queries, mobile data content, etc.
- *Process mediated/transaction data*: commercial transactions, banking/ stock prices records, e-commerce, credit cards, medical records, etc.
- *Machine-generated data* (Internet of Things): sensor data (weather/ pollution, traffic, security/surveillance,…), tracking devices (GPS systems, mobile phone location, satellite images), data from computer systems (logs & web logs, …).

**Figure 1** – *Infographic of big-data generated on Internet (source: intel 2014).*

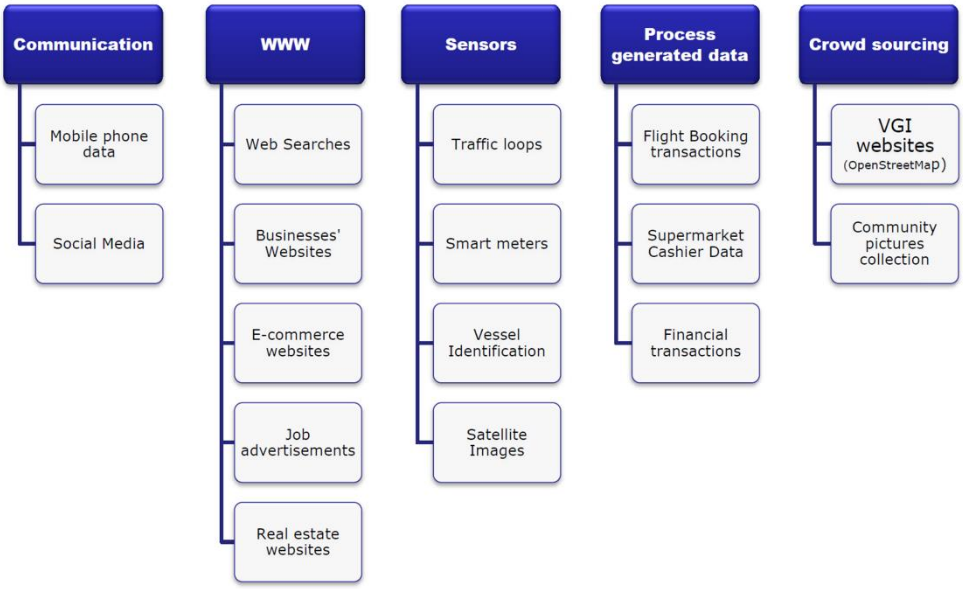


Of course, the Internet is at the moment one of the most prolific sources of data, as is shown in fig. 1. A classification of sources as suggested by EUROSTAT (Skaliotis, 2015) is proposed in fig. 2.

The characteristic of big-data is not just the size. They are often new data types, concerning people's behaviour and beliefs, new types of instruments, and new types of actors.

As pointed out in the introduction, currently we produce a huge amount of data but only a small percentage is actually analysed. In the last few years however, there has been a growing interest in the analysis of these data in various application fields. For example, the predictive power of "*social*" big-data is being used in many fields like public health or economic development. Global Pulse, an initiative by the United Nations (www.unglobalpulse.org), tries to leverage these data for global development. The group carried out analysis of messages in social networks for several projects, as "*Understanding Immunisation Awareness And Sentiment Through Analysis Of Social Media And News Content* (2015)" or "*Analysing Social Media Conversations To Understand Public Perceptions Of Sanitation* (2014)". The next paragraphs will deepen some applications of big-data analysis.

**Figure 2** – *Classification of big-data sources (Eurostat 2015).*



## *2.1. Textual analysis*

Recent years have seen a tremendous growth of text-based data, in particular web pages, news, e-mail and social media. A characteristic of textual data is that they are generated directly by humans, rather than for example by sensors, and are therefore very useful to analyse people's opinions and preferences.

The explosive growth of text data makes very difficult to analyse this data in a timely manner. Therefore, information retrieval systems, able to identify the information quickly and accurately, are necessary. This need has led, for example, to the creation of the well known web search engines. Moreover, it is very useful to analyse textual data, such as product reviews, forum discussions, and social media to get user's opinions.

Text constitutes unstructured data, which do not conform to well-defined patterns, and are therefore relatively complex to analyse. The optimal processing goal would be to understand the content encoded in the text, but the current technology does not yet allow a computer to accurately understand the natural language text. On the other hand, a wide variety of statistical approaches allow the analysis of textual data. For example, *recursive neural networks* have had significant successes in a number of tasks as in Socher et al. (2013) that used this technique to predict sentence sentiment, with good results.

### *2.2. Medicine*

The large amount of genome sequencing data now make it possible to uncover the genetic markers of rare disorders and find associations between diseases and rare sequence variants.

John Craig Venter was the first to sequence the human genome, with the enormous cost of about 100 million ($). The Venter's target now is to sequence at least one million genomes and use these data, together with information on the health of the DNA donor and the results of other medical tests, to identify new methods of treatment and prevention for many diseases. This approach, named also *precision medicine*, focuses on individual differences in the genes of patients, often in combination with information about their environment, health history and lifestyle. It is a big change in our current treatments, which focuses on generic approaches for as many people as possible, instead of personalized treatments. The potential of this approach has motivated an investment of 215 million ($) from the White House.

### *2.3. IoT, the Internet of Things*

Internet is entering a new phase of growth in which the “things” around us will be connected to the web. Internet of Things (IoT) will radically change the way we interact with our environment.

The large number of sensors, scattered in almost all sectors and connected to Internet, is beginning to result in a massive afflux of data. According to reliable estimates, by 2020, there should be 32 to 50 billion devices connected to Internet, and the volume of data generated is staggering.

Recently, Dutch Telco KPN announced, with great emphasis, that it has completed the national coverage of the Netherlands with a wireless Internet network of things. So far, KPN has inked contracts to connect 1.5 million devices. Similar IoT-networks are going up in France, Germany, South Korea, and elsewhere across the globe.

In its simplest form, the IoT, through networks and sensors integrated with cloud software, allows devices to communicate, analyse and share data.

To understand the relevance of IoT, we can list its application to some classical problems:

- *Predictive equipment maintenance*. Can be used, for example, to manage energy, predict equipment failure or detect other issues. This approach is also being used in the automobile industry, in which the same cars will be able to predict, and prevent, their failures.
- *Moving merchandise more efficiently*. This is one of the classical goals of smart transportation applications in retail, and can be obtained by a very accurate tracking and route optimization.
- *Warehouse automation and optimization*. IoT allows us to monitor sales opportunities in real time and track missed in-store sales. With IoT, we can also understand when the customer needs help or an incentive to purchase, and we can respond proactively.

IoT can create new services and business models. A key role is played by the widespread use of sensors, which can detect events or changes in quantities. Typically, data arising from sensors is in time series format and is often geotagged. Common *smart city* applications include transportation, energy grids, utilities like water, street lighting, parking etc.. For example Barcelona offers smart parking meters that operate on city-wide Wi-Fi, giving residents real-time updates on where to park and allowing them to pay with their phone.

IoT is having a key role also in the *smart home* applications: for example, the rolling shutters or heating can be controlled by a smartphone and some manufacturers already produce Internet-enabled appliances. Also the home security system can allow Internet-based monitoring of the environments and alarms.

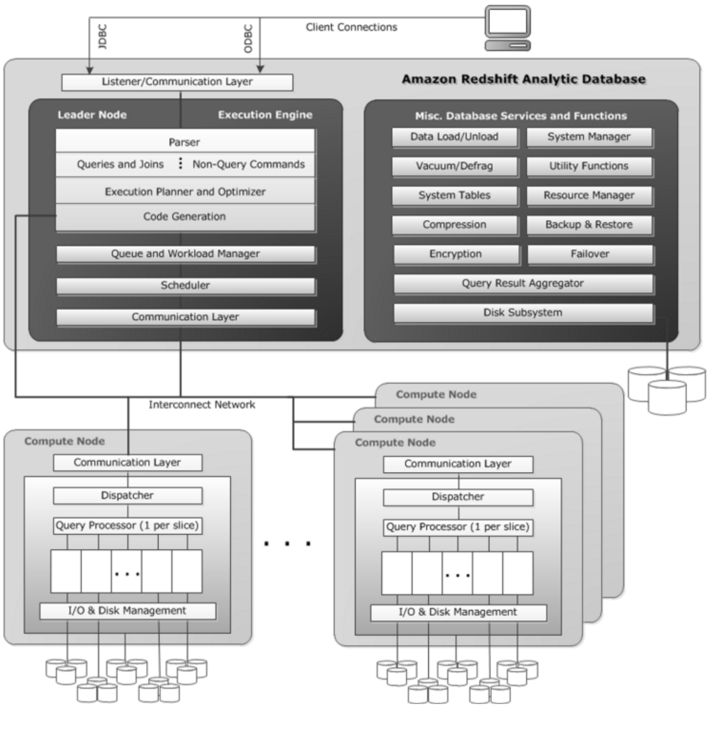
## 3. Big-data on the Cloud

If you need to analyse big-data, it is no longer necessary to have powerful hardware. In fact, you can use many services in the cloud that allows you to manage, structure and analyse large amounts of data. These services also allow the user not experienced in hardware and software to create procedures that lead to meaningful results.

Currently there are five main web platforms that can be used to analyse big-data on the Cloud. Using one of these platforms, in a few minutes, a cluster can be built for distributed analysis of big-data, obtaining performance and security without the cost of managing a complex hardware infrastructure:

- Amazon Web Services (AWS)
- Google Cloud Platform
- Microsoft Azure
- IBM Analytics
- SAP HANA Cloud Platform

**Figure 3 -** *High level view of internal components and functionality of the Amazon Redshift data warehouse (aws.amazon.com/it/documentation/redshift/)*

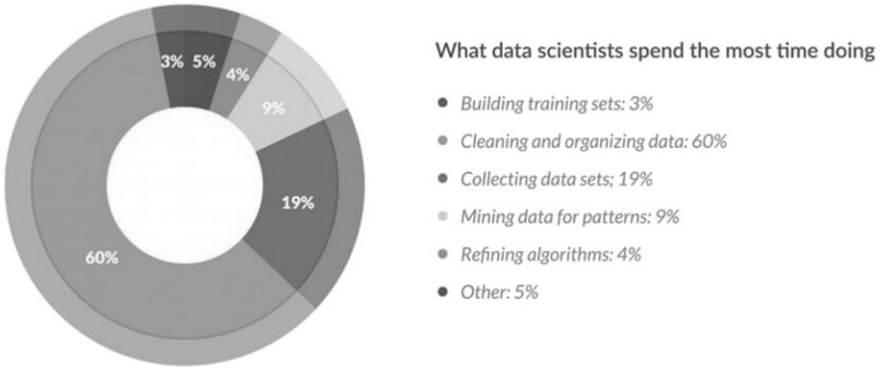


Each platform has its strengths. For example, Google Cloud allows to build complex machine learning models by the powerful TensorFlow framework. This software powers many Google products, with a managed scalable infrastructure, which is powered by GPUs. Another example is Amazon Redshift, a cloud-based data warehouse (see fig. 3). It is both scalable and, despite its complexity, relatively easy to use.

## 4. Unpleasant aspects of big-data

The characteristics that big-data possess, involve a number of problems that must be addressed, before and during the statistical analysis. For example, it was observed that '*data preparation*' accounts for about 80% of the work of data scientists (Forbes 2016). In particular, they spend 60% of their time on cleaning and organizing data, while collecting data sets comes second at 19% of their time (fig. 4). Most of them view data preparation as the least enjoyable part of their work.

**Figure 4 -** *Forbes, march 23, 2016*



Many other problematic aspects have to be considered:

- *Privacy*. Combining someone's personal information with vast external data sets, it is possible to infer new facts about that person, including sensitive information.
- *Coverage/selection bias*. Usually, data are not collected considering a statistical representativeness. The population of interest could have different characteristics than the observed collective.
- *Data security*. The enormous amount of data generated by IoT, often stored on the cloud, is a big problem for data security.

- *Data storage*. Despite the great technical advances, automated tools can create enormous archives making their storage too expensive.
- *Computing power*. Depending on the features, big-data may require fast processors, distributed big-data platforms (such as Hadoop), parallel processing, clustering, MPP, virtualization, large grid environments, high connectivity and high throughputs.

As an example of new privacy and security issues, at the end of 2015, Vtech, a manufacturer of digital toys based in Hong Kong, admitted that cybercriminals had access to the personal data of 6.4 million children by remote control of their toys connected to Internet.

## 5. Statistical challenges in big-data

The primary value of big-data does not come from the data in their raw form, it comes from the elaboration of data and the products and services that can emerge from the analysis. The radical changes that are emerging in the data management technologies must be accompanied by changes in analytical techniques and the way in which these support decisions. Following the Scheveningen Memorandum on big-data and official statistics (2013), the general directors of the National Statistical Institutes "*Acknowledge that the use of Big-data in the context of official statistics requires new developments in methodology, quality assessment and IT related issues. The European Statistical System should make a special effort to supports these Developments*".

With big-data we have a great opportunity to take advantage of computational and statistical methods to transform raw data into knowledge in several areas such as health and medicine, security, education, and business intelligence.

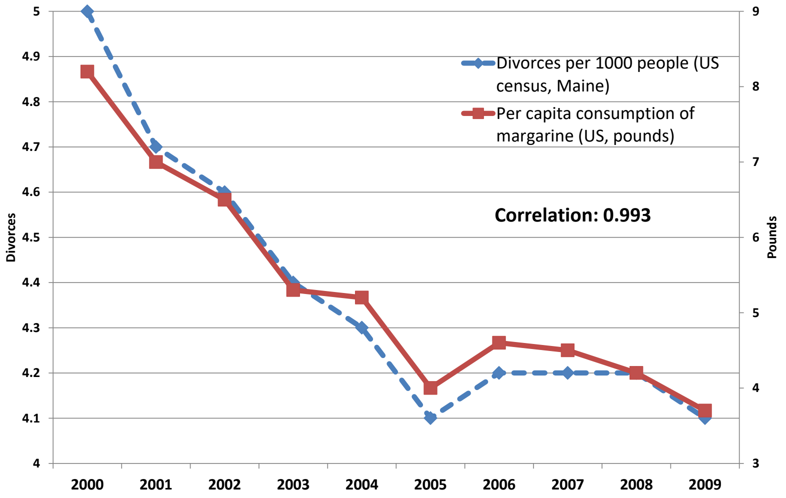
On the other hand, the high dimensionality introduces unique computational and statistical challenges.

- If there is not sampling, then there are not sampling errors. Rather, in this case, we should evaluate the model bias and its reliability or analyse the quality of data.
- In some cases, big-data have much more features than observations. For example, the standard set of microarray data are typically composed of thousands of features (the genes) with only few units.
- The number of fake correlations usually grows with the number of features. Spurious correlation may cause wrong statistical results (see fig. 5).

- When testing sequentially many hypotheses, we must correct for multiple testing. In fact, classical hypothesis procedure defines the test significant 5% (or 1%) of the time, even when the null is true.
- The traditional statistical modelling assumptions are hardly satisfied, which implies biased model parameters and inaccurate statistical tests.
- With big-data, common problems such as sampling bias, missing or incomplete data and sparsity must also be addressed.

Additional problems have been reported in the literature: the population can be unknown or difficult to define, errors can be placed in the pre-processing phase, data coming from social media usually include irrelevant babble or social bots. So, new tools and statistical methods will need to be developed for the pre-processing, classification, summarisation, feature extraction, anonymization and visualization of big-data.

**Figure 5** – *A famous spurious correlations(tylervigen.com).*

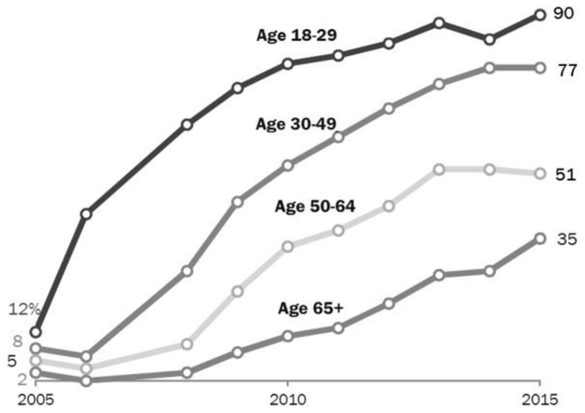


Usually, the major obstacle in the statistical analysis of big-data is the frequent lack of statistical representativity of the observed data. The selection bias can have a different relevance, depending of the analysed phenomena and the aim of analysis.

In recent years, particular attention has been paid to the composition of the collective that uses social networks. Following a research of the Pew Research Center (Perrin, 2015) on USA population, the bias concerns mainly *age* and *socio-economic status*. It is not surprising that young adults are the most likely to use social media but, today, 35% of all those 65 and older report using social media. Moreover, those in higher-income households were more likely to use social media, while no significant difference was found with respect to gender, educational attainment, racial or ethnic group. A slight difference was found with respect to the urbanization level of the residence place.

Knowing the type of selection bias in the data, allows us to modify the analysis and avoid erroneous results.

**Figure 6 -** *Percentage of American adults using social media by age.*



Source: Pew Research Center surveys, 2005-2006, 2008-2015. No data are available for 2007.

## 6. Big-data analytics

In the last years, companies are no longer satisfied to extract detailed information from their archives, but now they require the application of complex predictive models (the analytics).

Among the models suitable for the analysis of big-data, we must consider *neural networks*. One of the advantages of this technique is that it does not require any distributional assumptions on the explanatory variables and the target (if we adopt a supervised approach). On the contrary, neural networks require a great amount of data to assure convergence to the correct parameter estimates. The downside is the difficulty to interpret the model or measure the importance of the variables. Big convolutional or recurrent neural networks have proven to be very effective in classification and cluster analysis (LeCun et al. 2015) and for this purpose they are used, for example, by Google and Facebook.

If the goal is prediction accuracy, averaging many prediction models together, can be a good choice. The idea is that by averaging (or majority voting) several prediction algorithms it is possible to reduce variability without giving up bias. One of the earliest proposals is based on bootstrapping samples and building multiple prediction functions (*Bagging - bootstrap aggregating*, Breiman 1996). *Random forests* extended this idea with classification trees, by a randomization of the features (Breiman 2001), while a further extension that allows to better avoid overfitting are the *extremely randomized trees* (Geurts et al. 2006).

Usually, the prediction algorithms that most frequently win Kaggle statistical competitions, and won the Netflix prize, blend multiple models together. The Netflix movie-rating challenge has become one of the most famous examples for big-data analytics (Bennett and Lanning 2007). Netflix is a movie-rental company that launched a competition in 2006 to try to improve their system for recommending movies to their customers. The Netflix dataset has 17,770 movies (columns) and 480,189 customers (rows). The data matrix is very sparse with “only” 100 million (1%) of the ratings present in the training set. The goal is to predict the ratings for unrated movies, so as to better recommend movies to customers. The winning algorithm used a combination of a very large number of statistical techniques, together with a complex pre-processing (Töscher et al. 2009).

It is known that the use of complex models with thousands of parameters can lead to overfitting, greatly reducing the predictive ability of the model and its reliability.

Overfitting occurs when a non-linear model fits observed random errors instead of the “underlying relationships”. Given a sample *D*, once estimated a supervised model $\hat{f}_D$, we want to know its prediction capability on all the possible values of the features and the target. We define Prediction Error (PE):

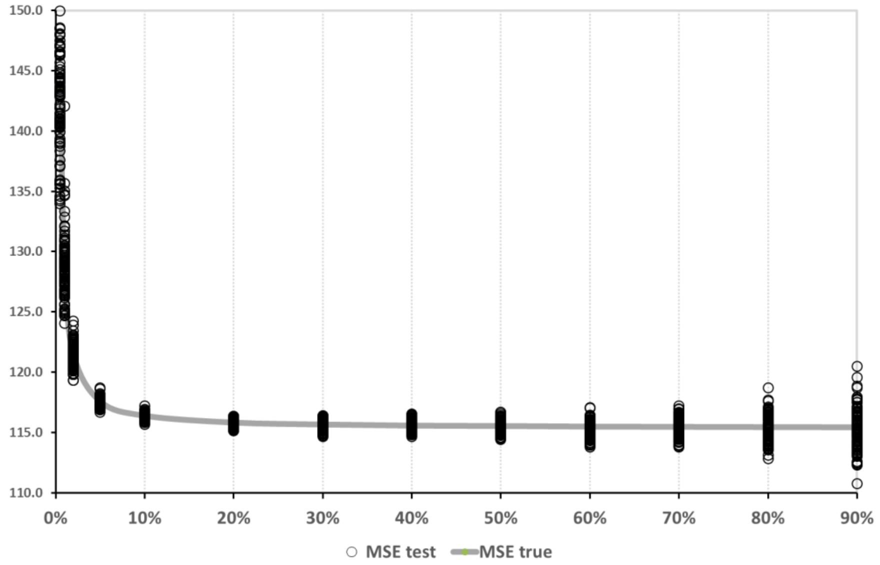
$$PE = E_{\mathbf{x}} E_Y [L(Y, \hat{f}_D(\mathbf{x})) \mid \hat{f}_D]$$

where L(.) is the loss function chosen. With a quadratic loss, the expression is the Mean Squared Error (MSE) computed on the population. In effect, PE is our evaluation of the fit loss for the estimated model on the unobserved population. To assess the model we need to estimate PE, usually by a cross-validation procedure (CV). This consists of splitting the original data-set in two or more parts in order to train the model on a data-set that is different from the data-set used to evaluate the model. In this way, we are able to obtain an evaluation that does not reward the overfitting models.

Several kind of cross-validation procedure can be used to estimate PE: Leave-one-out CV, Hold-out CV, K-fold CV (Di Ciaccio & Borra 2010). If we have a very large number of units, as it is common with big-data, the procedure less expensive is the Hold-out CV, that splits the sample in only two parts, the training and the test sets.

An excessive number of units can be considered an obstacle for statistical analysis. Actually, even when you have archives with millions of units, it is more effective to have a relatively small training set (for example, 30,000 units) leaving all the other units in the test-set. This is a big advantage from a computational point of view because it allows to estimate the model on a reduced data set, but using the other data for the model evaluation. We can see this property, which may seem counterintuitive, through a simulation.

**Figure 7** – *Estimated and true MSE for different hold-out splits (% of training set).*

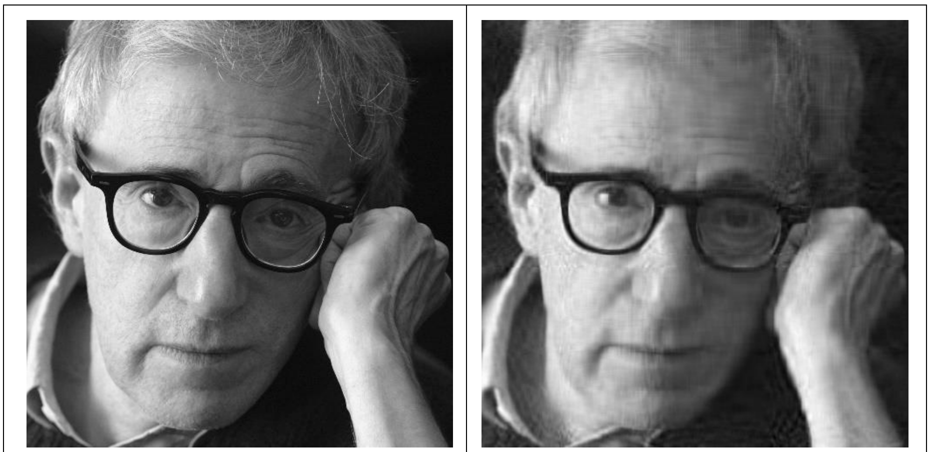


Given a big artificial dataset (100,000 units) with a quantitative target, we applied a small Neural Network to carry out a regression analysis. Hold-out with several split percentages was used to estimate the MSE, replicating the analysis 100 times for each percentage. The ‘true’ MSE was computed on an independent big-dataset. The results of the simulation, shown in fig. 7, display that with a training-size greater than 20,000 (20% of the sample) we obtain only an increase of the estimator variability of MSE (the circles). We can also deduce that with a large training-size and a large test-size, the hold-out estimator is essentially unbiased with low variability.

## 7. Dimensionality reduction

Analysing a database of millions of observations is not impossible for the statistical methods (e.g., by sampling the units). With thousands of variables, the number of input features should be reduced before a machine-learning algorithm can be successfully applied. Moreover, automatic statistical methodologies may be required to provide fast, even real-time, predictions, which would require parsimonious models. Indeed, until a few years ago, all calculations were made off-line, and researchers had time to process data with a relatively small sample size. Currently, with many big-data and online procedures, a different approach is needed.

**Figure 8 -** *Unsupervised dimensionality reduction by PCA (32 components)*



Original image (3800x3800 pixels)

Image rebuilt by PCA

Dimensionality reduction can be performed in two main ways:

- *Unsupervised feature extraction*. Create a smaller set of new features exploiting redundancies in the input data. This can be useful to visualize high-dimension data, seek the intrinsic dimensionality, reduce big-data to manageable dimensions.
- *Supervised feature selection*. Maintain only the most significant features from the original dataset. This allow to discard the irrelevant features, reduce the big-data dimensions, simplify the model interpretation.

In the first approach we can cite the well known *Principal Component Analysis* (PCA), that is a very effective technique if the data structure is almost linear. In fig. 8 it is shown the application of the PCA to a matrix with 3800 rows and 3800 columns corresponding to the image on the left (8-bit grey-scale). The right image is obtained considering 32 columns (the first components of PCA) instead of the original 3800 columns. Effectively, examining fig. 9, it can be observed that 32 components are enough.

If the data structure is non-linear, PCA will overestimate the intrinsic dimensionality of the data. In this case, local approaches are more effective: *local* PCA (Kambhatla & Leen, 1997) or *nearest neighbor algorithm* (Pettis et al. 1979). In the global approaches, in addition to PCA, we can use *manifold methods* or *autoencoder neural networks*.

**Figure 9** – *PCA for the image compression: explained variability ratio of each component.*

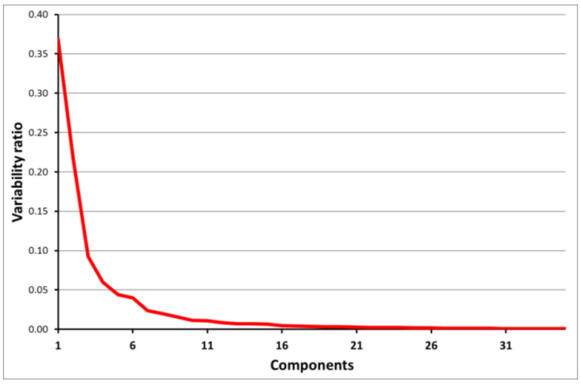
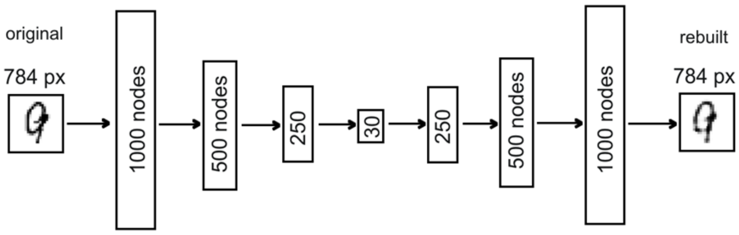
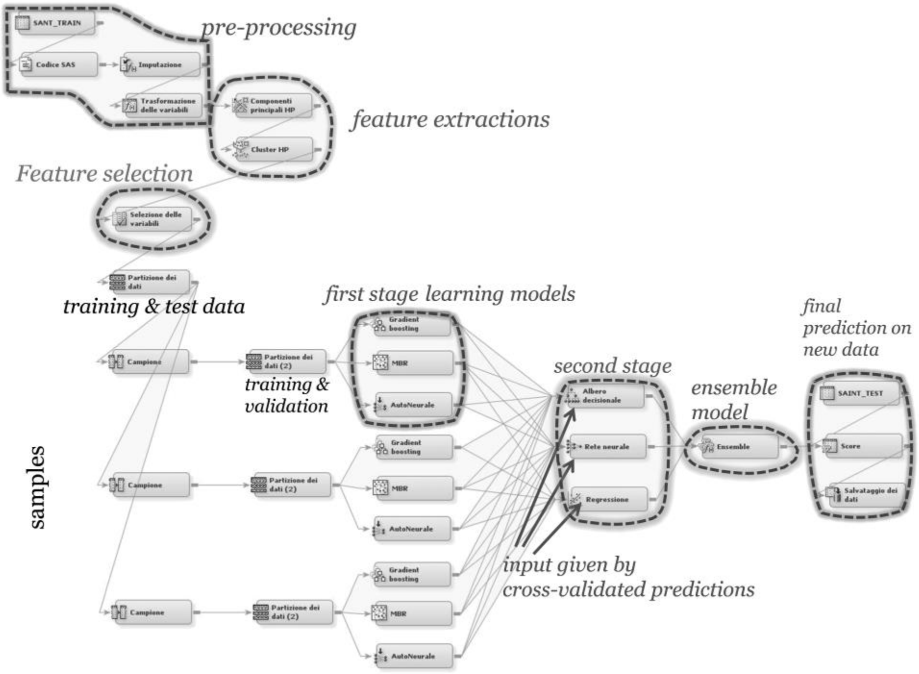


**Figure 10** – *The autoencoder used for MNIST data, with 7 hidden layers: 1000-500-250-30-250-500-1000 nodes.*



An autoencoder (Hinton & Zemel 1994) is a multilayer neural network in which the target values are equal to the inputs. This unsupervised learning algorithm has a small central layer to reconstruct high-dimensional input vectors.

**Figure 11** – *A complex scheme of analysis with re-sampling and ensemble learning (stacking) using SAS Enterprise Miner.*



In fig. 10 it is shown the structure of an autoencoder applied to the MNIST data set (LeCun et al., 1998). It was shown in literature that high-dimensional data can be converted to low-dimensional codes by training an autoencoder network that works much better than PCA (Hinton et al. 2006).

In general, having a lot of units and variables, we can adopt complex analysis schemes, as shown in fig. 11, where pre-processing, feature extraction and feature selection, sub-sampling, ensemble learning, are used together.

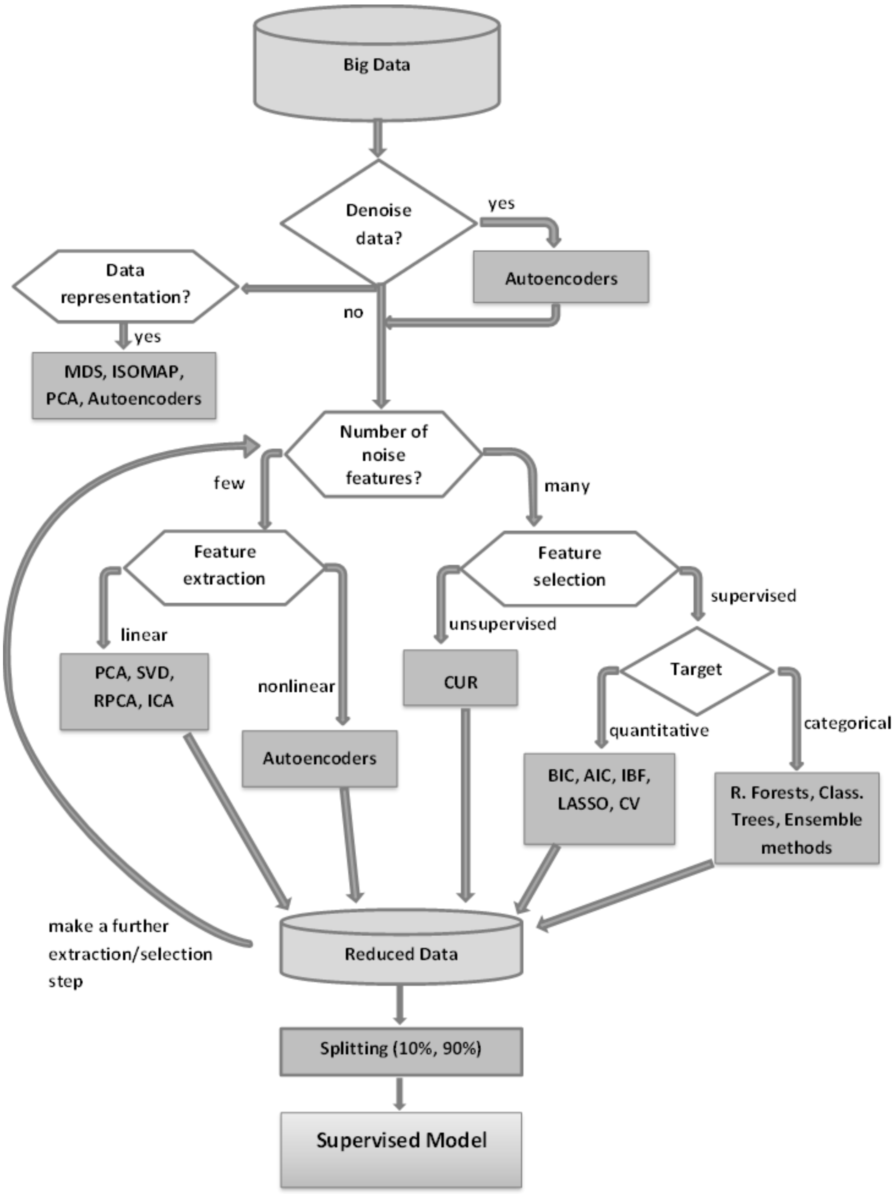
Finally, we may note that also the distinction between supervised and unsupervised approach might not be so strict in the future, if it were possible to obtain a correct classification without having labelled data. Some experiments with very large neural networks on big-data pictures showed that it is possible to train neurons to be selective for high-level concepts using entirely unlabelled data (Le, Ranzato et al. 2012).

## 8. Conclusion

The analysis of big-data can be approached in several ways, but the underlying problem is again a statistical problem. Learning methods for big-data are statistical methods that generalize and modify the classical techniques for the new data sets. With big-data, the underlying aim remain the same, although with more emphasis on data quality, dimensionality reduction, predictive capability and reliability of the learning models. The analysis of big-data requires the ability to assess the effectiveness of the model with a non-parametric inferential approach, usually referred as generalizability and regularization in the language of machine learning. The model can take advantage of the large amount of units available, but an excessive number of variables requires a dimensionality reduction, to avoid the inclusion a large amount of noise that can make ineffective the estimation of the model. As the flowchart of fig. 12 shows, the dimensionality reduction to be chosen depends on the characteristics of data and the objective of analysis.

Finally, it is useless to look with scepticism these huge data and the techniques that are applied, because this will be the main application field of the statistical analysis in the future, "*There is no need to distinguish big-data analytics from data analytics, as data will continue growing, and it will never be small again.*" (Fan & Bifet 2012).

**Figure 12** – *A typical flowchart of a big-data analysis.*



## References

BENNETT, J., LANNING, S. 2007. The Netflix Prize. Paper presented at the KDD Cup Workshop, San Jose, CA, 12 August.

BREIMAN, L. 2001. Random Forests. *Machine Learning* 45 (1): 5–32. doi:10.1023/A: 1010933404324.

BREIMAN, L. 1996. Bagging predictors. *Machine Learning* 24 (2): 123–140. doi:10.1007/BF00058655

DI CIACCIO, A., BORRA, S. 2010. Measuring the prediction error. A comparison of cross-validation, bootstrap and covariance penalty methods, *Computational Statistics & Data Analysis, 2010, vol. 54, issue 12*, pages 2976-2989.

FAN, W., BIFET, A. 2012. Mining Big-data: Current Status, and Forecast to the Future, *SIGKDD Explorations, vol. 14, issue 2*, pp. 1-5.

GEURTS, P., ERNST, D., WEHENKEL, L. 2006. Extremely randomized trees, *Machine Learning*, issue 1, pp 3-42.

HINTON, G.E., ZEMEL, R.S. 1994. Autoencoders, Minimum Description Length, and Helmholtz Free Energy. *Advances in Neural Information Processing Systems 6*. Cowan, Tesauro & Alspector (Eds.), Morgan Kaufmann: San Mateo, CA.

HINTON G.E., SALAKHUTDINOV R.R. 2006. Reducing the dimensionality of data with Neural Networks, *Science, vol. 313, issue 5786*, pp. 504-507.

KAMBHATLA, N., LEEN T.K. 1997. Dimension reduction by local principal component analysis, *Neural Computation 9 (7)*, 1493–1516.

LANEY, D. 2001. 3D Data Management: Controlling Data Volume, Velocity and Variety, *Gartner, Application Delivery Strategies*, 6 February 2001.

LECUN, Y., BENGIO, Y., HINTON, G., 2015. Deep Learning, *Nature 521*, 436–444.

LECUN, Y., BOTTOU, L., BENGIO, Y., HAFFNER, P. 1998. Gradient-based learning applied to document recognition, Proceedings of the IEEE, 86(11):2278-2324.

LE Q.V., RANZATO M., MONGA R, DEVIN M. .et al. 2012. Building High-level Features Using Large Scale Unsupervised Learning, *International Conference in Machine Learning 2012*.

SCHEVENINGEN Memorandum on Big-data and Official Statistics, 2013. DGINS conference.

SKALIOTIS, M., 2015. Big-data in the European Statistical System. Conference by STATEC and EUROSTAT, World Statistics Day 20.10.2015.

SONDERGAARD, P., 2011. Gartner Symposium/ITxpo 2011, October 16-20, Orlando.

TÖSCHER, A., JAHRER, M., 2009. The BigChaos Solution to the Netflix Grand Prize. http://www.stat.osu.edu/~dmsl/GrandPrize2009_BPC_BigChaos.pdf

PEDREGOSA F., VAROQUAUX G., GRAMFORT A. et al., 2011. SCIKIT-LEARN: Machine Learning in Python, JMLR 12, pp. 2825-2830.

SOCHER R., PERELYGIN A., WU J., CHUANG J., MANNING C., NG A., POTTS C. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In EMNLP.

PERRIN A. 2015. “Social Networking Usage: 2005-2015.” Pew Research Center. October 2015. Available at: http://www.pewInternet.org/2015/10/08/2015/Social-Networking-Usage-2005-2015/

PETTIS K., BAILEY, T., JAIN, A., DUBES, R. 1979. An intrinsic dimensionality estimator from near-neighbor information. *IEEE Trans. Pattern Analysis and Machine Intelligence, 1(1)*, pp. 25–37.

UNECE 2013. Classification of Types of Big-data. http://www1.unece.org/stat/platform/display/bigdata/Classification+of+Types+of+Big+Data

## SUMMARY

It is estimated that about 90% of the currently available data have been produced over the last two years. Of these, only 0.5% is effectively analysed and used. However, this data can be a great wealth, the oil of 21st century, when analysed with the right approach. In this article, we illustrate some specificities of these data and the great interest that they can represent in many fields. Then we consider some challenges to statistical analysis that emerge from their analysis, suggesting some strategies.

Agostino DI CIACCIO, Department of Statistical Science,
agostino.diciaccio@uniroma1.it
Giovanni Maria GIORGI, Department of Statistical Science,
giovanni.giorgi@uniroma1.it

# ON THE USE OF INTERNET AS A DATA SOURCE FOR OFFICIAL STATISTICS: A STRATEGY FOR IDENTIFYING ENTERPRISES ON THE WEB[1]

Giulio Barcaroli, Monica Scannapieco, Donato Summa

## 1. Introduction

Internet as a Data Source is gaining more and more importance in Official Statistics. An increasing number of Statistical Institutes are indeed experimenting the use of new sources of data (also known as Big Data) in order to produce the same or new statistical information in a multisource environment, more efficiently and with higher levels of quality (Citro, 2014).

Many examples of the use of Internet data sources can be reported. For instance:

- Internet queries: the use of Google Trends has been evaluated in order to produce now-casting estimates of unemployment indicators (Fasulo et al., 2015);
- Web prices: Web scraping is already in use in order to collect prices related to goods and services for the construction of Consumer Prices Indexes (Cavallo, 2013);
- Social media: posts in social media, like Twitter or Facebook, can be used in order to support the production of traditional Official Statistics indexes like, e.g., the Consumer Confidence Index (Daas et al., 2012).

The nature of data originating from Internet data sources is such that a number of problems have to be addressed, before making use of them for statistical purposes. The most relevant issue is related to representativeness and selectivity (Buelens et al., 2014), In this respect, different important questions arise in order to assess the usability of a given source:

1. Is it possible to refer collected data to specific units in the population of reference, and, if this is the case, with which level of uncertainty?
2. To what extent the population of reference is covered by a specific Internet data source?

[1] Invited paper to the 53rd SIDES Scientific Meeting – Rome 2016.

3. With respect to the target estimates, is the distribution of related variables in the subpopulation covered by the source significantly different from the one not covered?

In this paper, we will consider a specific case of Internet as a data source, that is the one related to the information contained in enterprises' websites. The representativeness of these data is analyzed with particular regard to points (1) and (2) of the above questions. Also, methodological and technological solutions of the problem concerning the minimization of linkage uncertainty and the maximization of population coverage are described.

The Italian National Institute of Statistics (Istat) is experimenting the use of Internet data to obtain a subset of the estimates currently produced by the sampling survey on survey "ICT usage in Enterprises", yearly carried out by Istat and by the other member states in the EU. Target estimates of this survey include the characteristics of websites used by enterprises to present their business (for instance, if the websites offer e-commerce facilities or job vacancies). The aim of the experiment is to evaluate the possibility to use the sample of surveyed data as a training set in order to fit models that will be applied to the generality of websites (Barcaroli et al., 2015-a) (Barcaroli et al., 2016).

Let us consider the task of correctly referring Internet data to population units (in our case, respectively, data collected from websites need to be referred to the population unit "enterprise"). The availability of a population frame[2] containing all the enterprises included in the target population suggests to use it as the basis for a search of the corresponding websites. Actually, for a subset of enterprises the indication of corresponding websites is already available, as they can be obtained by the indications that sampled enterprises gave in the "ICT usage in Enterprises" survey, and from other sources that can be integrated with such information. However, the size of this subset is insufficient to adequately cover the total amount of enterprises that own a website (estimated in about two thirds of the total).

In this paper, we will detail a strategy to address such an issue, which we will call in short "the URLs retrieval problem". In particular, we will describe a specific process that we designed and implemented able to retrieve, for any given enterprise, the URL of the corresponding website, if any.

The paper is organized as follows. In Section 2, background information is given related to the research and development of Big Data projects within which Istat was involved and where this problem was tackled. Section 3 describes the particular strategy adopted for ensuring the maximization of the coverage of the population of interest, while containing errors related to URLs retrieval. In Section

[2] In Istat the available frame is ASIA ("Archivio Statistico delle Imprese Attive", Statistical Archive of Active Enterprises).

4, the results of the application of the strategy are reported and analyzed. Finally, Section 5 contains conclusions and indications on the future work.

## 2. Background Information

### *2.1. Web information extraction pipeline*

Obtaining information from the Web is enabled by using Web scraping techniques. Web scraping (Web harvesting or Web data extraction) is basically a process that permits to extract information from websites.

It is possible to distinguish between two different kinds of Web scraping, namely: *specific Web scraping* and *generic Web scraping*.

*Specific Web scraping* is referred to the case when both structure and content of websites to be scraped are perfectly known, and scraping programs just have to replicate the behavior of a human being visiting the website and collecting the information of interest. A typical area of application is the data collection for price consumer indices; in that case one is typically interested to collect very specific information contained in a website (e.g. a value in a row of a table in a page).

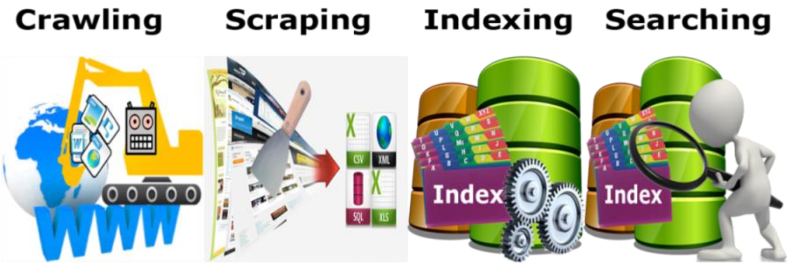
*Generic Web scraping*, instead, assumes that no a priori knowledge on the content is available, and the whole website is scraped and subsequently processed in order to infer information of interest. In such cases it is up to subsequent phases to retrieve any specific information that could be needed. An example of application of generic Web scraping has been performed for the survey “ICT usage in enterprises”.

The Web information extraction process is composed by four main different phases (see Figure 1):

- Crawling: a Web crawler (also called Web spider or ant or robot) is a software program that systematically browses the Web starting from an Internet address (or a set of Internet addresses) and some pre-defined conditions (e.g., how many links navigate, the depth, types of files to ignore, etc.).
- Scraping: a scraper takes Web resources (documents, images, etc.), and engages a process for extracting data from those resources, finalized to data storage for subsequent elaboration purposes.
- Indexing/Searching: searching operations on a huge amount of data can be very slow, so it is necessary (through crawler) to index contents.

Once these phases are completed it is possible to analyze the indexed contents in order to extract relevant information from the collected data and produce the needed outputs.

**Figure 1 –** *Web information extraction process*



## *2.2. Official Statistics Projects*

In Section 1, we cited the National project concerning the use of Internet as a data source for the "Survey on ICT usage in Enterprises". This case can be classified as a case of generic Web scraping. Indeed, on one side the number of websites that was necessary to scrape for the purpose of this project was consistent (more than 8000), while on the other side it was not possible to rely on a fixed and known in advance structure of the websites (Barcaroli et al., 2015-a) (Barcaroli et al., 2016).

Further projects at National level were carried out within CBS, the Dutch National Statistical Office, as well as by Istat: these project are rather examples of specific Web scraping. In particular, in (Ten Bosh et al., 2014) a first domain of experimentation was related to air tickets: the prices of air tickets were collected daily by Internet robots, developed by Statistics Netherlands supported by two external companies, and the results were stored for several months. The experiment showed that there was a common trend between the ticket prices collected by robots and existing manual collection. Two additional domains of experimentation were Dutch property market and Clothes prices, the first exhibiting more regularity in the sites structure, the latter more challenging with respect to automatic classification due to lack of a standard naming of the items, and variability in the sites organization. Similarly, in Italy a scraping activity was performed to get on consumer electronics prices and airfares (Giannini et al., 2014).

More recently, a Eurostat funded project started (ESSnet Big data, 2016) with two work-packages dedicated to Web scraping, namely one to “Web scraping of job vacancy” and another one to “Web scraping of enterprise characteristics”.

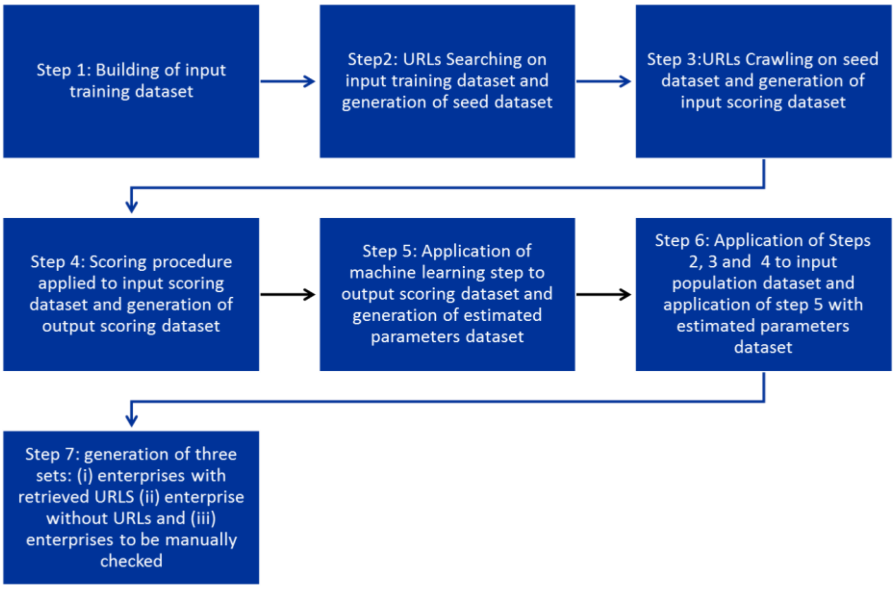
## 3. The Proposed Strategy

This Section will describe the strategy adopted for solving the URLs retrieval problem. We will start by providing an overview of the approach (Section 3.1), then we will detail the scraping process (Section 3.2, 3.3, 3.4, 3.5), and  we later detail the learning process (Section 3.6, 3.7).

### *3.1. Overview of the Strategy*

In order to address the URLs retrieval problem, we developed a step-wise strategy shown in Figure 2.

**Figure 2 –** *Steps of the overall strategy*



Step 1: Building of input training dataset

Step2: URLs Searching on input training dataset and generation of seed dataset

Step 3:URLs Crawling on seed dataset and generation of input scoring dataset

Step 4: Scoring procedure applied to input scoring dataset and generation of output scoring dataset

Step 5: Application of machine learning step to output scoring dataset and generation of estimated parameters dataset

Step 6: Application of Steps 2, 3 and  4 to input population dataset and application of step 5 with estimated parameters dataset

Step 7: generation of three sets: (i) enterprises with retrieved URLS (ii) enterprise without URLs and (iii) enterprises to be manually checked

In the following sections we will detail each step and we will provide a concrete running example of the strategy as applied to solve the URLs retrieval problem in the case of the survey “ICT Usage in enterprises”.

### *3.2. Step 1: Building the input training dataset*

In this step, the input training dataset is built by considering the units for which the URLs are known as well as the specific features of such units that could be useful to identify them on the Web.

In our case study, we built such a dataset by integrating different sources, namely: third party information (Consodata) and “ICT in enterprises” survey editions of 2013, 2014, 2015. The resulting dataset was composed by 81912 enterprises with URLs.

### *3.3. Step 2: URLs Searching*

In this step, the objective is to retrieve for each unit in the input training dataset, one or more URLs to scrape based on identifying information that is present in such a dataset.

In our case study, we decided to set up an automated procedure that used “enterprises’ denominations” (in Italian “Ragioni Sociali”) for searching the enterprise on the Internet via a search engine. In particular, we used the denomination of the enterprise as a search string, then we queried the search engine and collected the first ten links returned as the result of the search query.

Each of the link was visited in order to estimate the official one. These operations might look very simple if done by humans, but they are indeed very time consuming especially if you have to deal with a huge number of enterprises. Hence, we decided to automate this step by writing a custom Java program that takes as input two files containing the list of the enterprises’ names and the corresponding list of enterprises’ IDs. For each enterprise the program queries a search engine (we are currently using Bing search engine) and retrieves the list of the first ten URLs provided by the search engine.

These URLs are then stored in a txt file, thus having one txt file for each enterprise. At the end, the program reads each produced file and creates the seed dataset, containing the following fields: URL_retrieved, enterprise_Id and position_of_the_URL. Basically we can say that the seed dataset contains the list of potential enterprise official URLs to be scraped.

When we had to choose the search engine to use we tested Google, Istella, Bing and Yahoo in order to determine the most suitable for our objectives. We tried to execute a batch of 10000 automatic queries on each of them, Google and Yahoo blocked us after about 1000 queries so we had to discard them. We focused on Bing and Istella, in both cases we considered about 7000 enterprises with a known site; for each enterprise name we collected the first 10 results from the search engine and compared them with the known site. Using Istella we obtained 50% of matches and not always in the first positions, using Bing we obtained a success rate of 65% so we chose it due to the better obtained results.

### *3.4. Step 3: URLs Crawling*

In this phase the objective is to retrieve for each potential URL of the seed dataset the textual content of the corresponding page. We tested different software solutions for crawling already available off the shelf but they did not satisfy completely our needs, in particular they were not as flexible as we needed (Barcaroli et al., 2015-b).

We decided to develop a custom Java program based on crawler4j (https://github.com/yasserg/crawler4j) and jsoup (https://jsoup.org/). The program takes as input 3 files:

- the seed file produced in the search phase
- a list of URL domains to filter out (directories domains)
- a configuration file

For each row of the seed file, if the URL is not in the list of the domains to filter out, the program tries to acquire the HTML content of the page. From each acquired HTML page the program extracts just the textual content of the HTML fields we are interested in and write a line in a TSV file.

The TSV file just produced is then indexed and loaded in SOLR (http://lucene.apache.org/solr/), an open source enterprise search platform built on top of Apache Lucene.

The resulting document base constitutes the input scoring dataset, to be passed to the following step.

### *3.5. Step 4: URLs Scoring*

In this step, for each document of the input scoring dataset, a score vector is computed and a score is assigned.

For our case study, the resulting dataset, i.e. the output scoring dataset, contains the following fields: *enterpriseId, linkPosition, URL, scoreVector, score.*

In the rest of this section we detail the procedure to compute the score, named URLScorer (see **Errore. L'origine riferimento non è stata trovata.** ).

Starting from the input scoring dataset, for each link URLScorer generates a vector of numbers. Every position in the vector contains a number that means “a specific characteristic in the Web page was found or not” (e.g. presence of the telephone number).

For our case study, the elements/characteristics that we considered are the following:

- Simple URL (is the URL in the form www.name.com or not ?)
- VAT (is it present in the page or not ?)
- city (is it present in the page or not ?)
- province code (e.g. NA, is it present in the page or not ?)
- link position (from 0 that means that that was the first link provided by the search engine to 9)
- telephone number (is it present in the page or not ?)
- zip code (is it present in the page or not ?)

For each element/characteristic we computed the confusion matrix obtained by using just that element for classification purposes. Based on that confusion matrix we computed the standard performance measures, i.e. precision, recall and f-measure. Then, we assigned the f-measures of the corresponding confusion matrixes as raw weights to elements; lastly we normalized the raw weights so that the sum of the final (adjusted) weights is 1000 (i.e. normalized weights in **Errore. L'origine riferimento non è stata trovata.**). In order to assign a final score to a link, we summed up the normalized weights of those elements that were actually present in the vector. In a nutshell, we multiplied each number of the vector with a specific coefficient and summed up all the results in order to obtain a score for that link.

To validate the results of URL Scorer, we performed a validation task as follows:

- we selected 20000 enterprises with a known website;
- we compared the URL estimated by the procedure with the highest score with the official URL that we already knew. In particular, comparing only the domain part of the URL (e.g. "rossi.it" if the URL is "www.rossi.it/aboutUs") we found exact matches in 64% of the cases.

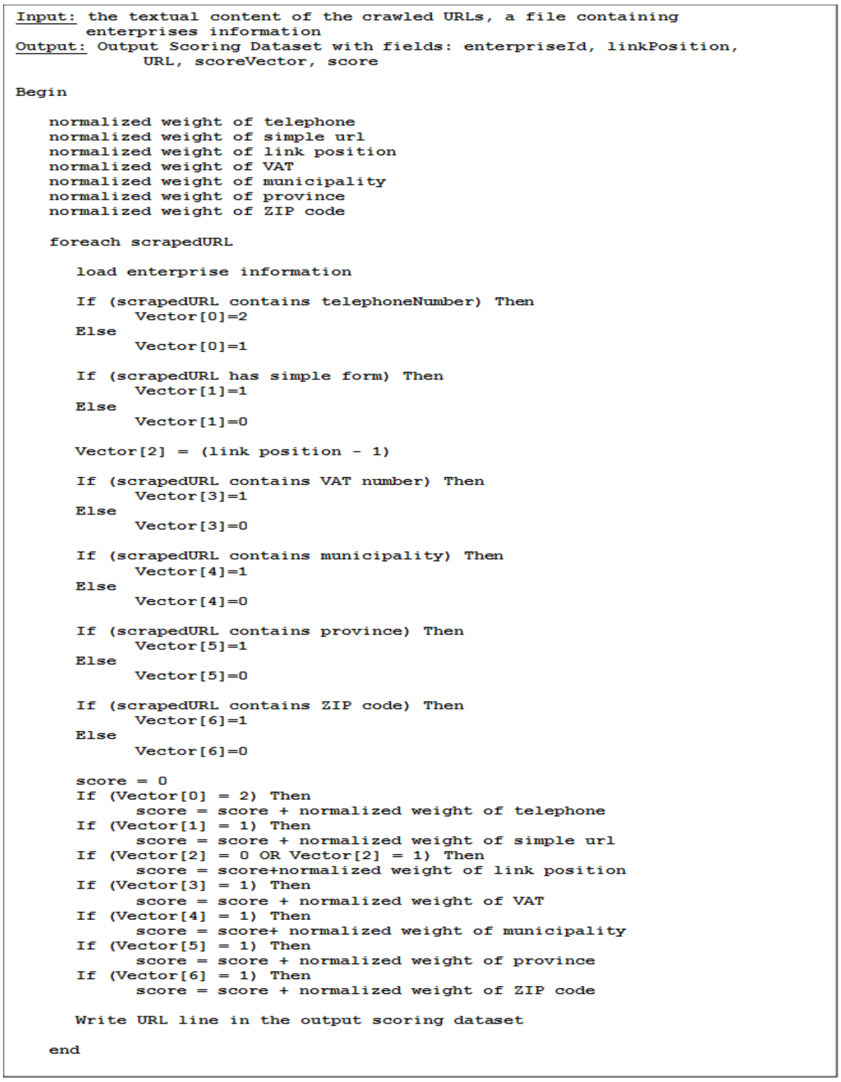
The real success percentage is probably higher than the one obtained because sometimes the official Web site that we know is not correct due to several reasons, including:

- it is outdated (the URL is changed);
- it does not exist anymore;
- wrong domain (e.g. “rossi.it” instead of “rossi.com”);
- it is only an information page with contacts (e.g. enterprise contact portals like Italian “paginegialle.it))

- it is the site that sells the products of the enterprise (e.g. enterprise e-commerce portals, like Italian “mediaworld.it”)
- it is the site of the mother company (franchising enterprises).

In any case, we chose to improve such a result by performing the machine learning step detailed in the following section.

**Figure 3 –** *Pseudo-code of URLScorer*



```
Input: the textual content of the crawled URLs, a file containing
       enterprises information
Output: Output Scoring Dataset with fields: enterpriseId, linkPosition,
            URL, scoreVector, score

Begin

   normalized weight of telephone
   normalized weight of simple url
   normalized weight of link position
   normalized weight of VAT
   normalized weight of municipality
   normalized weight of province
   normalized weight of ZIP code

   foreach scrapedURL

      load enterprise information

      If (scrapedURL contains telephoneNumber) Then
            Vector[0]=2
      Else
            Vector[0]=1

      If (scrapedURL has simple form) Then
            Vector[1]=1
      Else
            Vector[1]=0

      Vector[2] = (link position - 1)

      If (scrapedURL contains VAT number) Then
            Vector[3]=1
      Else
            Vector[3]=0

      If (scrapedURL contains municipality) Then
            Vector[4]=1
      Else
            Vector[4]=0

      If (scrapedURL contains province) Then
            Vector[5]=1
      Else
            Vector[5]=0

      If (scrapedURL contains ZIP code) Then
            Vector[6]=1
      Else
            Vector[6]=0

      score = 0
      If (Vector[0] = 2) Then
            score = score + normalized weight of telephone
      If (Vector[1] = 1) Then
            score = score + normalized weight of simple url
      If (Vector[2] = 0 OR Vector[2] = 1) Then
            score = score+normalized weight of link position
      If (Vector[3] = 1) Then
            score = score + normalized weight of VAT
      If (Vector[4] = 1) Then
            score = score+ normalized weight of municipality
      If (Vector[5] = 1) Then
            score = score + normalized weight of province
      If (Vector[6] = 1) Then
            score = score + normalized weight of ZIP code

      Write URL line in the output scoring dataset

   end
```

### *3.6. Using a Machine Learning approach to associate URLs to enterprises*

The scoring procedure implemented by URL Scorer produces, for each link associated to a given enterprise, a scoring vector, in which each single position indicates success or failure of the search of a particular information in the website indicated by the link, plus a total score derived from the scoring vector.

A natural choice to select a link as the correct one for the given enterprise could be the one with the maximum score. Unfortunately, this choice would not prevent to choose a non-correct link. In fact, consider that:

- about one third of enterprises do not own a website: for those enterprises, all the links obtained by the search engine are not correct by definition;
- the correct website might not be included in the set of the first 10 links for a number of reasons.

For these reasons, we have to define a more complex decision rule.

Taking into account that a subset of enterprises for which the correct link is already indicated is available, it is possible to adopt a *machine learning* approach under which a model is fitted in this “training” set, and then applied to the set represented by all other enterprises.

In our case study, our input training dataset consisted of 81912, of which 73006 records had at least one page fetched. On the basis of the output scoring dataset we first associated to each enterprise of the 73006 sized set the link with the highest score. As we know if the link is correct or not, a dichotomous variable *correct_Yes_No* says if the URL is the right one or not: this variable plays the role of the Y variable, to be predicted by the model. Together with this information, variables indicating success or failure of the search of *telephone*, *VAT code*, *municipality*, *province* and *zip code* play the role of the X variables (predictors), together with the *link position* and coincidence of the central part of the URL with the name of the enterprise (*simple URL*).

This initial set is split into two equal size subsets, the first acting as the proper training set to fit the model, the second as the test set used to evaluate the performance of the model.

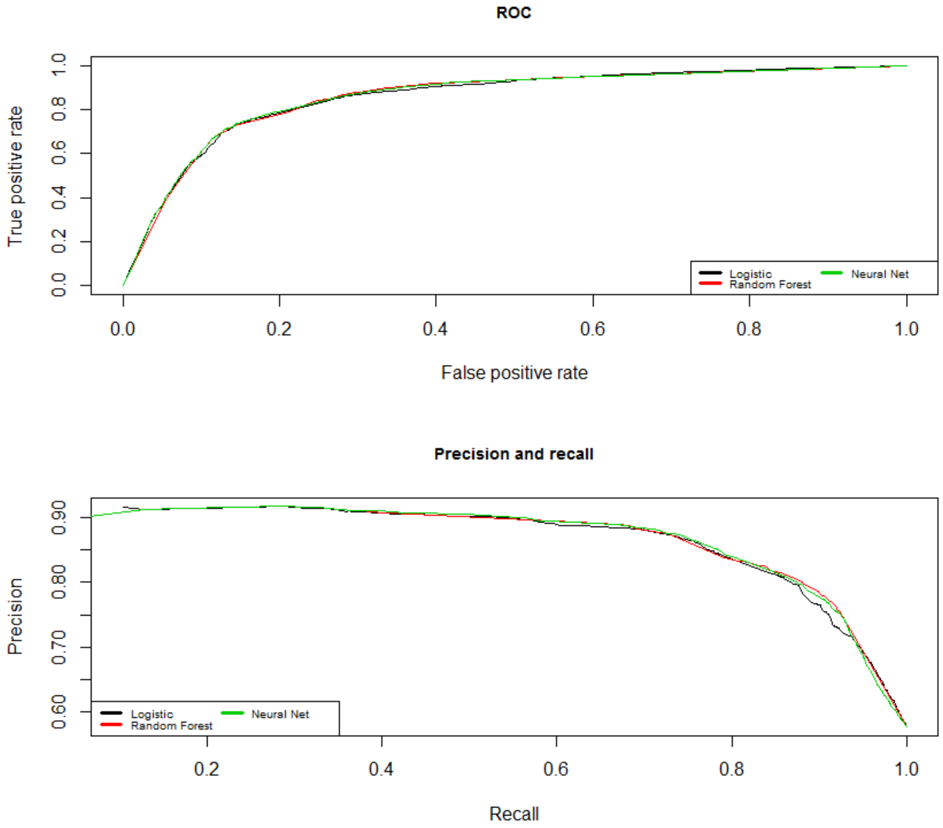
**Table 1 – Evaluation of Neural Networks, Random Forest and Logistic Model.**

| Learner | Accuracy | Sensitivity | Specificity | F-measure |
|---|---|---|---|---|
| Neural Networks | 0.7960 | 0.8011 | 0.7890 | 0.8194 |
| Random Forest | 0.7999 | 0.8278 | 0.7616 | 0.8270 |
| Logistic Model | 0.7918 | 0.7857 | 0.8002 | 0.8135 |

Different learners have been fitted and evaluated, namely Neural Networks, Random Forest and Logistic Model. Their performance has been evaluated by considering the classic indicators, that is accuracy, sensitivity, specificity and F-measure (harmonic mean of recall and precision). Their values are reported in Table 1.

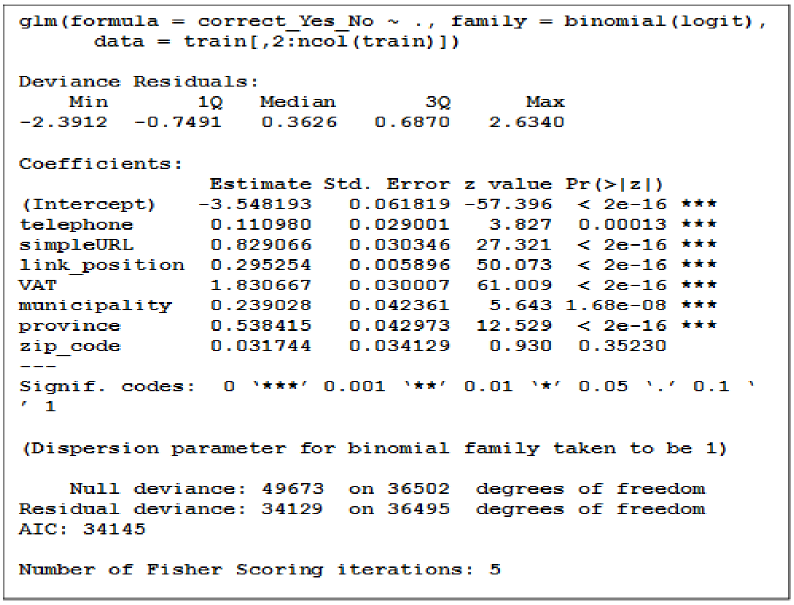
The difference in performance is not significantly different for the three learners, this can be seen also visualizing the ROC and the curves of precision/recall in Figure 4.

**Figure 4 –** *Performance of the three learners*



Taking into account the statistical properties of the logistic model, this learner has therefore been preferred to the others, also because of the interpretation of the score as a probability. In Figure , the fitting results of the logistic model applied to the training set are shown.

**Figure 5 –** *Logistic model fitting*



```
glm(formula = correct_Yes_No ~ ., family = binomial(logit),
      data = train[,2:ncol(train)])

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3912  -0.7491   0.3626   0.6870   2.6340

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)   -3.548193   0.061819 -57.396  < 2e-16 ***
telephone      0.110980   0.029001   3.827  0.00013 ***
simpleURL      0.829066   0.030346  27.321  < 2e-16 ***
link_position  0.295254   0.005896  50.073  < 2e-16 ***
VAT            1.830667   0.030007  61.009  < 2e-16 ***
municipality   0.239028   0.042361   5.643 1.68e-08 ***
province       0.538415   0.042973  12.529  < 2e-16 ***
zip_code       0.031744   0.034129   0.930  0.35230
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 49673  on 36502  degrees of freedom
Residual deviance: 34129  on 36495  degrees of freedom
AIC: 34145

Number of Fisher Scoring iterations: 5
```

Once applied to the test set, units have been sorted in ascending order with respect to the score assigned by the logistic model, and have been grouped in 10 balanced classes (Table 2).

**Table 2 –** *Class of units by their scores*

| Group | scores | True positives | False positives | Error rate |
|---|---|---|---|---|
| 1 | [0.0312,0.124] | 440 | 3239 | 0.88 |
| 2 | (0.124,0.254] | 628 | 3312 | 0.88 |
| 3 | (0.254,0.369] | 859 | 2569 | 0.75 |
| 4 | (0.369,0.507] | 1473 | 2290 | 0.61 |
| 5 | (0.507,0.573] | 1895 | 1555 | 0.45 |
| 6 | (0.573,0.725] | 3038 | 830 | 0.21 |
| 7 | (0.725,0.862] | 2958 | 561 | 0.16 |
| 8 | (0.862,0.921] | 3188 | 428 | 0.12 |
| 9 | (0.921,0.936] | 1397 | 141 | 0.09 |
| 10 | (0.936,0.943] | 5222 | 480 | 0.08 |

By taking all the links in a given class, the error rate depends on the number of false positives in that class. It is clear that the error rate decreases as the score (i.e.

the probability of correct link) increases. If the *acceptation* threshold value is set to of 0.573 as the one to decide if a link is correct or not, the upper five classes are accepted *in toto* and the mean error that can be expected is 0.13, and the total recall is 0.75. In other words, 75% of correct links can be found, together with 13% or erroneous ones.

If the *refusal* threshold value is set to 0.507, the lower five classes are discarded, losing in this 23% of correct links.

It is possible to choose the 5th class (containing about 9.5% of cases), where true and false positives are balanced, as the one to be controlled interactively (for instance, by adopting a *crowdsourcing* platform that is one of our planned future steps).

### *3.7. Step 6: application of the whole procedure to the input population and Step 7: generation of the results*

Once learned the model in Step 5, this step consists of the application of steps 2,3,4 and 5 to the whole <u>input population.</u>

For our case study, i.e. the survey "ICT usage in enterprises", the population of interest of the survey is composed by enterprises with at least 10 employees and operating in different branches of industry and services, the size of such a population is around 200000. By the ICT survey estimates, it is known that about 70% of these enterprises do own a website, used for different purposes.

As described in Section 3.6, starting from our <u>input train set</u> (specifically the part of it with at least one page fetched, of size 73003) a logistic model has been fitted, and threshold values chosen, to be used in order to find additional URLs for remaining enterprise websites.

So, on the complementary set of enterprises for which the URLs are not available, the three steps procedure (searching, crawling and scoring) has been applied. Here, we report the results of the application of the logistic model to the 106019 enterprises for which URLs were not available (i.e. 205759-73006=132753, of these at least one link has been crawled for 106019):

- 26097 (24.6%) URLs have been evaluated as reliable, and associated to the corresponding enterprises;
- 68885 (64.9%) have been considered as erroneous, and excluded;
- 11037 (10.4%) have been addressed to interactive controls.

This latter component is expected to contain about 45% of false positive (see Table 2, the 5th class). So, after the controls about 55% of the 11037 URLs, let us say 6000, should be individuated as correct.

In conclusion, at the end of the process we should obtain a total number of identified URLs equal to about 105000. If we consider a total amount of 130,000 websites pertaining to the enterprises population, we obtain a coverage of near 81%, which can be deemed a satisfactory one.

## 4. URLs Searching and Crawling Performance

We set up a testing environment with the characteristics shown in Table 3. Let us notice that, being the task massive and resource consuming, this environment is quite under-sized (in terms of RAM and CPU).

**Table 3 –** *Environment configuration*

| | |
|---|---:|
| *CPU* | 2 cores running at 2.2 GHz |
| RAM | 16 GB |
| O.S. | Red Hat Enterprise Linux 7 |
| Kernel version | 3.10.0-327.22.2.el7.x86_64 |

Table 3 shows the performance for our running case study.

The execution time of searching and crawling programs takes several hours: this means that explicitly programmatic control to manage failures have been designed and developed in order to manage this long-running feature and get at a result. In terms of dimension of the generated files, being it several Giga bytes it was necessary to adopt a dedicated storage platform (Apache SOLR), as anticipated in Section 3.4. The usage of this platform permitted an indexed access to the generated document base.

## 5. Conclusions and Future Work

In this paper, we showed an overall strategy for solving the URLs retrieval problem, i.e. the problem of finding a URL corresponding to a statistical unit we are interested to. We detailed the strategy for the case of retrieving URLs of enterprises, with respect to a specific survey run at National and European level, i.e. the survey "ICT usage in enterprises". The strategy involved several steps with a mix of techniques, ranging from scraping and crawling techniques to machine learning ones. Our results show the feasibility of addressing this problem with a partially automated solution that gets good results both in terms of quality and efficiency.

We have planned as future work to adopt a crowdsourcing platform to manage the interactive task related to enterprises we were not able to manage automatically. Indeed, in our case study around ten thousands enterprises required such a manual effort, which is quite a huge number. We are currently evaluating Crowdsearcher (http://crowdsearcher.search-computing.it/) as a platform for this task.

**Table 4 –** *Performances*

| | TrainSet | TestSet |
|---|---:|---:|
| # of enterprises | 81912 | 132753 |
| UrlSearcher execution time | 14h 3min | 22h 17min |
| # urls in seed file | 814577 | 1321323 |
| UrlCrawler execution time | 8h 39min | 13h 4min |
| # urls filtered out | 470039 | 846052 |
| # urls after filter | 344538 | 475271 |
| # urls reached | 241202 | 305488 |
| % of reached urls | 70.01 | 64.27 |
| # of enterprises found | 76976 | 117998 |
| # of enterprises with 0 pages fetched | 3970 | 11979 |
| # of enterprises with at least 1 page fetched | 73006 | 106019 |
| Output CSV file size | 8.6 GB | 10.1 GB |

## References

BARCAROLI G., NURRA A., SALAMONE S., SCANNAPIECO M., SCARNÒ M., SUMMA D. 2015-a. Internet as Data Source in the Istat Survey on ICT in Enterprises. *Austrian Journal of Statistics*, Vol. 44, pp. 31-43.

BARCAROLI G., SCANNAPIECO M., SUMMA D., SCARNÒ M. 2015-b. Using Internet as a Data Source for Official Statistics: a Comparative Analysis of Web Scraping Technologies. Available at: http://www.websm.org/uploadi/editor/doc/1437484121Barcaroli-etal_WebScraping_Final_unblinded.pdf

BARCAROLI G., BIANCHI G., BRUNI R., NURRA A., SALAMONE S., SCARNÒ M. 2016. *Machine learning and statistical inference: the case of Istat survey on ICT*. 48th Scientific Meeting of the Italian Statistical Society. Proceedings ISBN: 9788861970618

BERESEWICZ M. E. 2015. On representativeness of internet data sources for real estate market in Poland. *Austrian Journal of Statistics*, Vol. 44, N.2, pp.45–57.

BUELENS B., DAAS P. J. H., BURGER J., PUTS M., VAN DEN BRAKEL J. 2014. Selectivity of Big Data. Available at: https://www.cbs.nl/nl-nl/achtergrond/2014/14/selectivityof-big-data.

CAVALLO A. 2013. Online and official price indexes: Measuring Argentina’s inflation. *Journal of Monetary Economics*, Vol.60, N.2, pp.152–165.

CITRO C. F. 2014. From multiple modes for surveys to multiple data sources for estimates. *Survey Methodology*, Vol.40, N.2, pp.137–161.

DAAS P., ROOS M., VAN DE VEN M., NERONI J. 2012. Twitter as a potential data source for statistics. *Discussion paper (201221)*. Central Bureau of Statistics. Den Haag/Heerlen, 2012

ESSNET BIG DATA 2016. Eurostat funded Project – Essnet on Big Data, https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/index.php/ESSnet_Big_Data

FASULO A., D’ALÒ M., FALORSI S. 2015. Provisional estimates of monthly unemployment rate using Google Trend. *Proceedings of ITACOSM*. Rome, June 2015.

GIANNINI R., LO CONTE R., MOSCA S., POLIDORO F., ROSSETTI F. 2014. Web scraping techniques to collect data on consumer electronics and airfares for Italian. *HICP compilation*, Q2014 European Conference on quality in official statistics Vienna, 2-5 June 2014.

TEN BOSH O., WINDMEIJER D. 2014. On the Use of Internet Robots for Official Statistics. In MSIS-2014. URL:http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.50/2014/Topic_3_NL.pdf.

## SUMMARY

### On the Use of Internet as a Data Source for Official Statistics: a Strategy for Identifying Enterprises on the Web

In this paper, we propose an overall strategy for solving the URLs retrieval problem, i.e. the problem of finding a URL corresponding to a statistical unit we are interested to. We detail the strategy for the case of retrieving URLs of enterprises, with respect to a specific survey run at National and European level, i.e. the survey “ICT usage in enterprises”. The strategy involves several steps with a mix of techniques, ranging from scraping and crawling techniques to machine learning ones. Our results show the feasibility of addressing this problem with a partially automated solution that gets good results both in terms of quality and efficiency.

Giulio BARCAROLI, Istat, barcarol@istat.it
Monica SCANNAPIECO, Istat, scannapi@istat.it
Donato SUMMA, Istat, donato.summa@istat.it

# THE SCANNER DATA IN THE CONSUMER PRICE SURVEY: THE DATA PROCESSING

Antonella Bernardini, Alessio Guandalini, Francesca Inglese, Marco D. Terribili

## Introduction

The Italian National Institute of Statistics (ISTAT) is planning a redesign of the Consumer Price Survey (CPS). The main aim of the project is to modernise the survey. One of goal is to improving and unburdening the data collection phase. Furthermore, another goal is to progressively introduce more rigorous sampling procedures, probabilistic where possible, for the selection of outlets and products for the sectors where this is feasible (Bernardini *et al.*, 2016).

Register of local units and availability of scanner data (SD) from retail modern distribution, provided through an agreement with Nielsen[1], are the starting point of this transformation. The SD represents a big opportunity for introducing improvements in terms of both data collection and sampling perspective. Furthermore, the use of web-scraping for collecting prices from online market (i.e, tourism, mobile phone) is involved in the review of the survey.

Since the end of 2014, ISTAT, through a contract with Nielsen and an agreement with the six main retail chains operating in Italy, started receiving SD referred to food and grocery markets and treating them with the objective of experimenting the computation of the consumer price index (CPI).

The use of SD for this purpose implies the definition and the implementation of different checks to be made in both data acquisition and processing phases.

This work aims to describe the scanner data checks implemented in the experimental phase. The collection of high-frequency data, as SD are, requires the use of systematic and automatic checks, both formal and quality checks on the data flow . Formal checks have to ensure completeness of data collected at provincial level, distribution chains, outlets, products and weeks. Quality checks introduce editing rules which identify inadmissible values on the variables of interest (quantities sold, turnover and prices).

[1] Nielsen is a global information, data, and measurement company that gathers data on consumers in more than 100 countries to give view of trends and habits worldwide. For more details see e.g. http://www.nielsen.com/eu/en.html.

Completeness and correctness of the data are two important pillars for a correct use of SD and for an accurate computation of the price index over time (Vermeulen and Herren, 2006; Van der Grient et al., 2010).

The paper is organized as follows: section 2 describes the SD structure; section 3 focuses on the checks defined for treating SD and presents the methods used in the data processing phase to identify inadmissible prices; section 4 shows some important results regarding data cleaning; in section 4 some assessments carried out on SD after data cleaning phase are reported; section 5 presents the analysis of the impact of data cleaning on micro-indices; finally, some conclusions and future developments are exposed.

## 1. Scanner data structure

The customary furniture of SD files contains elementary information referred to single European Article Number (EAN) codes (GTIN) for specific outlets, consisting in turnover and quantities sold during a week. This information does not provide the “shelf price” of the product identified by the EAN code and outlet, but it allows to define a unit price or average weekly price. This feature introduces a crucial issue in the discussion among economists, which goes beyond the focus of this paper. Furthermore, usually SD do not include information about discounts or special sales.

Actually, ISTAT contract with Nielsen foresees the provision of weekly data of turnover and quantities, at EAN code and outlet level, for hypermarket and supermarket of six modern retail distribution chains (Conad, Coop, Esselunga, Auchan, Carrefour, Selex) operating in the food and grocery market, in 35 Italian provinces. For the experimental phase Nielsen provided backward data, for at least one full year and the preceding month of December, starting from December 2013 or 2014 (depending on the starting point of delivery of each province). The coverage of the six chains with respect to the whole modern distribution is quite high in all provinces with, obviously, variability among geographical areas of the country.

Furthermore, Nielsen provides and updates the dictionary for the classification of EAN codes to GS1-ECR-Indicod product classification. 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 and gathers together homogeneous products. The COIOPOP related to food and grocery consumption segments are 121 out of a total of 324.

Moreover, for operational constraints a restriction is introduced regarding the observable weeks: only the relevant weeks are considered, defined as the first three full weeks (composed of seven days) in each month (ILO, 2004). Also in the recommendation drafted by EUROSTAT, the use of data referred to the first two full week for each month is advised.

## 2. The treatment of data

### *2.1. Formal and quality checks*

In the experimental stage formal checks have been studied and implemented on the flow of SD by Nielsen both during and after the data loading.

The formal checks during the data loading enables:

- the presence of only occurrences (a product identified by EAN sold in an outlet in a week) with full numeric code for the outlets and products;
- the presence of only occurrences with numeric and valid decimal values for turnover and quantities;
- the presence of only occurrences related to the period of interest.

The formal checks after the data loading concern:

- the presence of duplicates for outlets, products and weeks;
- the presence of a product not included in the list updated every two months;
- the absence of an outlet among the required outlets;
- the presence of null fields of turnover and quantities;
- the presence of unauthorized data such as outlets which do not belong to the authorized provinces or to the allowed chains, or outlets not classified as a hypermarket or supermarket.

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

The quality checks follow the phase of formal checks. They aim to introduce editing rules for identifying inadmissible values among the variables of interest (quantities sold, turnover and unit prices).

First-type quality checks are implemented to identify and eliminate the problematic occurrences (outlet, EAN code, week) in which:

- quantity<1 are not motivated by unit of measurement;
- decimal values on quantities>1 are not motivated by unit of measurement;
- unit prices <= 0,01 €.

Subsequently, in order to maintain an accurate and stable price index over time, a second-type of quality check has been implemented. The prices of each product are validated considering the price labels, during the relevant weeks of every month at provincial level.

## *2.2. Trimming methods*

To identify inadmissible unit prices of occurrences, several methods – more or less sophisticated – have been tested. For the sake of simplicity, also in terms of computational burden, the choice has been reduced between two methods.

Both methods are based on the computation of the median unit prices, considering the quantities sold for each single occurrence.

A simple solution could consist in a fixed trimming method that defines the tolerance interval of prices as:

$$\left( \frac{Median_w}{K_1}, K_2 * Median_w \right).$$

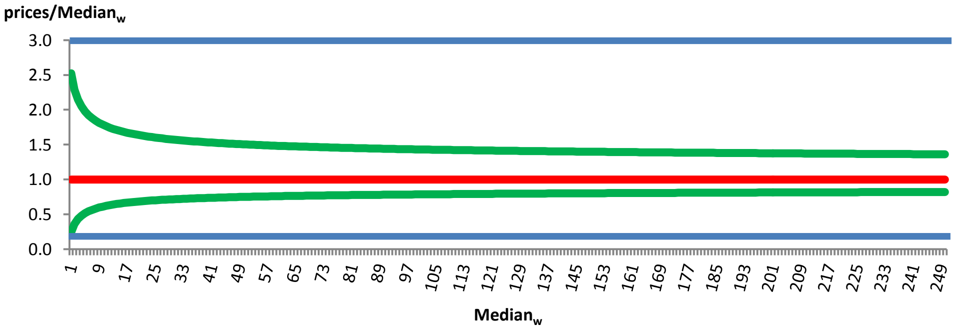
A most suitable of fixed trimming can be named moving trimming and the related tolerance interval is defined as:

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

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 unit price of the product considered acceptable.

For food and grocery, $K_1$=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.

**Figure 1 –** *Relative tolerance interval limits as the weighted (with quantities sold) median of unit price of occurrences sold in the month in a province, Median$_{w,}$ with fixed (blue lines) and moving (green lines) trimming.*



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 (see Figure 1). Therefore, the moving trimming is preferable with respect to the fixed one because it narrows down, thanks to the log function, the extremes as the median price of the product increases.

## 3. Main results

### *3.1. Example of application*

In the moving trimming method, as said above, the relative allowed range shrinks contextually with the increase of the median unit price of occurrences sold in the month in the province (Figure 1).

Herein, just a few examples of how the method works when applied to products with different unit prices are presented.

In the Figures below, the density functions of the unit prices and the distribution of prices and sold quantities for 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 unweighted median price (median unit price of occurrences in a month in the province) and weighted median price (median unit price taking into account the sold quantities of the occurrences in a month in the province) of the product. The values of inadmissible occurrences, that is occurrences with inadmissible unit prices, are indicated by red dots, while green dots indicate the presence of discounts.

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

**Figure 2 –** *Density function of the weekly prices and distribution of prices and sold quantities of "Powdered milk for babies Danone" sold in May 2014 in the province of Turin.*

| | |
|---|---|
| Coicop | 01.1.4.2.0 |
| Product | Powdered milk for babies Danone |
| Province | Turin |
| Year | 2014 |
| Month | 5 |

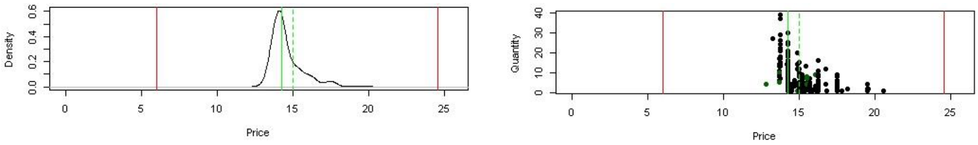


**Figure 3** – *Density function of the unit prices and distribution of prices and sold quantities of "Aged rum, Havana Club, 700 ml", sold in December 2014 in the province of Palermo.*

| | |
|---|---|
| Coicop | 02.1.1.1.2 |
| Product | Aged rum, Havana Club, 700 ml |
| Province | Palermo |
| Year | 2014 |
| Month | 12 |

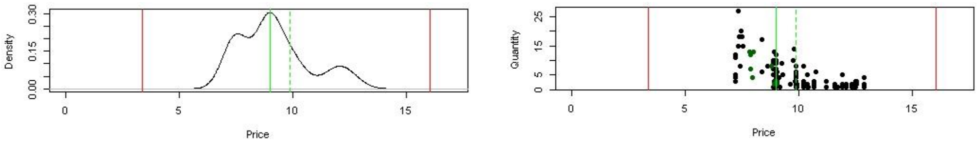


**Figure 4** – *Density function of the unit prices and distribution of prices and sold quantities of "Rice, PrivateLabel, 1000 gr", sold in December 2013 in the province of Ancona.*

| | |
|---|---|
| Coicop | 01.1.1.1.0 |
| Product | Rice, PL, 1000 gr |
| Province | Ancona |
| Year | 2013 |
| Month | 12 |

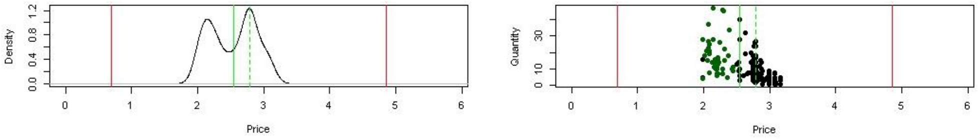


**Figure 5** – *Density function of the unit prices and distribution of prices and sold quantities of "Sparkling water, Rocchetta, 1.5 lt", in May 2013 in the province of Cagliari.*

| | |
|---|---|
| Coicop | 01.2.2.1.0 |
| Product | Sparkling water, Rocchetta, 1,5 lt |
| Province | Cagliari |
| Year | 2013 |
| Month | 5 |

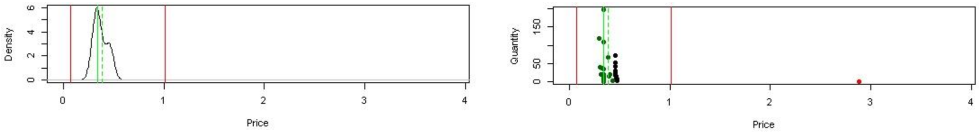
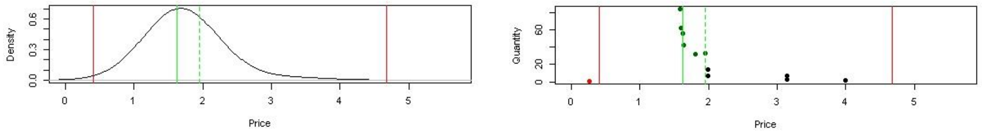


**Figure 6 –** *Density function of the unit prices and distribution of prices and sold quantities of "Yogurt Actimel, LC1", in June 2014 in the province of Piacenza.*

| | |
|---|---|
| Coicop | 01.1.4.3.0 |
| Product | Yogurt Actimel, LC1 |
| Province | Piacenza |
| Year | 2014 |
| Month | 6 |



## *4.2. Removal of inadmissible unit prices effect*

The moving trimming method has been applied on 2013 and 2014 data. The data consists in weekly prices and quantities of food and grocery market products sold in the hypermarket or supermarket of six modern retail distribution chains (Conad, Coop, Esselunga, Auchan, Carrefour, Selex) in five Italian provinces (Ancona, Cagliari, Palermo, Piacenza, Turin).

The results highlight the number of removed occurrences, the related sold quantities and the related turnover, and detects the consumption segments most affected by the deletion of occurrences.

**Table 1 –** *Total amount of occurrences, quantities sold and turnover – percentage of occurrences, quantities sold and turnover removed, by year and province.*

| Province | Year | Total amount of | | | Removal of inadmissible unit prices effect in terms of | | |
|---|---|---|---|---|---|---|---|
| | | Occurrences | Quantities sold | Turnover | Occurrences (%) | Quantities sold (%) | Turnover (%) |
| ANCONA | 2013 | 16'314'683 | 179'261'691 | 302'928'869.00 | 0.042 | 0.027 | 0.024 |
| | 2014 | 16'972'117 | 183'593'696 | 309'328'337.97 | 0.033 | 0.021 | 0.019 |
| CAGLIARI | 2013 | 11'236'941 | 145'630'320 | 256'859'395.95 | 0.044 | 0.027 | 0.028 |
| | 2014 | 11'383'752 | 145'612'723 | 253'942'082.32 | 0.044 | 0.032 | 0.030 |
| PALERMO | 2013 | 10'671'436 | 139'756'595 | 222'335'145.53 | 0.064 | 0.030 | 0.026 |
| | 2014 | 12'094'184 | 152'094'573 | 240'512'170.38 | 0.098 | 0.033 | 0.030 |
| PIACENZA | 2013 | 6'474'872 | 93'362'079 | 167'777'673.59 | 0.021 | 0.009 | 0.017 |
| | 2014 | 7'521'222 | 100'610'744 | 180'572'592.89 | 0.030 | 0.020 | 0.027 |
| TURIN | 2013 | 45'458'148 | 671'423'702 | 1'242'080'444.40 | 0.048 | 0.017 | 0.023 |
| | 2014 | 48'621'536 | 679'459'389 | 1'248'849'275.00 | 0.062 | 0.026 | 0.026 |

Table 1 presents the effect of removing inadmissible weekly prices on the total amount of occurrences, quantities sold and turnover.

Generally the number of removed occurrences is very low (Table 1) in all provinces but is not constant between the years under review. The exception is represented by the province of Cagliari for which the percentage of deleted occurrences is 0.04% in both years.

**Table 2 –** *Consumption segments (COICOP-6digit) with percentage of turnover removed greater than 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 |
| TURIN | 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 |

The turnover share lost due to the elimination of occurrences is very contained, in fact, is never more than 0.03%. Analyzing the problem by consumption segment, the situation changes. However, a very low proportion of consumption segments lost more than 0.05% of turnover. Therefore, for the majority of consumption segments the loss of turnover share is negligible.

In table 2, for each province and each year, the consumption segments that suffer of a loss of turnover share greater than 0.05%, due to the removal of inadmissible unit prices related to specific occurrences, are listed. It looks clear that consumption segments like *mineral water*, *perfumes* and *razors* are more problematic than other ones. In fact, in this segments, a greater number of occurrences with inadmissible unit prices can be identified.

## 4. The impact of data cleaning on micro-indices

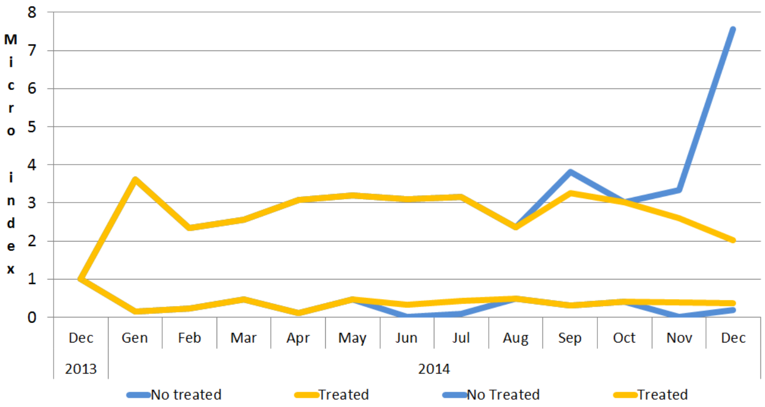
For an assessment of data cleaning, an analysis on micro-indices has been performed. In particular, micro-indices have been computed and their maximum and minimum values in each month for all the considered consumption segments, before and after the treatment, have been compared. The micro-indices are defined as the ratio between the price of the occurrence at time *t* and the price of the same occurrence at the base (usually December of the previous year). Their values represent the variation of price with respect to the price collected in the reference month. Micro-indices greater than 300% and lower than 20% are usually uncommon in food and grocery market.

In Figure 6 the maximum and minimum value of micro-indices, calculated before and after the phase of the data processing in the consumption segments *wine* and *pasta* for the province of Turin, have been compared.

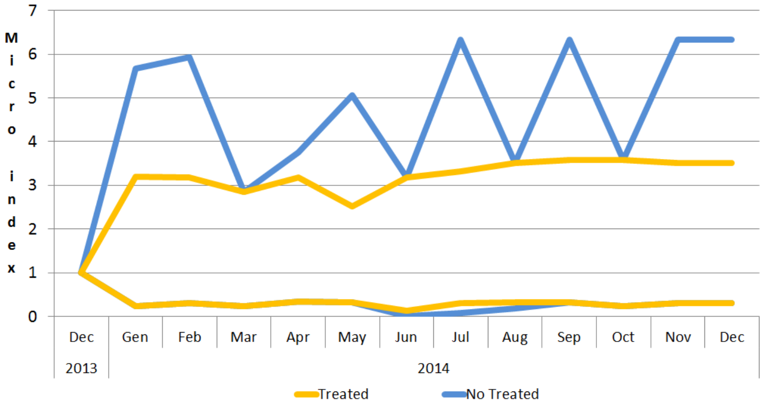
The impact of data cleaning on the micro-indices can be quite important, especially for the maximum. The changes for the minimum are less important. This is actually a symptom of the fact that further steps for improving the moving trimming have to focus in this direction. However, also in this version of the moving trimming, a great improvement in the quality and stability of the micro-indices, therefore in the quality and stability of the whole CPI, can be appreciate. In fact, after the application of moving trimming, the micro-indices come into a more stable and realistic range. For instance, not making the treatment, an increase of 7 times for the price of wine – that is clearly not realistic – would be introduced in the computation of the CPI in December 2014.

**Figure 7 –** *Maximum and minimum of micro-indices before and after the treatment with moving trimming method for a consumption segment.*

***TURIN 2014***
*consumption segment:* ***WINE***



***TURIN 2014***
*consumption segment:* ***PASTA***



The impact of moving trimming in pasta segment is more consistent during the year, while for wine segment is remarkable just since September 2014, and very important in December 2014. This could be due to the fact that wine segment have quite “seasonal” trend, because it increases the sales until December because of Christmas and New Year. In general, each consumption segment has its peculiarity and must be analysed carefully.

## 5. Conclusions

For a correct use of SD in the computation of an accurate price index over time, the correctness of the data becomes an important issue.

The data processing phase on SD is complex and articulated because of the considerable amount of data to be analyzed and the reduced timeframe during which the quality check process must be performed. For this reason, even methods well-known and well-established in statistical theory must be revised and adapted to this particular context.

In this work, just few aspects of the problem concerning data quality have been presented. A solution for identifying incorrect records and inadmissible unit prices of occurrences have been implemented.

The proposed method is named moving trimming. It is based on thresholds depending on the median of unit prices, taking into account the quantities, at which a product is sold in the weeks of the month in the whole province and two parameters that implicitly refer to the maximum discount and the maximum increase of price considered acceptable. The SD actually available are related to food a grocery retail sectors therefore, in this case, two plausible values for parameters are 5 (i.e. 1/5=0.2, that is 80% of discount) and 3 (300% of increase). However, increasing the median price of the product, the maximum discount and the maximum increase are reduced thanks to a logarithmic factor. Therefore, as the median price increases, the relative width for the interval of tolerance narrows down.

Some practical examples have been illustrated and synthetic results for the detection of inadmissible value are presented. Generally the number of removed occurrences is very low and, moreover, the turnover share lost due to the elimination of occurrences is very contained, is never more than 0.03%. The impact on micro-indices is remarkable and it can help to make the whole computation of CPI more stable and robust.

However, further and in depth studies are required in order to improve the moving trimming method and to identify ad hoc solutions when the density function of the prices assumes special shapes, in particular for products and consumption segment with high price elasticity of demand.

## References

BERNARDINI A., DE VITIIS C., GUANDALINI A., INGLESE F., TERRIBILI M.D. 2016. Measuring inflation through different sampling designs implemented on scanner data. *UNECE meeting of the group of experts, Geneve 2-4 May 2016..*

DE VITIIS C., CASCIANO M. C., GUANDALINI A., INGLESE F., SERI G., TERRIBILI M.D., TIERO F. 2015. Sampling design issues in the first Italian experience on scanner data, *Working group report on redesign CPS survey*.

ILO, IMF, OECD, Eurostat, United Nations, World Bank. 2004. Consumer Price Index Manual: Theory and Practice, Geneva: ILO Publications.

VAN DER GRIENT, H. A., DE HAAN J.. 2010. The Use of Supermarket Scanner Data in the Dutch CPI, *Joint ECE/ILO Workshop on Scanner Data, Eurostat*.

VERMEULEN B. C., HERREN H. M.,. 2006. Rents in Switzerland: sampling and quality adjustment, *11th Meeting, Ottawa Group, Neuchâtel 27-29 May*.

## SUMMARY

### The scanner data in the consumer price survey: the data processing

The Italian National Institute of Statistics (ISTAT) is planning a redesign of the Consumer Price Survey (CPS). The aim of the project is to modernise the survey. One of goal is to improving and unburdening the data collection phase.

Register of local units and availability of scanner data (SD) from retail modern distribution, provided through an agreement with Nielsen, are the starting point of this transformation.

The use of SD for this purpose implies the definition and the implementation of different checks to be made in both data acquisition and processing phases.

This work aims to describe the SD checks implemented in the experimental phase. The collection of high-frequency data, as are SD, requires the use of systematic and automatic checks, both formal and quality checks on the data flow. Formal checks have to ensure completeness of data collected at provincial level, distribution chains, outlets, products and weeks. Quality checks introduce editing rules which identify inadmissible values on the variables of interest (quantities sold, turnover and prices).

The moving trimming, a method for detecting inadmissible weekly prices of products, is described and discussed in the work. The main feature of the method is that the relative length of the tolerance interval shrinks as the median – taking into account the sold quantities – of weekly prices of the product in the month increases.

Antonella BERNARDINI, Istat, bernardini@istat.it
Alessio GUANDALINI, Istat, alessio.guandalini@istat.it
Francesca INGLESE, Istat, fringles@istat.it
Marco D. TERRIBILI, Istat, terribili@istat.it

# GLI ERRORI CAMPIONARI NELLE SERIE STORICHE DESTAGIONALIZZATE: ALCUNE ANALISI SUI DATI MENSILI DELLA RILEVAZIONE SULLE FORZE LAVORO

Claudio Ceccarelli, Cinzia Graziani, Silvia Loriga, Michele Antonio Salvatore, Andrea Spizzichino

## 1. Introduzione

L'analisi in serie storica di stime campionarie è divenuta pratica corrente per molti istituti nazionali di statistica, tra cui anche l'Istat. In particolare, l'analisi congiunturale (cioè di breve periodo) viene condotta sulle serie storiche destagionalizzate, ottenute identificando e poi eliminando dalle serie originarie la componente stagionale. Assume particolare rilievo pertanto la necessità di valutare l'accuratezza delle stime destagionalizzate che vengono diffuse, tenendo conto dell'errore campionario di cui sono affette le stime.

Nonostante la necessità di disporre della stima della varianza di una serie storica destagionalizzata proveniente da una serie di stime campionarie, né l'approccio *model-based seasonal adjustment* (implementato in TRAMO-SEATS) né tantomeno il metodo X-12 ARIMA[1] offrono attualmente la possibilità di tenere conto della eventuale natura campionaria dei dati e dell'errore che ne deriva (Bell, 2005). Prendendo spunto da tale problematica, il presente lavoro effettua un'analisi empirica sui dati mensili della Rilevazione sulle forze di lavoro al fine di valutare se gli errori campionari, stimati per i dati non destagionalizzati, possano essere applicati anche ai dati destagionalizzati per un'analisi della loro accuratezza.

A questo riguardo non esiste attualmente una pratica condivisa a livello internazionale: tra i pochi istituti nazionali di statistica che risulta abbiano affrontato tale problema, Statistics Canada afferma che sulla base di studi da loro effettuati, è risultato che l'errore campionario stimato per i dati non destagionalizzati possa essere applicato anche ai corrispondenti dati

[1] TRAMO-SEATS (*Time series Regression with Arima noise, Missing observations and Outliers - Signal Extraction in Arima Times Series*) e X-12 ARIMA sono i due metodi più diffusi per la destagionalizzazione delle serie storiche. TRAMO-SEATS è un metodo *model-based*: si ipotizza l'esistenza di un modello statistico parametrico e si assume che la serie storica osservata sia la parte finita della realizzazione di un processo stocastico, la cui struttura probabilistica è adeguatamente descritta dal suddetto modello. X-12 ARIMA invece segue un approccio non parametrico di tipo *filter-based*, si basa cioè sull'applicazione ripetuta di una serie di filtri lineari costituiti da medie mobili. In Istat si utilizza quasi esclusivamente il metodo TRAMO-SEATS.

destagionalizzati. Analogamente anche ABS (Australian bureau of statistics), che nel comunicato stampa mensile sulla Labour force survey applica ai dati destagionalizzati gli stessi intervalli di confidenza stimati per i dati non destagionalizzati.

L'analisi qui condotta, effettuata sulle serie degli errori campionari dei principali aggregati della partecipazione al mercato del lavoro, è finalizzata innanzitutto a verificare l'eventuale presenza di stagionalità nelle serie: in caso di assenza di stagionalità si dimostra come gli stessi errori campionari stimati per i dati non destagionalizzati possano essere ritenuti validi anche per i corrispondenti dati destagionalizzati; l'altro caso, ovvero quello in cui si verifichi la presenza di stagionalità nelle serie degli errori campionari, sarà invece oggetto di studi e sviluppi futuri.

## 2. Gli errori campionari

La Rilevazione sulle forze di lavoro è un'indagine campionaria, l'accuratezza delle stime prodotte deve quindi essere valutata tenendo conto dell'errore campionario derivante dall'aver osservato la variabile di interesse solo su una parte (campione) della popolazione.

Questo errore può essere espresso in termini di errore assoluto (standard error) o di errore relativo (cioè l'errore assoluto diviso per la stima, che prende il nome di coefficiente di variazione, CV). A partire da questi è possibile costruire l'intervallo di confidenza, all'interno del quale, con un prefissato livello di fiducia, è contenuto il valore vero, ma ignoto, del parametro oggetto di stima.

L'intervallo di confidenza è calcolato aggiungendo e sottraendo alla stima puntuale il suo errore campionario assoluto, moltiplicato per un coefficiente che dipende dal livello di fiducia; ad esempio, considerando il tradizionale livello di fiducia del 95%, il coefficiente corrispondente è pari a 1,96.

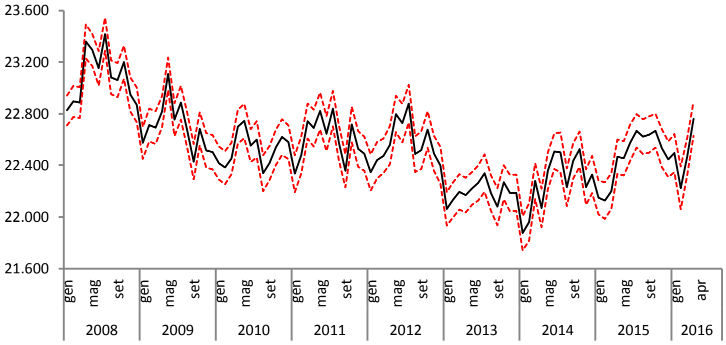
Da alcuni mesi nella nota metodologica in calce al comunicato stampa mensile 'Occupati e disoccupati'[2] l'Istat diffonde gli errori campionari (in particolare quelli relativi) delle stime delle principali variabili di interesse della partecipazione al mercato del lavoro. Da questi errori è possibile, come sopra descritto, costruire gli intervalli di confidenza riportati graficamente nelle figure seguenti (linee tratteggiate in rosso nelle figure1, 2 e 3).

La valutazione dell'errore campionario e dell'intervallo di confidenza è importante per una corretta interpretazione delle stime prodotte da un'indagine

[2] I comunicati mensili 'Occupati e disoccupati' sono consultabili nell'archivio Istat dei comunicati stampa all'indirizzo http://www.istat.it/it/archivio/occupati+e+disoccupati .

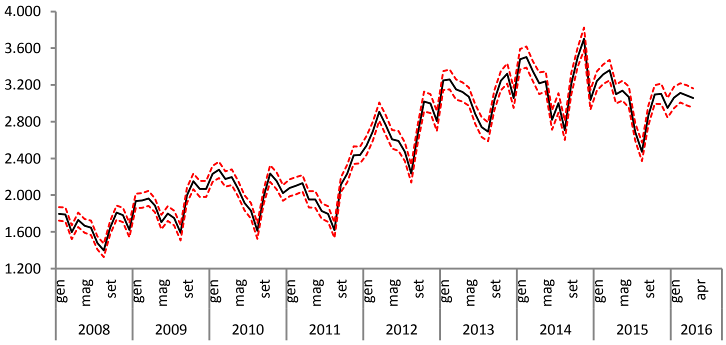
campionaria, ad esempio al fine di valutare la significatività della differenza tra due stime riferite a diversi domini territoriali o a diversi periodi di riferimento.

**Figura 1 –** *Occupati. Gennaio 2008 - aprile 2016, dati non destagionalizzati, valori assoluti in migliaia di unità*



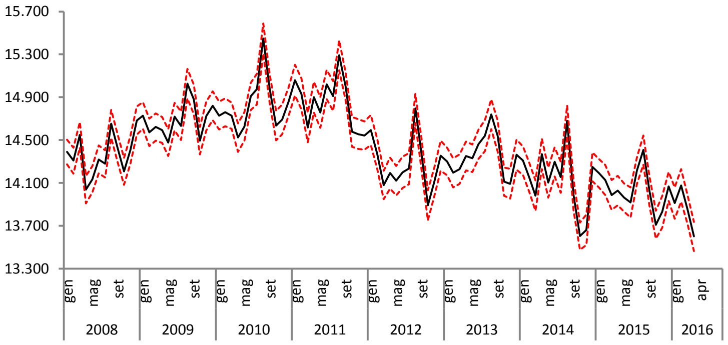
*Fonte: Istat. Rilevazione sulle forze di lavoro.*

**Figura 2 –** *Disoccupati. Gennaio 2008 - aprile 2016, dati non destagionalizzati, valori assoluti in migliaia di unità*



*Fonte: Istat. Rilevazione sulle forze di lavoro.*

**Figura 3 –** *Inattivi. Gennaio 2008 - aprile 2016, dati non destagionalizzati, valori assoluti in migliaia di unità*



*Fonte: Istat. Rilevazione sulle forze di lavoro.*

## 3. La destagionalizzazione di una serie storica di stime campionarie

Il focus principale del comunicato mensile “Occupati e disoccupati” è l’analisi congiunturale (cioè di breve periodo) della partecipazione al mercato del lavoro; di conseguenza, per permettere confronti tra un mese e il precedente, tutti i dati riportati nel comunicato stampa sono destagionalizzati.

La destagionalizzazione di serie storiche di dati a cadenza infra-annuale si basa sull’ipotesi che tali serie siano rappresentabili come combinazione di diverse componenti tra loro ortogonali: il ciclo-trend (CT), che rappresenta la tendenza di medio-lungo termine della serie storica, non influenzata da oscillazioni di brevissimo periodo; la componente stagionale (S) che si manifesta nel corso dell’anno in modo ricorrente e che descrive le fluttuazioni attribuibili ad esempio a fattori meteorologici, consuetudinari o legislativi; la componente irregolare (I) dovuta a fattori erratici. Nell’ipotesi di modello additivo, la serie di dati Y è rappresentata dalla (1):

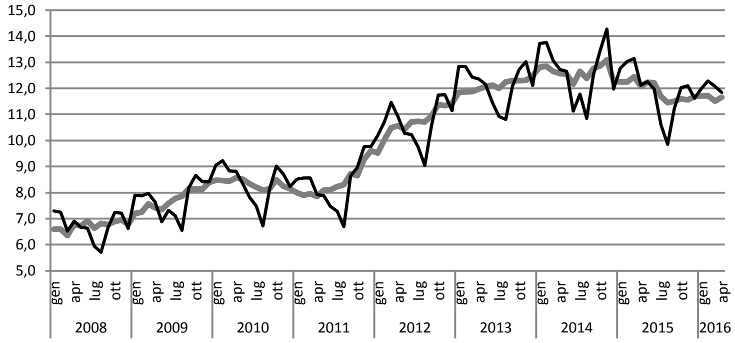
$$Y = CT + S + I \tag{1}$$

Una volta individuate le singole componenti mediante apposite tecniche statistiche (procedure di destagionalizzazione), la serie destagionalizzata è ottenuta sottraendo dalla serie storica originaria la componente stagionale (2).

$$Dest(Y) = Y - S = CT + I \tag{2}$$

Nella figura 4 è riportata a titolo di esempio la serie storica mensile non destagionalizzata (in nero) del tasso di disoccupazione e la corrispondente serie al netto della componente stagionale (in grigio).

**Figura 4** – *Tasso di disoccupazione. Gennaio 2008 - aprile 2016, dati non destagionalizzati e destagionalizzati, valori percentuali*



*Fonte: Istat. Rilevazione sulle forze di lavoro.*

La (1) riporta la scomposizione classica di una serie storica di dati non campionari. Nel caso in cui invece la serie storica derivi da stime prodotte da una indagine campionaria, si può assumere la presenza di un errore di campionamento (e) da considerare quando si stima con y il valore vero Y. Per comodità espositiva e senza perdere in generalità, si può assumere che le quantità sopra descritte siano legate dalla seguente relazione:

$$y = Y + e \tag{3}$$

Nel caso quindi di serie storiche di dati campionari, unendo la (1) con la (3), si ha:

$$y = CT + S + I + e \tag{4}$$

dove compaiono due componenti erratiche: I che rappresenta la componente irregolare del modello di destagionalizzazione ed e che rappresenta l’errore campionario commesso nello stimare Y con y.

Procedendo con la destagionalizzazione di y si avrà

$$Dest(y) = Dest(Y) + Dest(e) \tag{5}$$

La serie destagionalizzata ottenuta nella (5) è quindi uguale alla (2) a meno della destagionalizzazione della serie degli errori campionari.

Allo stato attuale i consueti metodi di destagionalizzazione (TRAMO-SEATS e X-12 Arima) non tengono conto della natura campionaria dei dati, di conseguenza la scomposizione di y non tiene conto di $e$ venendo così la (4) a coincidere con la (1).

Se ipotizziamo che la serie degli errori campionari non abbia un andamento stagionale ($Dest(e) = e$), considerando la (2) e la (5) si ha:

$$Dest(y) = Dest(Y) + e = CT + I + e \tag{6}$$

da cui si ricava che l'errore campionario delle stime non destagionalizzate può essere applicato anche ai corrispondenti dati destagionalizzati.

## 4. Analisi della stagionalità nella serie degli errori campionari

Adottando un approccio empirico, a partire dalla (5), si vuole verificare l'eventuale presenza di stagionalità nella serie degli errori campionari dei dati non destagionalizzati; come dimostrato nel paragrafo precedente, se gli errori non presentano un andamento stagionale, allora sono applicabili anche ai corrispondenti dati destagionalizzati.

Le serie storiche degli errori campionari assoluti dei principali aggregati della partecipazione al mercato del lavoro sono state pertanto sottoposte a specifici test per la verifica della presenza di stagionalità. In particolare sono stati utilizzati i test di stagionalità implementati nel *software* JDemetra+[3].

I test effettuati verificano se nelle serie sono presenti movimenti stagionali, prerequisito essenziale per procedere poi con la destagionalizzazione. La tabella 1 riporta una sintesi dei risultati dei principali test.

[3] Grudkowska (2015).

**Tabella 1 –** *Sintesi dei risultati dei principali test di stagionalità sulle serie degli errori assoluti di occupati, disoccupati e inattivi.*

| Test di stagionalità | Occupati | Disoccupati | Inattivi |
|---|---|---|---|
| 1. Autocorrelazioni dei ritardi stagionali | ? | ? | YES |
| 2.Test di Friedman (non parametrico) | ? | ? | ? |
| 3.Test di Kruskal-Wallis (non parametrico) | ? | ? | NO |
| 4.Picchi spettrali | NO | YES | NO |
| 5.Periodogramma | NO | NO | NO |
| 6.Test di regressione con dummies stagionali fisse | NO | YES | NO |
| 7.Test di stagionalità combinato | Assenza di stagionalità identificabile | Assenza di stagionalità identificabile | Assenza di stagionalità identificabile |

Il primo test implementato dal *software* è il test di Ljung-Box sulle autocorrelazioni dei ritardi stagionali: si verifica l'ipotesi nulla che a determinati *lag* temporali l'autocorrelazione non sia significativamente diversa da zero. In tal caso non c'è stagionalità al ritardo temporale considerato (nel caso in esame, trattandosi di serie mensili, si considera il 12° ritardo temporale ). Ciò tuttavia non esclude del tutto la presenza di stagionalità, in quanto potrebbe esserci autocorrelazione a *lag* temporali superiori a quello considerato. Questo è il caso delle serie degli errori assoluti di occupati e disoccupati (dove "?" indica che il risultato del test è incerto), mentre per gli inattivi il test evidenzia la presenza di autocorrelazione al 12° ritardo temporale.

Incertezza sulla presenza di stagionalità stabile emerge per le tre serie dai test non parametrici di Friedman[4] e di Kruskal-Wallis[5], i cui risultati sono riportati nel dettaglio nella tabella 2.

[4] Per maggiori dettagli sul test di Friedman si può consultare il sito http://ec.europa.eu/eurostat/sa-elearning/stable-seasonality-test.

[5] Per maggiori dettagli sul test di Kruskal-Wallis si può consultare il sito http://ec.europa.eu/eurostat/sa-elearning/kruskal-wallis-test.

**Tabella 2 –** *Test non parametrici di stagionalità sulle serie degli errori assoluti dei principali aggregati del mercato del lavoro.*

| Test di stagionalità | Occupati | Disoccupati | Inattivi |
|---|---|---|---|
| Test di Friedman | 24,6503 | 24,2168 | 20,7483 |
| *p-value** | *0,0103* | *0,0118* | *0,036* |
| Test di Kruskal-Wallis | 21,1169 | 20,4086 | 16,5111 |
| *p-value*** | *0,0322* | *0,040* | *0,1232* |

**p-value <0,01 si rifiuta l'ipotesi nulla di assenza di stagionalità stabile. Tuttavia se 0,01<=p-value<=0,05 il risultato del test è incerto.*
*** p-value >0,05 si rifiuta l'ipotesi nulla di presenza di stagionalità stabile. Tuttavia se 0,01<=p-value<=0,1 il risultato del test è incerto.*

Anche l'analisi dei periodogrammi non evidenzia presenza di stagionalità nelle serie considerate (test 4 e 5 in tabella 1). In particolare non risultano picchi riconducibili a movimenti stagionali né per la serie degli errori assoluti degli occupati né per quella degli inattivi (test 4); per i disoccupati invece, pur essendo stati individuati picchi associabili alla presenza di una componente stagionale prevedibile, questi non sono poi risultati statisticamente significativi (test 5).

Assenza di stagionalità deterministica negli errori assoluti emerge per occupati e inattivi, ma non per i disoccupati, applicando il test di regressione con variabili *dummies* stagionali fisse[6].

I risultati dei test fin qui esposti non permettono dunque di stabilire in maniera univoca se le serie degli errori campionari assoluti siano o meno affette da stagionalità.

Per ottenere una chiara evidenza è risultato utile considerare il test di stagionalità combinato[7], ritenuto il test più accurato per determinare se i movimenti identificati in una serie storica siano abbastanza stabili e regolari da essere classificabili come stagionali; questo test combina il test di Kruskal-Wallis con altri test per la verifica di stagionalità stabile.

Il test permette di verificare se la serie è caratterizzata da:

1. presenza di stagionalità identificabile;
2. probabile non presenza di stagionalità identificabile;
3. non presenza di stagionalità identificabile.

[6] L'andamento della trasformata logaritmica di ciascuna serie viene approssimato con un modello di regressione che utilizza variabili *dummies*, 12 nel caso mensile. Se queste variabili considerate congiuntamente risultano statisticamente significative si deduce che c'è stagionalità deterministica.

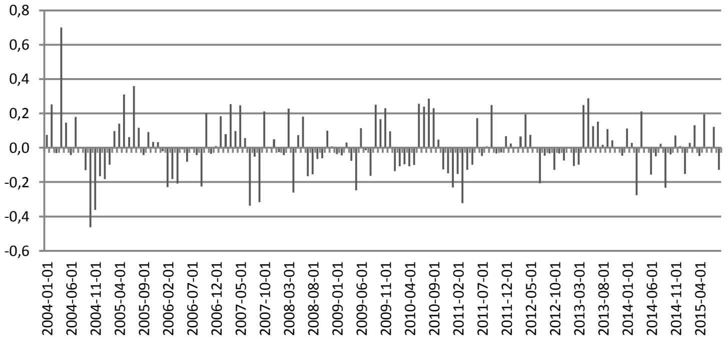
[7] Dagum E, et al. (1999). Per maggiori dettagli si può inoltre consultare il sito http://ec.europa.eu/eurostat/sa-elearning/combined-seasonality-test.

In particolare si raccomanda di non procedere a destagionalizzazione nel caso il test verifichi la non presenza di stagionalità identificabile[8]. Come si osserva dai risultati riportati in tabella 1, per tutte le serie oggetto di analisi non è stata rilevata stagionalità identificabile.

Per avere una ulteriore conferma dell'assenza di stagionalità, si è deciso, nonostante i risultati dei test, di procedere comunque alla destagionalizzazione delle serie. In tal modo è stato possibile isolare le diverse componenti (il ciclo-trend, la componente stagionale e la componente irregolare) e osservare, in particolare, l'entità della componente stagionale. Dai grafici che seguono appare evidente la residualità della componente stagionale per le serie degli errori campionari assoluti.

In particolare per gli occupati, il cui errore campionario assoluto ammonta in media a quasi 70 mila unità, la componente stagionale ha un campo di variazione pari a circa 1.200 unità (figura 5).

**Figura 5 –** *Occupati. Componente stagionale dell'errore campionario assoluto - Gennaio 2004 - agosto 2015, valori in migliaia di unità.*



La residualità della componente stagionale è evidente anche per le serie mensili degli errori campionari assoluti dei disoccupati (figura 6) e degli inattivi (figura 7).

Per i disoccupati il range di variazione è di circa 6.800 unità a fronte di un errore assoluto pari in media a oltre 45 mila unità; per gli inattivi invece il campo di variazione è di circa 1.500 unità mentre l'errore assoluto medio è di quasi 70 mila unità.

[8] ONS (2007), Guide to seasonal adjustment with the X-12-ARIMA.

**Figura 6 –** *Disoccupati. Componente stagionale dell'errore campionario assoluto - Gennaio 2004 - agosto 2015, valori in migliaia di unità.*

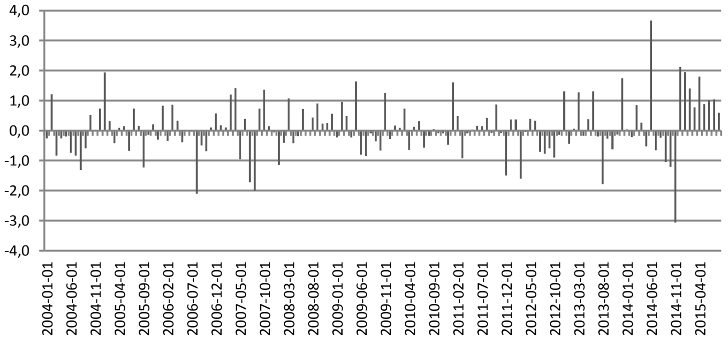
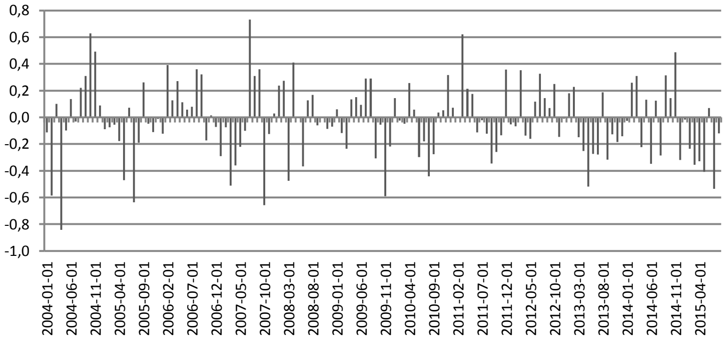


**Figura 7 –** *Inattivi. Componente stagionale dell'errore campionario assoluto - Gennaio 2004 - agosto 2015, valori in migliaia di unità.*



## 5. Stato dell'arte e sviluppi futuri

Dalle analisi effettuate emerge l'assenza di stagionalità nelle serie degli errori assoluti dei principali indicatori della partecipazione al mercato del lavoro.

Per tali indicatori, è pertanto possibile applicare ai dati destagionalizzati l'errore relativo stimato per i corrispondenti dati non destagionalizzati, al fine di fornire una misura dell'accuratezza delle stime diffuse, che tenga conto dell'errore campionario.

Tuttavia, nel caso in cui si riscontri stagionalità nelle serie degli errori assoluti, è necessario procedere con ulteriori studi al fine di formalizzare la misura dell'errore.

Gli studi e gli sviluppi futuri saranno pertanto volti a generalizzare i risultati qui ottenuti anche nel caso di presenza di stagionalità nelle serie degli errori assoluti, oltre che a esplorare la possibilità di stimare una misura dell'errore che tenga conto congiuntamente delle due fonti di errore di cui sono affetti i dati destagionalizzati diffusi:

- errore campionario
- errore derivante dalla stima del modello di destagionalizzazione.

## Riferimenti bibliografici

ABS. 2016. Labour Force Australia, June 2016 n. 6202.0, http://www.ausstats.abs.gov.au/ausstats/meisubs.nsf/0/702ADCA7FC70AE7BCA257FEF007D96B4/$File/62020_jun%202016.pdf.

BELL W.R.. 2005. Some Consideration of Seasonal Adjustment Variances, *ASA proceeding 2005*, http://www.census.gov/ts/papers/jsm2005wrb.pdf.

DAGUM E, et al.1999. X11ARIMA version 2000: Foundations and User's Manual.

EUROSTAT. Seasonality tests, *Take eLearning course on Seasonal Adjustment*, http://ec.europa.eu/eurostat/sa-elearning/seasonality-tests.

EUROSTAT. 2009. ESS Guidelines on Seasonal Adjustment, http://ec.europa.eu/eurostat/documents/3859598/5910549/KS-RA-09-006-EN.PDF.

GRUDKOWSKA S.. 2015. *JDemetra+ User Guide*. Narodowy Bank Polski, Department of Statistics.

ONS. 2007. Guide to seasonal adjustment with the X-12-ARIMA.

STATISTICS CANADA. 2016. Guide to the Labour Force Survey, Catalogue n. 71-543-G, http://www.statcan.gc.ca/pub/71-543-g/71-543-g2016001-eng.pdf.

## SUMMARY

## Sampling errors in seasonally adjusted time series: some analysis on Labour Force Survey monthly data

The time series analysis, even of indicators which are sampling estimates, has become common practice for many national statistical offices, including Istat. In particular, the short-term analysis is conducted on the seasonally adjusted time series, obtained by identifying and then eliminating the seasonal component from the original series. It is particularly relevant, therefore, the need to evaluate the accuracy of seasonally adjusted estimates that are disseminated, taking into account sampling errors which affect the estimates.

Despite the need of estimating the variance of a seasonally adjusted time series composed by sampling estimates, nor the model-based approach to seasonal adjustment (implemented in TRAMO-SEATS) nor the X-12 ARIMA method currently offer the possibility to take account of the possible sampling nature of the data and of their error (Bell, 2005). Taking a cue from this issue, this paper makes an empirical analysis on the monthly data of the Labour Force Survey in order to assess whether the sampling errors, estimated for not seasonally adjusted data, can also be applied to seasonally adjusted data to analyze their accuracy.

In this respect currently there are no common practices at the international level: among the few national statistical offices that have addressed this issue, Statistics Canada stated that they found that the estimated sampling errors for not seasonally adjusted data can also be applied to the corresponding seasonally adjusted. Similarly, ABS (Australian Bureau of statistics), that in the monthly press release on the Labour Force Survey applies to seasonally adjusted data the same confidence intervals estimated for the not seasonally adjusted.

The analysis conducted here, on the series of sampling errors of the main indicators of the participation in the labour market, is primarily aimed to check for the presence of seasonality in the series: in case of absence of seasonality it is demonstrated how the sampling errors estimated for not seasonally adjusted data may be applied as well to the corresponding seasonally adjusted; the opposite case, namely the presence of seasonality in the series of sampling errors, will be the subject of further studies and developments.

Claudio CECCARELLI, Istat, clceccar@istat.it
Cinzia GRAZIANI, Istat, cingraziani@istat.it
Silvia LORIGA, Istat, siloriga@istat.it
Michele Antonio SALVATORE, Istat, salvatore@istat.it
Andrea SPIZZICHINO, Istat, spizzich@istat.it

# BONFERRONI INDEX DECOMPOSITION AND THE SHAPLEY METHOD

Giovanni Maria Giorgi, Alessio Guandalini

## Introduction

In the last years the Bonferroni inequality index ($B$, Bonferroni, 1930) has been analysed with more attention for its particular characteristics. The Bonferroni, as the Gini (1914) index, has been identified as special cases of a general formula by De Vergottini (1956). Further studies on $B$ have been conducted by Piesch (1975) and Nygård and Sandström (1981). Recently, new and interesting interpretations and extensions of $B$ have been proposed.

A widespread topic in the literature on inequality measures is their decomposition. Many contributions are related to the Gini index $R$ (Gini, 1914). Tarsitano (1990) showed various standard results used to investigate the decomposition of $B$ and Bárcena-Martin and Silber (2013) derived an algorithm that greatly simplifies it.

In the field of inequality index decomposition two main lines of research can be distinguished: decomposition by income sources and by population subgroups. The former is widely treated, whilst less attention has been paid on the latter (Giorgi, 2011). The reason lies in the difficult to decompose additively some inequality indices, such as $R$ and $B$, by population subgroups. To overcome this drawback Deutsch and Silber (2007) used the so-called Shapley method on $R$.

In the present paper the same method has been applied on $B$. Several interesting similarities and differences among the two indices are highlighted. Furthermore, some properties of $B$ have been object of deeper investigation.

The paper is organized as follows: in section 1 the original expression and the main properties of $B$ are presented. In section 2, the Shapley method is quickly surveyed and a numerical illustration is provided. In section 4 the method is applied to real data (It-SILC data referred to 2009). Finally, conclusions and future prospects of research are discussed in section 5.

## 1. The Bonferroni inequality index

The expression of $B$ proposed by Bonferroni (1930, p. 55 and p. 85) is a function of partial means:

$$B = \frac{1}{N-1}\sum_{i=1}^{N-1}\frac{(\mu - \mu_i)}{\mu} = 1 - \frac{1}{N-1}\sum_{i=1}^{N-1}\frac{\mu_i}{\mu},$$

where $0 \leq B \leq 1$, and

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i \qquad \mu_i = \frac{1}{N}\sum_{j=1}^{i} x_j \qquad i = 1, 2, \ldots, N$$

are the general and the partial means for units sorted in non-decreasing order with respect to the $X$ variable.

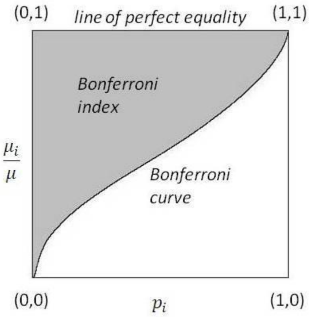
The $B$ index satisfies the axiomatic properties required for inequality indices [i.e. the principle of transfer, of proportional addition to incomes, of proportional addition to persons, of symmetry, of normalization and of operationality] (Giorgi, 1998, p. 142) and $B \geq R$ holds, because it gives bigger weights to units with lower values in the $X$ ranking (see, e.g., De Vergottini, 1950 pp. 318-319 and Pizzetti, 1951 p. 302). Therefore, $B$ is more sensitive to lower levels of the income distribution (see, e.g., Giorgi and Mondani, 1995).

The Bonferroni index is linked to the Bonferroni curve (Figure 1) which is obtained by plotting the cumulative proportion of recipients ($p_i = i/N$), arranged in non-decreasing values of $X$, versus the corresponding ratio between partial mean and total mean ($\mu_i/\mu$).

The polygonal line joining the points ($p_i$, $\mu_i/\mu$) is the Bonferroni curve. If all the recipients in the population have the same quantity of $X$ (i.e equal to $\mu$) the Bonferroni curve coincides with the line of perfect equality that joins the coordinate points (0,0), (0,1), (1,1).

The area between the Bonferroni curve and the line of perfect equality is the concentration area, which is equal to the value of $B$ (Giorgi and Crescenzi, 2001, p. 572-573).

**Figure 1 –** *An example of Bonferroni curve.*



## 2. The Shapley decomposition

To overcome the problem of additive decomposition of $R$ by population subgroups, Deutsch and Silber in 2007 used the Shapley decomposition, first introduced in this field by Shorrocks (1999). They derived the impact of four components: inequality within population subgroups ($w$), inequality between population subgroups ($b$), ranking ($r$) and the relative size in each population subgroup ($n$).

Shapley decomposition is based on the well-known concept of Shapley value in cooperative game theory (Shapley, 1953). The idea of Shapley value is to remove from time to time the contribution of all possible combinations of considered factors for determining their marginal contribution. Therefore, when the method is applied to inequality indices, considering factors in symmetrical manner, it allows to derive the expected marginal contribution to inequality of each factor. Moreover, the contributions sum to the exactly amount of inequality index considered (Shorrocks 1999 and 2013).

For comparing the results obtained by Deutsch and Silber (2007) on $R$, the same factors have been considered for decomposing $B$ (i.e. $w$, $b$, $r$, $n$).

Let assume to have a population $P$ partitioned in $k$ population subgroups $P_j$ ($j = 1, \dots, k$) where $y_{ji}$ is the income of recipient $i$ ($i = 1, \dots, n_j$) in the population subgroup $j$.

For removing $w$ is sufficient to replace $y_{ji}$ with $\mu_j$, that is the average income of the population subgroup $j$ to which the recipient $i$ belongs ($y_{ij} \rightarrow \mu_j$). While, for removing $b$, a kind of standardization is done and $y_{ij}$ must be replaced with $y_{ij}\ \frac{\mu_j}{\mu}$ ($y_{ij} \rightarrow y_{ij}\ \frac{\mu_j}{\mu}$). For closing the effect of difference in size ($n$) of population subgroups off, the population subgroups must be brought to have the same sizes. Therefore, the least common multiple of size of each population subgroup is done and the $y_{ij}$ are repeated by the number that leads equality in size between the population subgroups. When applying on $B$, an objection usually raised is that $B$, on the contrary of $R$, does not satisfy the Dalton principle of replication invariant. However, with a simulation study – not included here for the sake of brevity – has been verified that, when the size increases (since 100 units), the effect of replications becomes negligible in $B$. Finally, for removing the effect of ranking ($r$) it is sufficient to sort, firstly, the population subgroups by their average income, $\mu_j$, and then the recipients by their income within each population subgroup. Obviously for removing the effect of two or more factors at the same time, the methods just illustrated must be applied together.

The marginal impact ($SV$) of each factor is derived computing the following weighted means of the indices ($I$=$R$,$B$) derived when, from time to time, the effect of components is removed:

$$SV_w = \frac{1}{4}(I - I_w) + \frac{1}{12}[(I_b - I_{wb}) + (I_n - I_{wn}) + (I_r - I_{wr})] + \frac{1}{12}[(I_{bn} - I_{wbn}) + (I_{br} - I_{wbr}) + (I_{rn} - I_{wrn})] + \frac{1}{4}(I_{bnr} - I_{wbnr}) \tag{1}$$

$$SV_b = \frac{1}{4}(I - I_b) + \frac{1}{12}[(I_w - I_{wb}) + (I_n - I_{bn}) + (I_r - I_{br})] + \frac{1}{12}[(I_{wn} - I_{wbn}) + (I_{wr} - I_{wbr}) + (I_{rn} - I_{bnr})] + \frac{1}{4}(I_{wnr} - I_{wbnr}) \tag{2}$$

$$SV_n = \frac{1}{4}(I - I_n) + \frac{1}{12}[(I_w - I_{wn}) + (I_b - I_{bn}) + (I_r - I_{nr})] + \frac{1}{12}[(I_{wb} - I_{wbn}) + (I_{wr} - I_{wnr}) + (I_{br} - I_{bnr})] + \frac{1}{4}(I_{wbr} - I_{wbnr}) \tag{3}$$

$$SV_r = \frac{1}{4}(I - I_r) + \frac{1}{12}[(I_w - I_{wr}) + (I_b - I_{br}) + (I_n - I_{nr})] + \frac{1}{12}[(I_{wb} - I_{wbr}) + (I_{wn} - I_{wnr}) + (I_{bn} - I_{bnr})] + \frac{1}{4}(I_{wbn} - I_{wbnr}) \tag{4}$$

In the expressions (1)-(4) the subscript of $I$ denotes which factor has been removed (for instance $I_w$ is the index computed when the component of within inequality, $w$, has been removed).

### *2.1. Numerical illustration*

To better explain how the Shapley decomposition works, the example in Deutsch and Silber (2007) is recovered and all the computational steps are exhaustively explained. Let us consider a population with 5 recipients with related income 2, 4, 14, 30 and 50. Assume that individuals with income 2, 14 and 50 belong to population subgroup A and those with income 4 and 30 belong to population subgroup B.

Table 1 shows all the scenarios when removing factors separately, in pairs, in tern and all together. Furthermore, the related income distribution and the values of $R$ and $B$ are presented. Then, the marginal contributions for each factor ($SV$) have been derived with the expressions (1)-(4) and the values are reported in Table 2.

Because this is an illustrative example on the application of the Shapley decomposition, here the replication invariance principle property for $B$ is overlooked. Anyway, some important preliminary results can be stressed. In particular, when removing $w$, $B$ is negative (Table 1, case 2 and 7). It can occurs when there is negative correlation between mean income and mean rank (Frick and Goebel, 2008, p. 559). In fact, in the extreme case, when arranging the distribution of income in decreasing order, Rao (1969, p. 245) shows that $R$ is equal to $-R$, and the same occurs for $B$.

**Table 1 –** *Gini ($R$) and Bonferroni ($B$) indices in different scenarios in which the factors have been removed. Illustrative example related to the income of 5 recipients belonging to two different population subgroups: A={2, 14, 50} and B={4, 30}.*

| Removed factor | | Income distribution | $R$ | $B$ |
|---|---|---|---:|---:|
| 1 | $-$ | 2 4 14 30 50 | 0.488 | 0.698 |
| 2 | $w$ | 22 17 22 17 22 | 0.000 | -0.017 |
| 3 | $b$ | 1.82 4.71 12.73 35.29 45.45 | 0.471 | 0.686 |
| 4 | $n$ | 2 2 4 4 4 14 14 30 30 30 50 50 | 0.481 | 0.650 |
| 5 | $r$ | 4 30 2 14 50 | 0.304 | 0.431 |
| 6 | $wb$ | 20 20 20 20 20 | 0.000 | 0.000 |
| 7 | $wn$ | 22 22 17 17 17 22 22 17 17 17 22 22 | 0.000 | -0.022 |
| 8 | $wr$ | 17 17 22 22 22 | 0.060 | 0.098 |
| 9 | $bn$ | 1.82 1.82 4.71 4.71 4.71 12.73 12.73 35.29 35.29 35.29 45.45 45.45 | 0.462 | 0.638 |
| 10 | $br$ | 4.71 35.29 1.82 12.73 45.45 | 0.236 | 0.346 |
| 11 | $rn$ | 4 4 4 30 30 30 2 2 14 14 50 50 | 0.284 | 0.415 |
| 12 | $wbn$ | 19.5 19.5 19.5 19.5 19.5 19.5 19.5 19.5 19.5 19.5 19.5 19.5 | 0.000 | 0.000 |
| 13 | $wbr$ | 20 20 20 20 20 | 0.000 | 0.000 |
| 14 | $wnr$ | 17 17 17 17 17 17 22 22 22 22 22 22 | 0.064 | 0.091 |
| 15 | $bnr$ | 4.71 4.71 4.71 35.29 35.29 1.82 1.82 12.73 12.73 12.73 45.45 45.45 45.45 | 0.217 | 0.345 |
| 16 | $wbnr$ | 20 20 20 20 20 20 20 20 20 20 20 20 | 0.000 | 0.000 |

Note: $w$=inequality within, $b$=inequality between, $n$=size, $r$=ranking.

The results in Table 2 provide an initial idea on the hierarchy and the magnitude of the marginal contribution of each factor in determining $R$ and $B$. Without pursuing further the results of the example, it is just important to point out that the hierarchy of the factors is the same for both the indices and their magnitude change slightly.

**Table 2** – *Marginal impact of each component on Gini ($R$) and Bonferroni ($B$) indices. Illustrative example related to the income of 5 recipients belonging to two different population subgroup: A={2, 14, 50} and B={4, 30}.*

| Factor | Contribution on $R$ | | Contribution on $B$ | |
|---|---|---|---|---|
| | $SV$ | % | $SV$ | % |
| $w$ | 0.353 | 72.34 | 0.515 | 73.77 |
| $b$ | 0.038 | 7.78 | 0.045 | 6.39 |
| $n$ | 0.005 | 1.02 | 0.018 | 2.64 |
| $r$ | 0.092 | 18.86 | 0.120 | 17.20 |
| $I$ | 0.488 | 100.00 | 0.698 | 100.00 |

Note: $w$=inequality within, $b$=inequality between, $n$=size, $r$=ranking.

## 3. Application on real data

The Shapley decomposition of Gini ratio index ($R$) and Bonferroni index ($B$) have been applied on the data collected by Italian component of European Survey on Income and Living Condition of 2009. The Eu-SILC is a yearly survey carried out in all European countries and defined within the European Regulation no. 1177/2003. Its main aim is to provide data on income, poverty and social exclusion, both cross-sectional and longitudinal. The Italian sample of 2009 survey is 20,928 household and 52,433 individuals. We consider the whole Italian population divided into three population subgroups, that are the main geographical areas: North, Center and South. In table 3, some descriptive statistics on household income distribution for the whole population and for the population subgroups are presented.

The inequality measures have been computed with respect to the household incomes. The incomes have not been equivalised to take into account the different size of the households. The values of $R$ have been estimated through the expression of the sampling estimator defined by Eurostat (2004, p. 39), whilst $B$ through the expression of the sampling estimator derived in Giorgi and Guandalini (2013, p. 154).

Looking at Table 3, North and Center have a quite similar situation. Whilst in South there are lower incomes and higher inequality. Through Shapley decomposition the impact of within inequality ($w$), between inequality ($b$), and ranking ($r$), different size of subgroups ($n$), both on $R$ and $B$, has been derived. Then the contribution for each component on these two inequality measures have been compared. The sample size in the three subpopulations considered are larger than 4,000 sampling units, therefore the Dalton principle of replication invariance can be considered satisfied by $B$, too.

**Table 3** – *Some descriptive statistics on average Italian household income distribution by three population subgroups (North, Center and South). Eu-SILC, Italy 2009.*

| Total Geografical Area | Percentage of households | Q1 | Median | Q3 | Mean | $R$ | $B$ |
|---|---|---|---|---|---|---|---|
| North | 48.30 | 16,603 | 26,631 | 40,844 | 31,872 | 0.360 | 0.481 |
| Center | 19.74 | 16,566 | 26,330 | 41,066 | 31,609 | 0.358 | 0.481 |
| South | 31.96 | 12,819 | 20,285 | 31,272 | 24,434 | 0.367 | 0.488 |
| Italy | 100.00 | 15,148 | 24,118 | 38,233 | 29,442 | 0.367 | 0.487 |

Note: Sample size=20,928; Total number of the household=24,641,200.

In Figure 2 the Lorenz curve and the Bonferroni curve for each scenario have been reported. They show what happen to the income distribution when the components considered are removed separately, in pairs, in tern and all together. In this way it is easier to understand in which way the factors contribute to the inequality.

**Table 4** – *Marginal impact of each component on Gini (R) and Bonferroni (B) indices. Confidence interval at 95% in squared brackets. Application on average Italian household income distribution by three population subgroups (North, Center and South). Eu-SILC, Italy 2009.*

| Factor | Contribution on $R$ | | Contribution on $B$ | |
|---|---|---|---|---|
| | $SV$ | % | $SV$ | % |
| $w$ | 0.24111<br>[0.24100, 0.24122] | 65.70<br>[65.69, 65.71] | 0.33433<br>[0.33414, 0.33452] | 68.74<br>[68.71, 68.77] |
| $b$ | 0.03553<br>[0.03540, 0.03567] | 9.68<br>[9.65, 9.72] | 0.04603<br>[0.04585, 0.04622] | 9.46<br>[9.43, 9.50] |
| $n$ | 0.00131<br>[0.00128, 0.00134] | 0.36<br>[0.35, 0.36] | 0.00544<br>[0.00534, 0.00554] | 1.12<br>[1.10, 1.14] |
| $r$ | 0.08904<br>[0.08894, 0.08913] | 24.26<br>[24.24, 24.29] | 0.10057<br>[0.10041, 0.10073] | 20.68<br>[20.65, 20.71] |
| $I$ | 0.36699<br>[0.36682, 0.36716] | 100.00 | 0.48637<br>[0.48618, 0.48657] | 100.00 |

Note: $w$=inequality within, $b$=inequality between, $n$=size, $r$=ranking.

The index $R$ is equal to 0.367 and $B$ is equal to 0.502 (Figure 2: case 1). When removing the component $w$ both the indices are close to 0. Instead, when closing the $b$ component off, they decrease slightly. Removing $n$ has an unimportant effect on indices. Finally, when removing $r$ component, we have a particular trend of both curves, because in these cases we order first the population subgroups and then the units within the population subgroup (Figure 2: cases 5, 10, 11 and 15). In table 4 the absolute and relative impact of each of four components considered is shown.

## 4. Conclusion and further research

An important topic on inequality measures is their decomposition. Two main lines of research can be distinguished: decomposition by income sources and by population subgroups. In the literature less attention has been paid on decomposition by population subgroups because the inequality indices are not additively decomposable. To overcome this drawback Deutsch and Silber (2007) used the so-called Shapley method to decompose $R$. In the present paper the same method has been applied to decompose $B$.

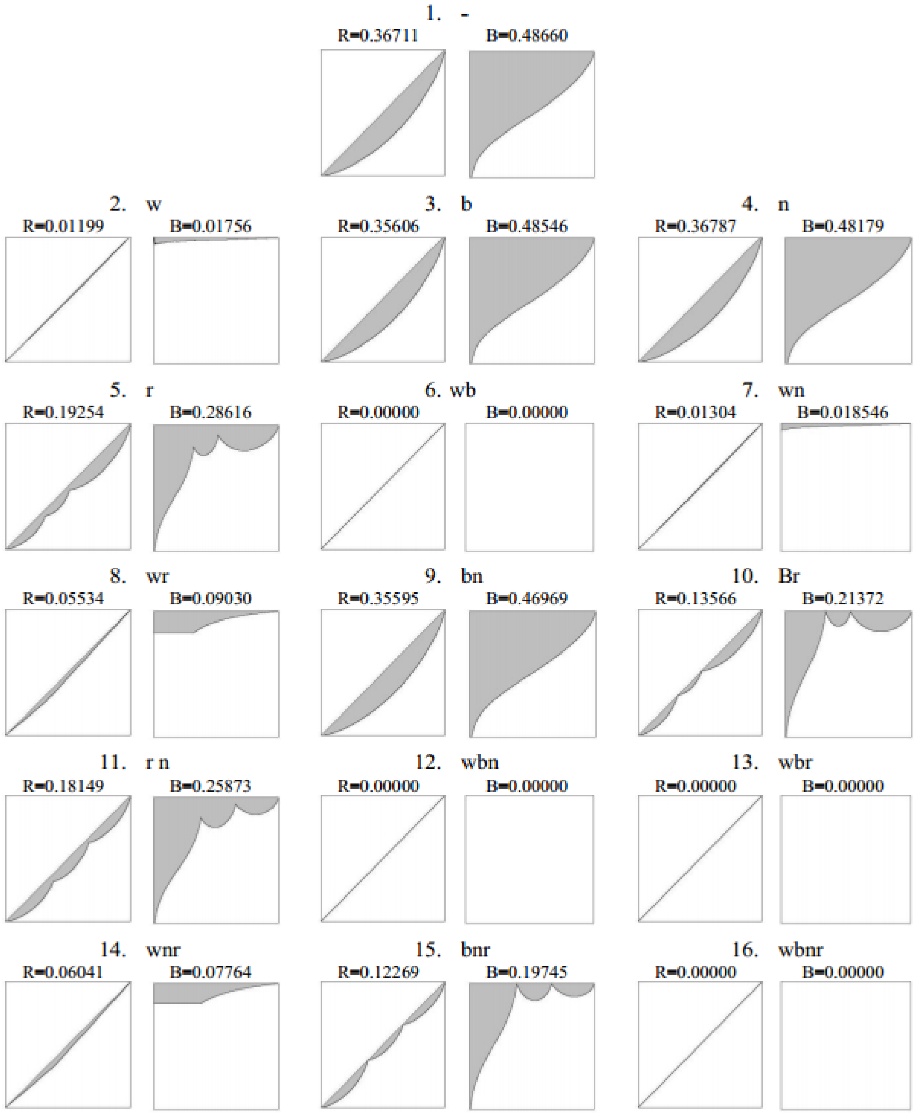
The Shapley decomposition has been useful also to highlight the difference between $R$ and $B$. The empirical illustration shows the decomposition of both indices in Italy in 2009 when the whole Italian population is divided considering the three main geographical areas: North, Center and South. Four components are considered in the decomposition: inequality within groups ($w$), inequality between groups ($b$), differences in size ($s$) and ranking ($r$). For both indices, most of the total inequality is due to $w$ followed by $r$, $b$ and $n$.

The relative contribution of between groups inequality is similar for both the indices. The within groups inequality has a higher contribution in determining $R$, whilst ranking and differences in size have a higher contribution in determining $B$. The study case shows that, besides the difference between $B$ and $R$ in assigning weights to the units and the consequent greater sensitivity for lower levels of income by $B$, other interesting features make the index different. In fact, the features of each group, such as homogeneity within - denoted by the component of inequality within ($w$) - and the size of the groups (component $n$), have higher influence on $B$ than on $R$. The hierarchy and the magnitude of these components in determining the inequality appear to be confirmed both in the application on real data and in the numerical illustration. However, a deeper investigation on the range of variation of the components under different income distributions is very interesting and it will be the object of further studies.

## Acknowledgment

The present work has been realized within the grant for the project “Indici di disuguaglianza e variabilità: nuove prospettive di ricerca” (Sapienza 2013).

**Figure 2 –** *Gini curve and Bonferroni curve when removing the component within (w), between (b), size (n) and ranking (r) separately, in pairs, in tern and all together.*



## References

BÁRCENA-MARTIN E., SILBER J. 2013. On the generalization and decomposition of the Bonferroni index. *Social Choice and Welfare*, Vol. 41, No. 4, pp. 763-787.

BONFERRONI C.E. 1930. *Elementi di statistica generale*. Firenze: Libreria Seber.

DEUTSCH J., SILBER J. 2007. Decomposing income inequality by population subgroups: a generalization. In J.A. Bishop and Y. Amiel (eds.), *Research on Economic Inequality: Inequality and Poverty*. Berlin: Springer, Vol. 14, pp. 237-253.

DE VERGOTTINI M. 1950. Sugli indici di concentrazione. *Statistica*, Vol. 10, pp. 445-454.

EUROSTAT 2004. Common cross-sectional EU indicators based on EU-SILC; the gender gap. *Working Group on Statistics on Income and Living Conditions (EU-SILC)*, 29-30 March, Luxemburg, pp. 1-42.

FRICK J.R., GOEBEL J. 2008. Regional Income Stratification in Unified Germany Using a Gini Decomposition Approach. *Regional Studies*, Vol. 42, No. 4, pp. 555-577.

GINI C. 1914. Sulla misura della concentrazione e della variabilità dei caratteri, *Atti del Reale Istituto Veneto di Scienze, Lettere ed Arti*, Vol. 73, pp. 1203-1248. (English translation in *Metron*, 2005, Vol. 63, pp. 3-38).

GIORGI G.M. 1998. Concentration index, Bonferroni. In Kotz S. *et. al* (Eds.) *Encyclopedia of Statistical Sciences*, Update 2. New York: Wiley-Intersciences, pp. 141-146.

GIORGI G.M. 2011. The Gini inequality index decomposition, an evolutionary study. In Deutsch J., Silber J. (Eds.) *The measurement of individual well-being and group inequality: essay in memory of Z.M. Berrebi*. London: Routledge, pp. 185-218.

GIORGI G.M., CRESCENZI M. 2001. A look at the Bonferroni inequality measure in a reliability framework. *Statistica*, Vol. 61, No. 4, pp. 571 -583.

GIORGI G.M., GUANDALINI A. 2013. A sampling estimator of the Bonferroni inequality index. *Rivista Italiana di Economia, Demografia e Statistica*, Vol. LXVII, No. 3/4, pp. 151-158.

GIORGI G.M., MONDANI R. 1995. Sampling distribution of Bonferroni inequality index from an exponential population. *Sankhya*, Series B, Vol. 57, No. 1, pp. 10-18.

NYGÅRD F., SANSTRÖM A. 1981. *Measuring income inequality*. Stockholm: Almqvist & Wiksell International.

PIESCH W. 1975. *Statistische Konzentrationsmasse*. Tübingen: J.B.C. Mohr (Paul Siebeck).

PIZZETTI E. 1951. Relazioni fra indici di concentrazione. *Statistica*, Anno XI, No. 3-4, pp. 294-316.

RAO V. 1969. Two decompositions of concentration ratio. *Journal of the Royal Statistical Society*, Vol 132A, pp. 418-425.

SHAPLEY L. 1953. A value for n-person games. In Kuhn H.K. and Tucker A.W. (Eds.) *Contributions to the theory of games, 2*. Princeton (N.J.): Princeton University Press, pp. 307-315.

SHORROCKS A.F.1999. Decomposition procedures for distributional analysis: a unified framework based on the Shapley value: *University of Essex – Department of Economics, Unpublished Paper*.

SHORROCKS A.F. 2013. Decomposition procedures for distributional analysis: a unified framework based on the Shapley value. *The Journal of Economic Inequality*, Vol. 11, No. 1, pp. 99-126.

TARSITANO A. 1990. The Bonferroni index of income inequality. In Dagum C. and Zenga M. (Eds.) *Income and Wealth Distribution, Inequality and Poverty*. Berlin: Springer-Verlag, pp. 228-242.

## SUMMARY

## Bonferroni Index Decomposition and the Shapley method

The Bonferroni inequality index ($B$) remained almost forgotten until the last two decades. Recently, it has been rediscovered and furthermore, new and interesting interpretations of $B$ have been proposed. An important topic in the literature on inequality measures is their decomposition. Two main important lines of research involve decomposition by income sources and by population subgroups. Many contributions are related to $R$, less to $B$. The Shapley decomposition enables to overcome the problem related to inequality index of not being additively decomposable into the sum of within and between groups components. In this perspective Deutsch and Silber (2007) use the Shapley decomposition for $R$. In this paper, the Shapley decomposition have been applied to $B$, too. The comparison among the results obtained for both the indices allows to highlight other interesting similarities and differences among the two indices.

Giovanni Maria GIORGI, Department of Statistical Sciences, “Sapienza” University of Rome, giovannimgiorgi@gmail.com

Alessio GUANDALINI, Italian National Institute of Statistics – ISTAT, alessio.guandalini@istat.it

# UN'ANALISI DELL'IMPATTO DELLA TECNICA MISTA SULLA QUALITA' DEI DATI NELL'ADULT EDUCATION SURVEY[1]

Barbara Baldazzi, Martina Lo Conte

## 1. Introduzione

La scelta della tecnica di rilevazione è uno degli aspetti più rilevanti in un'indagine statistica, alla quale sono legati costi e tempi a disposizione, mancate risposte, qualità dei dati ed errori di misura. Oggi, sempre più spesso, per contattare i rispondenti in una stessa rilevazione, vengono utilizzate più tecniche, cercando di massimizzare i vantaggi di ognuna di esse.

L'uso combinato di diverse modalità di intervista consente, allo stesso tempo, di ridurre i costi e raggiungere un maggior numero di persone, aumentando così i tassi di risposta e la copertura della popolazione obiettivo (De Leeuw, 2005). Affiancare più tecniche ha, tuttavia, delle implicazioni, in termini sia di maggiore complessità nell'organizzazione della rilevazione, sia di un possibile impatto sulla qualità dei dati.

Utilizzando diversi strumenti di rilevazione, si possono fornire stimoli diversi nel processo di risposta, che passa attraverso la comprensione della domanda, il recupero delle informazioni nella memoria e nelle proprie conoscenze, l'integrazione di informazioni diverse per arrivare ad un giudizio complessivo, fino alla formulazione della risposta finale (Tourangeau et al., 2000).

Input orali o visivi, maggiori o minori livelli di dettaglio nelle definizioni e classificazioni, possibilità di inserire controlli e filtri, nonché la presenza o meno di un rilevatore possono quindi far si che la risposta a cui si arrivi sia diversa. Tra gli effetti più comuni dovuti alla tecnica vi sono l'attitudine a modificare la risposta corretta per presentarsi in una maniera più positiva (fenomeno noto come "desiderabilità sociale") e la tendenza a trovare delle scorciatoie ("satisficing effect"), che si verificano soprattutto nelle indagini con intervistatore (Dillman, 2000).

Da non sottovalutare il fatto, inoltre, che tecniche diverse molto spesso raggiungono popolazioni differenti sia per gli strumenti utilizzati (ad esempio i soli

[1] Sebbene l'articolo sia frutto del lavoro congiunto dei due autori sono da attribuire a Barbara Baldazzi i paragrafi 2 e 3.1 e a Martina Lo Conte i paragrafi 1 e 3.2. Il paragrafo 4 è scritto da entrambi.

possessori di telefono o chi ha un accesso a Internet), sia per le competenze necessarie (a questionari autosomministrati rispondono più frequentemente individui con livelli di istruzione più elevati) (Roberts, 2007).

Tutti questi fattori possono, pertanto, mettere a rischio la confrontabilità dei dati raccolti con tecniche diverse, che a fine rilevazione devono essere messi insieme per produrre le stime complessive dell'indagine (Martin, 2011).

Le differenze nelle stime finali a seconda del metodo utilizzato per la raccolta possono derivare, allora, sia dall'errore di misura dovuto agli strumenti di rilevazione (effetto tecnica) sia dalla diversa composizione dei campioni di rispondenti (effetto selezione). Effetto tecnica ed effetto selezione sono spesso sovrapposti e risulta difficile comprendere se l'errore di misura complessivo sia causato dall'uno o dall'altro.

L'obiettivo del nostro studio consiste nel valutare se, nell'indagine Adult Education Survey, le risposte fornite dalle famiglie intervistate con le due tecniche di rilevazione CATI e CAPI siano significativamente diverse e se queste diversità siano dovute ad una distorsione causata dalla tecnica utilizzata o ad una diversa composizione dei campioni intervistati con le due tecniche. In questo contributo si illustrano i primi risultati. Dopo una descrizione dei dati utilizzati (par. 2), si analizza la partecipazione alla formazione nei campioni di individui che hanno risposto con tecnica CATI e CAPI (par. 3.1); nel par. 3.2 si esamina la composizione dei campioni rispetto alle variabili strutturali e la propensione degli individui a partecipare con il CATI. A termine del contributo, alcune considerazioni conclusive e gli sviluppi futuri con cui si intende proseguire la ricerca (par. 4).

## 2. I dati utilizzati: l'Adult Education Survey

La rilevazione europea sulla partecipazione degli adulti alle attività formative (Adult Educational Survey – AES) ha l'obiettivo di fornire dati confrontabili a livello europeo sulla partecipazione degli adulti ad attività d'istruzione e formazione e offrire dati di qualità a supporto delle politiche nazionali di aggiornamento e riqualificazione del capitale umano[2]. La rilevazione è stata condotta in tutti i paesi dell'Unione ed era disciplinata da un regolamento Eurostat. Nel nostro Paese, la rilevazione si è svolta tra settembre e dicembre del 2012 su un campione di circa 6.000 famiglie.

[2] Le attività formative a cui gli individui possono partecipare nell'arco della loro vita sono "tutte le attività intraprese al fine di migliorare la conoscenza, le abilità e le competenze, in una prospettiva personale, civica, sociale e/o legata a una prospettiva occupazionale" (definizione del "Life Long Learning" - LLL).

È stata adottata la tecnica mista CAPI/CATI, introducendo una sostanziale novità nel sistema di raccolta dei dati (Istat, 2014). Le unità di rilevazione - le famiglie di fatto - sono state estratte dalle Liste Anagrafiche Comunali costituite dall'insieme degli individui che hanno residenza in un determinato Comune. La raccolta dei recapiti telefonici delle famiglie estratte è avvenuta attraverso: 1) l'abbinamento dei nominativi delle famiglie ai recapiti presenti negli elenchi telefonici, 2) l'invio di una lettera informativa (e successivi due solleciti) a firma del Presidente dell'Istat, nella quale è stata offerta la possibilità di scegliere l'intervista telefonica, previa indicazione di un recapito telefonico dove essere contattati, in alternativa all'intervista a domicilio in modalità CAPI[3].

La famiglia ha potuto fornire il proprio recapito telefonico attraverso: i) un numero verde, ii) un servizio interattivo di risposta vocale - IVR, iii) un portale Web. L'architettura del processo ha fatto convergere in un database integrato le informazioni raccolte (Baldazzi et al., 2013).

Il sistema di raccolta dei recapiti telefonici ha consentito di creare tre sottoinsiemi di famiglie da intervistare così composti: il 33,3% delle famiglie dato direttamente dalla famiglia; il 34,9% con il recapito telefonico trovato negli archivi telefonici; il 31,8% privo di recapito telefonico. Da questi serbatoi sono state estratte le famiglie da intervistare: una lista di famiglie da intervistare con tecnica CATI (Telefono Dato - famiglie che hanno comunicato il loro recapito telefonico), una lista di famiglie da intervistare con tecnica CATI (Telefono Archivio - con recapito abbinato da elenchi telefonici) e una lista di famiglie da intervistare con tecnica CAPI. Le famiglie delle liste CATI sono state estratte in maniera casuale dai rispettivi serbatoi con un sovracampionamento basato sul tasso di partecipazione regionale rilevato nel corso dell'indagine pilota; le famiglie della lista CAPI sono state distribuite in quartine composte da una famiglia base e tre famiglie sostitute.

[3] "*Qualora Lei desiderasse evitare il disturbo di ricevere persone in casa può chiedere che l'intervista avvenga per telefono. In tal caso chiami subito l'apposito Numero Verde (gratuito) per fornire a un operatore o a un sistema di registrazione automatica il Codice Familiare Istat posto in alto a destra e il suo numero di telefono fisso o il suo numero di telefono cellulare che saranno tenuti assolutamente riservati, presso i quali contattare Lei e i suoi familiari. In alternativa potrà accedere al sito…*".

## 3. L’analisi dei dati

### *3.1. I comportamenti formativi degli adulti intervistati con le due tecniche*

La rilevazione AES ricava informazioni sulle attività di formazione svolte dagli individui di 25-64 anni nei 12 mesi precedenti l’intervista[4]. Si distinguono in attività di formazione formale, vale a dire i corsi del sistema d’istruzione volti al conseguimento di titoli riconosciuti dal sistema nazionale delle qualificazioni (dalla licenza elementare al dottorato di ricerca) e attività di formazione non formale, ossia attività strutturate e organizzate che, tuttavia, non permettono di acquisire un titolo di studio. Analizzando le variabili obiettivo (1- partecipazione ad almeno una attività sia formale sia non formale, 2 – partecipazione ad almeno una attività formale, 3- partecipazione ad almeno una attività di tipo non formale) differenziate per tecnica si rilevano alcune differenze tra chi è coinvolto in attività di formazione e chi non le svolge, in base alla tecnica con cui hanno risposto.

**Tabella 1 –** *La partecipazione degli adulti ad attività formative per tecnica di rilevazione.*

| Partecipazione ad attività di formazione | CAPI | CATI TD | CATI TA | $\chi^2$ | prob. |
|---|---|---|---|---|---|
| Almeno un'attività di formazione: SI | 29.5 | 41.0 | 33.2 | | |
| Almeno un'attività di formazione: NO | 70.5 | 59.0 | 66.8 | 127.24 | <.0001 |
| Almeno un'attività di formazione formale: SI | 2.1 | 2.8 | 2.5 | | |
| Almeno un'attività di formazione formale: NO | 97.9 | 97.3 | 97.5 | 2.63 | 0.269 |
| Almeno un'attività di formazione non formale SI | 29.4 | 41.2 | 35.1 | | |
| Almeno un'attività di formazione non formale NO | 70.7 | 58.8 | 64.9 | 96.17 | <.0001 |

*Il collettivo raggiunto è composto da 8703 individui di 25-64 anni così distribuiti: 2140 intervistati con tecnica CAPI, 5637 intervistati con tecnica CATI con telefono dato e 926 individui con tecnica CATI con telefono preso dagli archivi.*

Tra coloro che sono stati intervistati con la tecnica CAPI, oltre il 70% ha dichiarato di non aver partecipato a nessuna attività di formazione (Tabella 1). La stessa percentuale scende al di sotto del 60% per i rispondenti CATI-TD che hanno fornito volontariamente il proprio telefono. Le differenze tra le tecniche di intervista risultano significative: soltanto la partecipazione ad attività di formazione di tipo formale (i corsi di istruzione scolastica e universitaria) risulta non differenziata in base alla tecnica di intervista. Ciò deriva dalla scarsa numerosità di coloro che, nella fascia di età di 25-64 anni scelgono questo tipo di corsi che sono

[4] Vi rientrano le attività di formazione connotate dall’intenzionalità, cioè da una precisa volontà di migliorare il proprio livello di conoscenza, comprensione o abilità, finanziate con risorse pubbliche o private, dal datore di lavoro o dal diretto interessato, erogate con modalità diverse (usando strumenti tradizionali o le ICT) a cui si accede indipendentemente dall’età e dalla condizione nel mercato del lavoro.

solitamente destinati a fasce più giovani della popolazione. Già da una prima lettura emergono, quindi, comportamenti diversi per ciò che riguarda le variabili obiettivo della rilevazione, tra le persone che hanno risposto con tecniche differenti.

Applicando un modello di regressione alla variabile dipendente "partecipazione ad una attività di formazione", è possibile modellare la relazione tra l'esito dicotomico da esaminare (la partecipazione alla formazione o meno) e un set di variabili che possono essere sia dicotomiche sia categoriche per valutare gli effetti di ciascun indicatore considerato sul verificarsi dell'evento di interesse.

Il modello logistico applicato consente di valutare la probabilità che un individuo di 25-64 anni ha di partecipare ad una attività di formazione (formale o non formale), esaminata in riferimento ad ogni singolo indicatore considerato. Le variabili esplicative prese in esame sono relative alle dinamiche sociali ed economiche (genere, età, tipologia familiare, titolo di studio e condizione professionale e occupazionale), alla dimensione geografica e territoriale (ripartizione geografica e tipo di comune), all'abilità nell'uso delle nuove tecnologie che rappresentano uno dei tre strumenti utilizzabili per fornire il recapito telefonico (uso del PC e di Internet) e alla tecnica di intervista, inserita, appunto, per valutare l'influenza della tecnica al netto delle altre variabili.

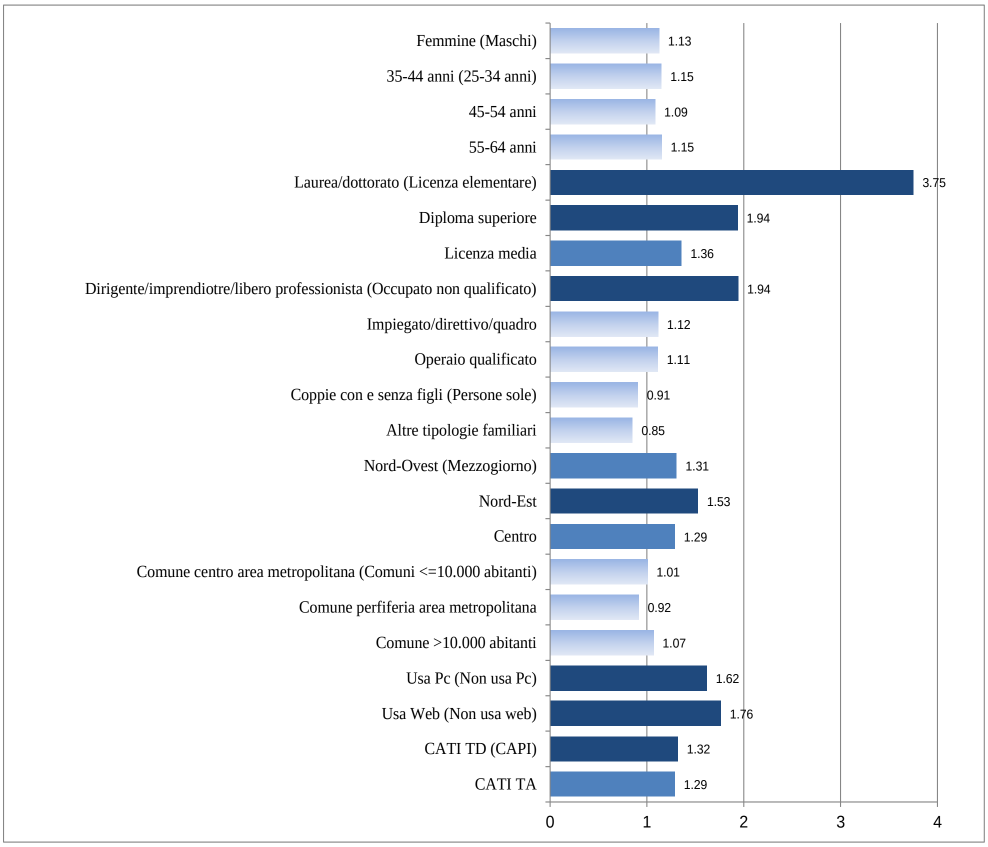
I risultati ottenuti (Figura 1) individuano nel titolo di studio, nella condizione professionale e occupazionale e nella ripartizione geografica i fattori che risultano essere maggiormente "correlati" con la partecipazione alla formazione. Significativi sono, inoltre, l'uso di internet, del pc e la differente tecnica di rilevazione. Le variabili genere, classe d'età, tipologia familiare e tipo di comune non risultano, invece, significative.

La partecipazione alla formazione è fortemente caratterizzata dal livello del titolo di studio, dalla professione e in parte dalla ripartizione geografica. Le odds[5] di partecipazione alla formazione per coloro che vivono nel Nord-Est sono di 1,5 volte superiori le odds di chi vive nel Mezzogiorno. La probabilità di fare un'attività di formazione, al netto dell'effetto degli altri fattori, è di quasi 4 volte maggiore tra coloro che hanno una istruzione elevata (Laurea o più) rispetto a coloro che hanno soltanto la licenza elementare; è di 2 volte superiore per coloro che hanno conseguito il diploma di scuola superiore.

La probabilità di partecipare alla formazione dei dirigenti, imprenditori e libero professionisti è significativamente più elevata: 2 volte superiore rispetto alla probabilità di partecipare alla formazione dei lavoratori non qualificati.

[5] L'odds ratio è il rapporto tra la frequenza con la quale un evento si verifica in un gruppo (ad esempio nei residenti nel Nord-Ovest) e la frequenza con la quale lo stesso evento si verifica in un gruppo di riferimento (i residenti nel Mezzogiorno).

**Figura 1** – *Modello di regressione logistica. Variabile risposta: Propensione alla partecipazione alle attività di formazione*



Relativamente alla tecnica di rilevazione un impatto si osserva per coloro che hanno fornito il loro recapito telefonico (CATI-TD): la probabilità di partecipare ad un'attività di formazione risulta essere maggiore del 32% rispetto a coloro che sono stati intervistati con tecnica CAPI. La partecipazione alla formazione degli intervistati con il CATI abbinato al telefono rilevato da archivio, non risulta, invece, significativamente diversa dagli intervistati con modalità CAPI. Il modello conferma, quindi, una differenza tra i 3 collettivi definitivi tramite la tecnica di rilevazione: il collettivo intervistato attraverso il CATI-TD, ossia con il telefono dato direttamente dall'individuo, ha una propensione maggiore a partecipare alla formazione rispetto agli altri due collettivi (tecnica CAPI e CATI-TA). Le analisi successive saranno volte a determinare quanto ciò sia dovuto all'effetto tecnica o ad un effetto selezione del campione.

### *3.2. Analisi della composizione dei campioni e della propensione a rispondere CATI*

Poiché gli individui hanno una diversa propensione a rispondere con una tecnica piuttosto che con un'altra, come si è detto, la composizione dei campioni intervistati con tecniche diverse spesso non è casuale e determina un effetto selezione nei diversi campioni di rispondenti. Ciò può verificarsi soprattutto nelle indagini a tecnica mista in cui venga data la possibilità di scegliere la modalità di intervista (come è avvenuto per l'AES 2011). La differenza nelle stime risultanti da dati raccolti con tecniche diverse, quindi, potrebbe essere dovuta alle caratteristiche delle diverse tipologie di individui che hanno risposto con una tecnica specifica.

In questo paragrafo si vuole valutare se esistono differenze strutturali significative tra rispondenti CATI e CAPI, e verificare così la presenza di un effetto selezione dei rispondenti che hanno usufruito di tecniche diverse. Infatti, se particolari tipologie di individui avessero una maggiore propensione a rispondere con una tecnica piuttosto che con un'altra, e partecipassero più frequentemente con una tecnica, occorrerebbe tenere conto di questo effetto di distorsione prima di valutare se esistano differenze dovute alla tecnica di rilevazione (Camillo et al., 2008).

L'ipotesi alla base è che i rispondenti con una più alta predisposizione a rispondere con una certa tecnica potrebbero avere in comune caratteristiche particolari e quindi appartenere a gruppi omogenei rispetto ad alcune variabili. Le differenze nelle distribuzioni per tecnica, pertanto, possono dipendere da queste caratteristiche e non propriamente dalla tecnica utilizzata.

Analizzando le principali variabili socio-demografiche nei tre sotto-campioni (Tabella 2), si osserva che, rispetto alle età, il collettivo che ha scelto il CATI-TD contiene più individui della fascia di età 35-44, mentre con il CATI-TA, ossia con telefono da archivio, e con il CAPI sono state raggiunte più persone con età più elevata.

I gruppi intervistati con il CAPI e il CATI-TA risultano simili per molte variabili: risiedono maggiormente nel Sud e nelle Isole, in comuni che non sono centro di aree metropolitane, né limitrofi a queste, hanno livelli di istruzione più bassi e utilizzano meno il computer e Internet. Di contro, coloro che hanno scelto il CATI-TD sono più spesso residenti nel Nord Ovest, possiedono titoli di studio più elevati (diploma, laurea e dottorato) e fanno un maggior uso di Internet e PC.

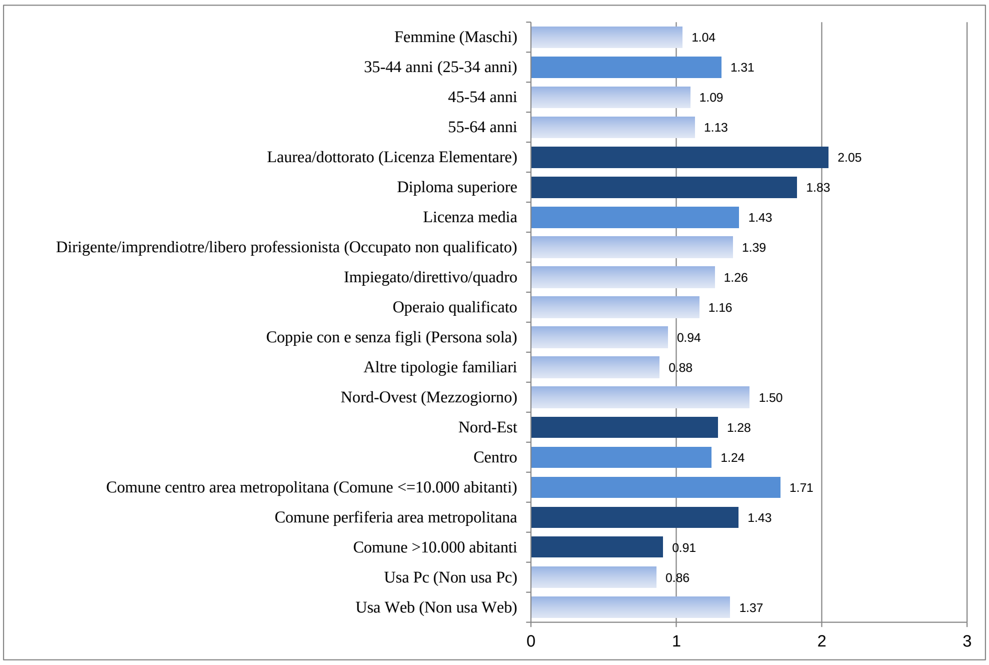
**Tabella 2** – *Caratteristiche strutturali dei campioni per tecnica di rilevazione.*

| Variabili strutturali del campione | CAPI | CATI TD | CATI TA | $\chi^2$ | prob. |
|---|---|---|---|---|---|
| Maschi | 50.0 | 47.9 | 46.0 | | |
| Femmine | 50.0 | 52.1 | 54.0 | 4.9 | 0.1 |
| 25-34 anni | 20.1 | 17.4 | 12.3 | | |
| 35-44 anni | 26.3 | 30.3 | 21.3 | | |
| 45-54 anni | 29.0 | 28.3 | 31.5 | | |
| 55-64 anni | 24.6 | 24.1 | 34.9 | 89.7 | <.0001 |
| Persone sole | 11.6 | 11.5 | 6.3 | | |
| Coppie (con e senza figli) | 75.0 | 76.6 | 83.1 | | |
| Altre tipologie | 13.5 | 11.9 | 10.7 | 31.8 | <.0001 |
| Nord-Ovest | 25.6 | 31.9 | 25.2 | | |
| Nord-Est | 25.1 | 22.1 | 18.3 | | |
| Centro | 19.7 | 21.5 | 19.1 | | |
| Sud e Isole | 29.6 | 24.6 | 37.5 | 100.4 | <.0001 |
| Comune Centro Area Metropolitana | 9.1 | 17.2 | 12.1 | | |
| Comune Periferia Area Metropolitana | 10.8 | 15.8 | 13.5 | | |
| Comune <=10.000 abitanti | 31.1 | 27.1 | 28.1 | | |
| Comune >10.000 abitanti | 49.1 | 39.9 | 46.3 | 143.4 | <.0001 |
| Laurea/dottorato | 12.6 | 21.8 | 19.0 | | |
| Diploma | 31.8 | 38.4 | 34.0 | | |
| Licenza media | 43.5 | 34.1 | 37.3 | | |
| Licenza elementare | 12.2 | 5.7 | 9.7 | 216.6 | <.0001 |
| Dirigenti, Imprenditori, Liberi professionisti | 20.6 | 31.0 | 25.3 | | |
| Direttivi, Quadri, Impiegati | 17.7 | 19.1 | 16.5 | | |
| Operai qualificati | 15.7 | 12.6 | 14.0 | | |
| Personale non qualificato | 8.0 | 4.5 | 3.1 | | |
| Non occupati | 38.0 | 32.9 | 41.0 | 145.9 | <.0001 |
| Uso pc NO | 37.9 | 26.8 | 35.6 | | |
| Uso Pc SI | 62.2 | 73.2 | 64.4 | 102.1 | <.0001 |
| Uso Internet 12 mesi NO | 34.6 | 22.1 | 31.1 | | |
| Uso Internet 12 mesi SI | 65.4 | 77.9 | 68.9 | 139.6 | <.0001 |

Alla luce di questa maggiore coerenza dei gruppi CAPI e CATI-TA, si è deciso di approfondire, più che la tecnica di intervista utilizzata, la propensione a rispondere telefonicamente, attraverso la variabile che indica se hanno fornito o meno il telefono volontariamente per partecipare all'indagine. È stato quindi applicato un modello di regressione logistica al fine di individuare le variabili che influiscono maggiormente sulla propensione a partecipare con il CATI (Figura 2).

In generale risultano più predisposti a fornire il proprio recapito e ad essere successivamente intervistati per telefono gli individui più istruiti (i laureati hanno una probabilità doppia rispetto a chi ha la licenza elementare e i diplomati dell'80% più alta). Nel territorio, i più propensi sono i residenti nel Nord Est e nel Centro, rispetto a chi vive nel Mezzogiorno, e chi abita nelle grandi aree metropolitane o nei comuni limitrofi rispetto a chi vive nei Comuni più piccoli. Non risultano significative, invece, le variabili relative all'uso del PC e di Internet.

**Figura 2** – *Modello di regressione logistica. Variabile risposta: Propensione a partecipare all'intervista con il CATI-TD fornendo il proprio recapito telefonico*



La presenza di profili differenti sia tra i rispondenti a tecniche diverse, che con una diversa propensione a fornire il proprio recapito telefonico e a rispondere telefonicamente, indica la presenza di un possibile effetto selezione che deve essere tenuto sotto controllo nel momento in cui si analizzi l'effetto tecnica.

## 4. Conclusioni

L’uso della tecnica mista nelle rilevazioni porta con sé gli indubbi vantaggi di ridurre i costi e aumentare la copertura, ma può inficiare la qualità dei dati. Pertanto è opportuno mettere insieme i dati con molta cautela, verificando l’equivalenza delle misurazioni effettuate con tecniche differenti (Martin e Lynn, 2011).

Quando si riscontrano delle differenze nelle stime da tecniche diverse, non è sempre facile distinguere se queste siano dovute alle caratteristiche del quesito (formulazione delle domande o modalità di risposta, grado di sensibilità o complessità) o alle caratteristiche della tecnica (la presenza o meno di un intervistatore o l’ausilio del computer,  il canale di comunicazione, visivo e orale) o anche alle caratteristiche del rispondente (la propensione a fornire risposte “socialmente desiderabili” o a trovare scorciatoie nelle risposte, con il “satisficing effect”) (Roberts et al., 2006).

Nelle indagini a tecnica mista, effetto tecnica ed effetto selezione sono spesso sovrapposti e risulta difficile comprendere se l’errore di misura sia causato dall’uno o dall’altro. In letteratura vi sono alcuni studi che cercano di separare l’effetto selezione dall’effetto tecnica (Vanniewenhuize et al., 2010; Vanniewenhuize e Loosveldt, 2012). Alcuni autori hanno utilizzato la *propensity score* di Rosenbaum e Rubin (1983) per tenere sotto controllo l’effetto selezione (Klaush et al., 2013; Camillo et al., 2008).

In questo contributo si sono illustrati i primi risultati di una ricerca che mira a verificare la presenza di un effetto tecnica nell’indagine AES 2011, che utilizza per la prima volta la tecnica mista. Le differenze significative nella partecipazione degli adulti a corsi di formazione  rispetto alla tecnica di intervista richiedono una analisi attenta delle cause di tali diversità. Si tratta di effetto tecnica o effetto selezione? Ad esempio, con il CAPI potrebbero aver risposto di fare meno corsi di formazione, perché con questa tecnica si sono intervistate soprattutto persone più anziane e con un titolo di studio più basso.  In alternativa, o in aggiunta, le differenze tra distribuzioni per tipo di intervista CAPI e CATI potrebbero attribuirsi agli strumenti di indagine. Pur basandosi entrambe sulla presenza di un rilevatore, infatti, al telefono l’intervista può provocare stimoli ben diversi rispetto ad una situazione faccia a faccia. Si pensi alla definizione stessa di corsi di formazione – non sempre semplice ed univoca – che può essere spiegata in maniera migliore in presenza piuttosto che al telefono.

Dalle analisi fin qui effettuate si è riscontrata una diversa composizione dei tre sotto-campioni (CAPI, CATI-TD con telefono dato volontariamente e CATI-TA con telefono da archivio). Inoltre, sono state individuate le variabili che più influiscono sulla propensione a rispondere telefonicamente. Tale propensione sarà utilizzata per tenere sotto controllo l’effetto selezione, raggruppando in cluster omogenei rispetto a tale propensione, seguendo il metodo usato da Camillo, Conti e

Ghiselli (2008), che si basa sulle propensity score di Rosenbaum e Rubin (1983). Saranno testati, inoltre, altri approcci per analizzare l'effetto tecnica controllando per l'effetto selezione.

Il numero crescente di indagini a tecnica mista, se da un lato offre l'opportunità di effettuare ricerche a minor costo, dall'altro richiede evidenze che aiutino a prendere decisioni sui molteplici aspetti organizzativi e metodologici da tenere in considerazione quando si affianchino più tecniche di rilevazione.

## Riferimenti bibliografici

BALDAZZI B., BIANCHI A., MARTINO A. E., PALADINI P. 2013. *Innovazioni di processo e uso delle variabili testuali: il caso dell'Adult Education Survey,* Rivista Italiana Economia Demografia e Statistica, Vol. LXVII, No. 3-4.

CAMILLO F., CONTI V., GHISELLI S. 2008. L'integrazione di differenti tecniche di rilevazione dei dati utilizzando la propensity score. Università di Bologna. Consorzio Interuniversitario Almalaurea.

DE LEEUW E. D. 2005. *To Mix or Not to Mix Data Collection Modes in Surveys*, Journal of Official Statistics, No. 21.

DILLMAN D. A. 2000. *Mail and Internet surveys--The tailored design method*. New York. John Wiley & Sons, Inc.

KLAUSH T., HOX J.J., SCHOUTEN B. 2013, Measurement effects of survey mode on the equivalence of attitudinal rating scale questions, Sociological methods and research 42 (3) 227-263, SAGE.

ISTAT 2014. *La modernizzazione delle tecniche di rilevazione nelle indagini socio-economiche sulle famiglie.*

MARTIN P. 2011. *What makes a good mix? Chances and challenges of mixed mode data collection in the ESS*, Centre for Comparative Social Surveys Working paper Series, Paper no. 02, February 2011.

MARTIN P., LYNN P. 2011. *The effects of mixed mode survey designs on simple and complex analyses*, Institute for Social & Economic Research, No. 2011-28, November 2011.

ROBERTS C. E., JACLE A., LYNN, P. 2006. *Causes of Mode Effects: Separating out Interviewer and Stimulus Effects in Comparisons of Face-to-Face and Telephone Surveys*. Proceedings of the Survey Research Methods Section. American Statistical Association. http://www.amstat.org/Sections/Srms/Proceedings/

ROBERTS C. 2007. *Mixing modes of data collection in surveys: A methodological review*, Southampton: ESRC National Centre for Research Methods. NCRM Methods Review Papers, NCRM/008.

ROSENBAUM P.R., RUBIN D.B. 1983. *The central role of the propensity score in observational studies for causal effects*, Biometrika, Vol. 70, No.1 (Apr. 1983), pp.41-55.

TOURANGEAU R., RIPS L. J., RASINSKI K. 2000, *The psychology of survey response*, Cambridge University Press.

VANNIEWENHUIZE J., LOOSVELDT G., MOLENBERGHS G. 2010, *A method for evaluating mode effects in mixed-mode surveys*, Public Opinion Quarterly, Vol.74, No.5, 2010, pp.1027-1045.

VANNIEWENHUIZE J., LOOSVELDT G. 2012, *Evaluating Relative mode effects in mixed-mode surveys: Three methods to disentangle selection and measurement effects*, Sociological methods and research 42 (1) 82-104, SAGE.

## SUMMARY

### The Analysis of the Impact of Mixed-Mode on the Quality of Italian Adult Education Survey

In this study we analyze data from the Italian Adult Education Survey that uses a combination of telephone and face-to-face interviews (CATI and CAPI modes). The aim of our research is to assess whether the answers given by the households interviewed with different techniques are significantly different and whether these differences are due to mode-related measurement error or to different samples' composition. In this paper we illustrate some preliminary results. The analysis shows that the differences in some estimates are due, at least partly, to different respondents selected through the mode (mode selection effect). We also identify the variables that most affect the propensity to respond by telephone.

Barbara BALDAZZI, Istat baldazzi@istat.it
Martina LO CONTE, Istat loconte@istat.it

# MEASURING LOCAL WELL-BEING: A COMPARISON AMONG AGGREGATIVE METHODS FOR THE EQUITABLE AND SUSTAINABLE WELL-BEING

Francesco M. Chelli, Mariateresa Ciommi, Alessandra Emili, Chiara Gigliarano, Stefania Taralli

## 1. Introduction

A specific requirement in analysing well-being at local level is to bring out territorial disparities in order to assess the equity dimension of well-being and the territorial cohesion. In this work we explore different aggregation methods using the elementary indicators released by Istat under the "Provinces' BES" Project[1].

After selecting 41 indicators from the original dataset of 88 indicators, available for the year 2014, we apply three different aggregation methods, consisting of a weighted average of the standardized indicators. Firstly, we compute the Adjusted Mazziotta-Pareto Index (AMPI) to account for horizontal variability. Then we propose a different aggregation procedure, based on the Gini index (GW) that accounts for vertical variability. Finally, we propose a mixed approach that accounts both for horizontal and vertical variability, based on the Adjusted Mazziotta-Pareto but modified using a weighting system based on the Gini coefficients of the elementary indicators (GAMPI).

The rest of the paper is structured as follows: Section 2 presents the methodology for the construction of composite indexes; in details after describing the adopted normalization method, we describe, in details, the three aggregation methods. Section 3 presents the application of the above-mentioned methods to the Provinces' BES dataset. We compute the ranking correlation among the ranking produced by GW, AMPI and GAMPI. In addition, we focus on the GAMPI methods, classifying provinces in decile according the value of the aggregate indicator for each domain and finally, we compute the correlation among composite indicators of each domain. Section 4 concludes.

[1] The Province's Bes Project was launched in 2011 as a pilot study by the Province of Pesaro e Urbino in partnership with Istat, and from 2013 it has been extended other Italian Provinces and Metropolitan Cities. In 2015 it was listed in the National Statistical Programme as Statistical Information System.

## 2. Methodology

Since data have different unit of measurement, a preliminary step is necessary, in order to ensure the comparability. Among the huge class of normalization methods, we apply the methodology adopted in Mazziotta and Pareto (2016). Here, we apply the so-called *re-scaling* approach according to two '*goalposts'*, such that the interval length is one and a reference value (i.e. the Italian average) is the central value of the range (that is 0.5). More in details, let us define a matrix $X_h$ for each well-being domain $h = 1, \dots, H$, whose the general element $x_{ijh}$ is the value the of the *j-th* elementary indicator of the domain *h* for the *i-th* local unit (e.g., the province), with $j = 1, \dots, n_h$ and $i = 1, \dots, N$. We denote by $\mathrm{Max}_{.\mathrm{jh}}$and $\mathrm{Min}_{.\mathrm{jh}}$ respectively the minimum and the maximum value of the indicator *j* of the domain *h* across all the local units, whereas and $\mathrm{Rif}_{.\mathrm{jh}}$ represents a reference value, that is the average value for any indicator. Following Mazziotta and Pareto (2016) we compute the two *'goalposts'*, that is $\underline{x}_{.jh} = Rif_{.jh} - \Delta_{.jh}$ and $\overline{x}_{.jh} = Rif_{.jh} + \Delta_{.jh}$ where $\Delta_{.jh} = (Max_{.jh}$ - $Min_{.jh}) / 2$.[2] The goalposts allow scaling indicators; in addition, the formulation captures changes over time since the reference values could be set as the extreme values of each indicator over the time period considered. We note that the goalposts are obtained adding or subtracting the quantity $\Delta_{.jh}$ obtaining the new minimum $\underline{x}_{.jh}$ and maximum $\overline{x}_{.jh}$, respectively. In addition, these values enter in the normalization step, that is, the normalized indicator *j* belonging to domain $h$ for the *i*-th province, denoted with $\mathrm{I}_{\mathrm{ijh}}$, is then computed as follows:

$$\begin{cases} I_{ijh} = \dfrac{x_{ijh} - \underline{x}_{.jh}}{\overline{x}_{.jh} - \underline{x}_{.jh}} & (1)\ if\ the\ indicator\ has\ a\ positive\ polarity \\ I_{ijh} = \dfrac{\overline{x}_{.jh} - x_{ijh}}{\overline{x}_{.jh} - \underline{x}_{.jh}} & (2)\ if\ the\ indicator\ has\ a\ negative\ polarity \end{cases}$$

Equations (1) and (2) allow the normalization of the indicators according to the relation, namely polarity, with the phenomenon to be measured. That is, if the indicator has positive 'polarity', indicating that an increase in the indicator corresponds to an increase in the overall domain, equation (1) is used. By contrast, if the indicator shows negative relationship with the phenomenon, then we adopt the normalization expressed by equation (2).

Finally, a Linear Scale Technique is here adopted to re-scale indicators into a fixed range, that is an interval of length 60, such that the Italian mean (goalposts) is

[2] As observed in Mazziotta and Pareto (2016), using goalposts the re-scaling of the elementary indicators takes account changes over time.

fixed and equal to 100 (Mazziotta and Pareto, 2016; Massoli et. al., 2014). Formally:

$$r_{ijh} = I_{ijh} \cdot 60 + 70$$

and $r_{ijh}$ ranges in [70,130].

Using $r_{ijh}$ coefficients, we compute the simple arithmetic mean of the elementary indicators among a domain $h$, that is:

$$EW_{ih} = \sum_{\mathrm{j}=1}^{n_h} \left( r_{ijh} \cdot \frac{1}{n_h} \right)$$

where $n_h$ is the number of elementary indicators in the h-th domain and $i$ denotes a generic local unit.

## 2.1. Horizontal variability: The Adjusted Mazziotta-Pareto Method

Since 2015, ISTAT (2015) has adopted a non-compensatory approach based on a penalty function that is the so-called adjusted Mazziotta and Pareto (2013a, 2016), hereafter (AMPI). It is based on the arithmetic mean of the elementary indicators, adjusted by a function that takes into account the horizontal variability of the indicators. Thus, the AMPI for the i-th local unit and the h-th well-being domain is given by:

$$AMPI_{ih} = EW_{ih} \pm S_{ih} \cdot cv_{ih}$$

where $EW_{ih}$ is the arithmetic mean of the normalized indicators, and $S_{ih}$ and $cv_{ih}$ are respectively the standard deviation and the coefficient of variation of all the normalized indicators $r_{ijh}$ belonging to domain h for local unit i, i.e. $S_{r_{ih}} = \sqrt{\frac{\sum_{j=1}^{n_h} (r_{ijh} - EW_{ih})^2}{n_h}}$; $cv_{ih} = \frac{S_{ih}}{EW_{ih}}$ and the double sign depends on the polarity of the composite index with respect to the well-being. The negative sign (-) is used when the composite indicator EW is positively related to the construct of well-being, while the positive sing (+) when EW is negatively related to well-being.

Thus, the index is a combination of the average effect (the additive function - $EW_{ih}$) and the penalty effect (the function of horizontal variability - $S_{ih} \cdot cv_{ih}$), based on the variability among indicators in that province.

This method penalizes the local units that, mean being equal, present a more unbalanced distribution among the indicators values. Therefore, if within a BES domain, an Italian province presents a low value in one indicator and a high value in another, then that province receives a penalty without compensation. On the contrary, if there is a low variability among the indicators for that province, than the penalty effect is minimal. Despite all the components have the same importance in computing the composite index, to obtain a high value of the composite index, all the elementary indicators must assume high values. (Mazziotta and Pareto, 2013b).

However, applying the AMPI method to the BES at NUTS3 level, what emerges is that, within each domain, the contribution of the factor $S_{ih} \cdot cv_{ih}$ is small. In fact, if we compute the ranking correlation produced by AMPI and the ranking obtained removing the penalty coefficient, we observe that for all the domains, it ranges in [0.87; 0.99].[3]

### 2.2. Vertical variability: weighting using Gini

To overcome this drawback, we propose a different approach, introducing a weighting schema based on the Gini index of concentration.

More in details, within each domain, the weight associated to each elementary indicator is calculated as the Gini index of that indicator normalized by the sum of the Gini indices of all indicators in the same domain. Therefore, the composite index for the i-th local unit and the h-th well-being domain is defined as:

$$GW_{ih} = \frac{1}{G_h} \cdot \sum_{j=1}^{n_h} \left( r_{ijh} \cdot G_{.jh} \right)$$

where $G_{.jh}$ represents the Gini index of the j-th indicator belonging to the h-th domain and $G_h = \sum_{j=1}^{n_h} G_{.jh}$.

A possible drawback in the use of the Gini index is that it is usually defined for transferable variables. However, according to Becchetti et al. (2014), the possibility of redistribution is not essential since in this situation, the Gini index can be considered as a synthetic measure of the distribution of resources.

In addition, the use of the Gini index has also a theoretical justification since it is in line with the recommendations by the Commission on the Measurement of Economic Performance and Social Progress, according to which average measures should be accompanied by indicators reflecting their distribution (Stiglitz, et al. 2009).

Compared to the AMPI approach, these weights may be considered as a vertical variability coefficient: a more unequal distribution of an elementary indicator among provinces implies a greater weight associated to this indicator (Chelli et. al 2015).

[3] The results of the rank-correlation are available upon request.

Therefore, this aggregation method emphasizes the differences resulting from to more variables indicators and benefits those provinces with the highest values in the indicators which are more unequally distributed among the provinces. The peculiarity of this approach is to assume that the relative importance of the elementary indicators depends only on their distributions among the Italian provinces.

### *2.3. An unified approach: the Gini-based adjusted Mazziotta-Pareto Index*

Since both methods have advantages, we decide to merge.

The ***Gini-based adjusted Mazziotta-Pareto Index (GAMPI)*** modifies the adjusted Mazziotta-Pareto Index, by computing first a Gini-based weighted average of the elementary indicators, and then adjusting the weighted mean by the *penalty* function. The GAMPI for the *i*-th local unit and the *h*-th domain is given by:

$$GAMPI_{ih} = \frac{1}{G_h} \cdot \sum_{j=1}^{n_h} \left( r_{ijh} \cdot G_{.jh} \right) \pm S_{ih} \cdot cv_{ih} = GW_{ih} \pm S_{ih} \cdot cv_{ih}$$

Compared to the AMPI, the *mean* effect is adjusted by the *vertical variability* of the elementary indicators, while the *penalty* function considers the *horizontal variability* of the indicators in each province.

Similar to the *GW* method, greater weights are given to those indicators with more unequal distributions across provinces, while the *penalty* function benefits those provinces with a balanced distribution of the indicators of the same domain. The advantage of this approach is to consider both the variability of the indicators belonging to the same BES domain for a specific province (*horizontal variability*) and the distribution of a specific indicator across all the Italian provinces (*vertical variability*). Therefore, this approach penalizes more the provinces having both low values in the indicators with a more unequally distribution among the Italian provinces and with greater variability among the indicators within the same domain.

## 3. An illustrative example

The starting point of the analysis is a subset of the 88 variables constituting the Provinces' BES dataset for 2014. (Table 1). The step of selection of the 41 indicators available for 110 Italian province allowed us to improve the relevance of the indicators to the BES construct, as well as the quality of the dataset since the reliability, robustness and significance at this territorial level are not always satisfactory (ISTAT, 2015 pp. 52 and 125).

**Table 1 –** *List of elementary indicators for each domain*

| Domain | Indicator | Sign |
|---|---|---|
| 1 - Health | Life expectancy at birth (male) | + |
| | Life expectancy at birth (female) | + |
| | Avoidable mortality rate | - |
| 2 - Education and Training | Early leavers from education and training | - |
| | People of working age with secondary education degree or lower | - |
| | Student’s level of literacy | + |
| | Student’s level of digital competences | + |
| | Participation in lifelong learning (25-64 years) | + |
| 3 - Work and life balance | Non participation rate (15-74 years) | - |
| | Gender difference in non-participation rates (F-M) | - |
| | Employment rate (20-64 years) | + |
| | Gender differences in employment rates (M-F) | - |
| | Youth employment rate (15-29 years) | + |
| | Risk of severe accidents at work | - |
| 4 - Economic well-being | Gross disposable income per household | + |
| | Average amount of family assets | + |
| | Gender differences in the average wage of employees (M-F) | - |
| | Age groups differences in the average wage of employees | - |
| 5 - Social relationship | Non-profit organizations | + |
| | Volunteers in no-profit organizations (per 100 residents aged 14 and over) | + |
| 6 - Politics and istitutions | Turnout in the European Parliament elections | + |
| | Turnout in Provincial councils’ elections | + |
| | Percentage of women elected in municipal councils | + |
| | Percentage of young people (<40 years old) elected in municipal councils | + |
| 7 - Security | Violent crimes reported | - |
| 8 - Landscape and cultural heritage | Conservation of the historic urban fabrics | + |
| | Density of urban parks of historical interest | + |
| | Museums and similar institutions | + |
| 9 - Environment | Urban green areas | + |
| | Overruns of the daily limits of air pollution - | - |
| | Energy produced from renewable sources (electricity) | + |
| | Landfill storage of waste | - |
| 10 - Research and innovation | Propensity to patent (applications) | + |
| | New graduates in S & T (total in the year) | + |
| | Industries specialization in knowledge-intensive sectors | + |
| 11 - Quality of services | Irregularities in electricity supply | - |
| | Children (0-2 years old) receiving services for early childhood | + |
| | Separate collection of municipal waste | + |
| | Prisons overcrowding index | - |
| | Regional health services outflow (hospital admittance) | - |
| | Urban public transport networks density | + |

**Table 2** – *Rank correlation among synthetic methods in each domain*

| **Health** | GW | AMPI | GAMPI |
|---|---|---|---|
| GW | 1 | | |
| AMPI | 0,9051 | 1 | |
| GAMPI | 0,9989 | 0,9099 | 1 |

| **Education and training** | GW | AMPI | GAMPI |
|---|---|---|---|
| GW | 1 | | |
| AMPI | 0,9133 | 1 | |
| GAMPI | 0,9983 | 0,9169 | 1 |

| **Work and life balance** | GW | AMPI | GAMPI |
|---|---|---|---|
| GW | 1 | | |
| AMPI | 0,995 | 1 | |
| GAMPI | 0,9982 | 0,9963 | 1 |

| **Economic well-being** | GW | AMPI | GAMPI |
|---|---|---|---|
| GW | 1 | | |
| AMPI | 0,9406 | 1 | |
| GAMPI | 0,9524 | 0,9895 | 1 |

| **Social relationships** | GW | AMPI | GAMPI |
|---|---|---|---|
| GW | 1 | | |
| AMPI | 0,9982 | 1 | |
| GAMPI | 0,9999 | 0,9983 | 1 |

| **Politics and institutions** | GW | AMPI | GAMPI |
|---|---|---|---|
| GW | 1 | | |
| AMPI | 0,9353 | 1 | |
| GAMPI | 0,9852 | 0,968 | 1 |

| **Landscape and culturale heritage** | GW | AMPI | GAMPI |
|---|---|---|---|
| GW | 1 | | |
| AMPI | 0,9025 | 1 | |
| GAMPI | 0,9887 | 0,9082 | 1 |

| **Environment** | GW | AMPI | GAMPI |
|---|---|---|---|
| GW | 1 | | |
| AMPI | 0,9895 | 1 | |
| GAMPI | 0,9943 | 0,9911 | 1 |

| **Research and innovation** | GW | AMPI | GAMPI |
|---|---|---|---|
| GW | 1 | | |
| AMPI | 0,8772 | 1 | |
| GAMPI | 0,9973 | 0,8716 | 1 |

| **Quality of services** | GW | AMPI | GAMPI |
|---|---|---|---|
| GW | 1 | | |
| AMPI | 0,9768 | 1 | |
| GAMPI | 0,9949 | 0,984 | 1 |

First of all, with the 41 indicators, we compute the correlation among GW, AMPI and GAMPI (Table 2). What emerges is the high correlation among the three methods.

As illustrative example, we compute the GAMPI index. Figure 1 provides a geographical comparison of the Italian provinces in each BES domain. The *"Work and life balance"* and *"Research and innovation"* are the domains that better differentiate the provinces. There is, in fact, a clear gap between the provinces in the North, in the Center and in the South of Italy, which highlight the detriment of the southern provinces. The latter are also disadvantaged in the domains *"Quality of services"* and *"Social relationships"*, excluding the Sardinian provinces. Conversely, the South of Italy is advantaged in environmental domains as much as some northern provinces, such as the provinces of Aosta, Trento and Bolzano. The provinces in the North of Italy show the best performances in the domains *"Health"* and *"Education and training"*. In addition, they are advantaged in *"Economic well-being"* domain, as a consequent of a more developed economic activity. Finally, the provinces in the Center present better performance in

*"Landscape and cultural heritage"*. In addition, we account for the degree of correlation among the BES domains (Table 3).

Table 3 shows the value of the correlation coefficients among the composite indicators of each domain resulting from the GAMPI approach. The lowest levels of correlation are in correspondence with the domain "*Environment*", which is poorly correlated with all the domains and in particular with "*Health*", "*Education and training*" and "*Work and life balance*", for which the coefficients are almost null. On the contrary, the highest level of correlation is registered between "*Work and life balance*" and "*Quality of services*" (0.8082).

**Table 3 –** *Correlation among composite indicators of each domain for GAMPI approach*

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **D1** | 1 | | | | | | | | | |
| **D2** | 0.395 | 1 | | | | | | | | |
| **D3** | 0.6473 | 0.5065 | 1 | | | | | | | |
| **D4** | 0.3313 | 0.1493 | 0.403 | 1 | | | | | | |
| **D5** | 0.511 | 0.3139 | 0.6738 | 0.4584 | 1 | | | | | |
| **D6** | 0.5425 | 0.2837 | 0.5337 | 0.3647 | 0.3039 | 1 | | | | |
| **D7** | 0.3271 | 0.2207 | 0.3889 | 0.2673 | 0.47 | 0.1694 | 1 | | | |
| **D8** | 0.0019 | -0.0936 | -0.068 | 0.1726 | 0.3744 | -0.1297 | 0.4061 | 1 | | |
| **D9** | 0.5716 | 0.4933 | 0.7173 | 0.1319 | 0.3559 | 0.5044 | 0.1298 | -0.285 | 1 | |
| **D10** | 0.628 | 0.3414 | 0.8082 | 0.3511 | 0.5537 | 0.4702 | 0.2039 | -0.2015 | 0.6934 | 1 |

The fact that none of the values of the correlation coefficient is particularly high may, interestingly, reveal that the BES domains are not substitutable but rather complementary. Hence, a possible further step of aggregating all the domain-specific composite indices into an overall composite index of well-being may reveal to be a dangerous choice, since it would lose important information.

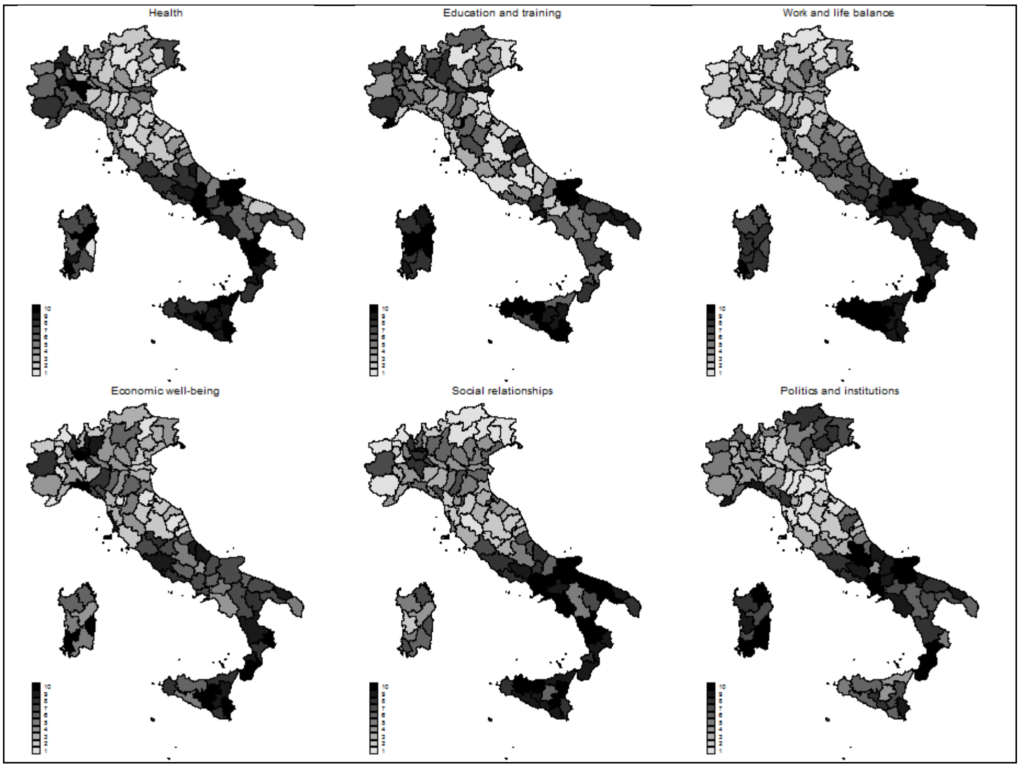
## 4. Conclusive remarks

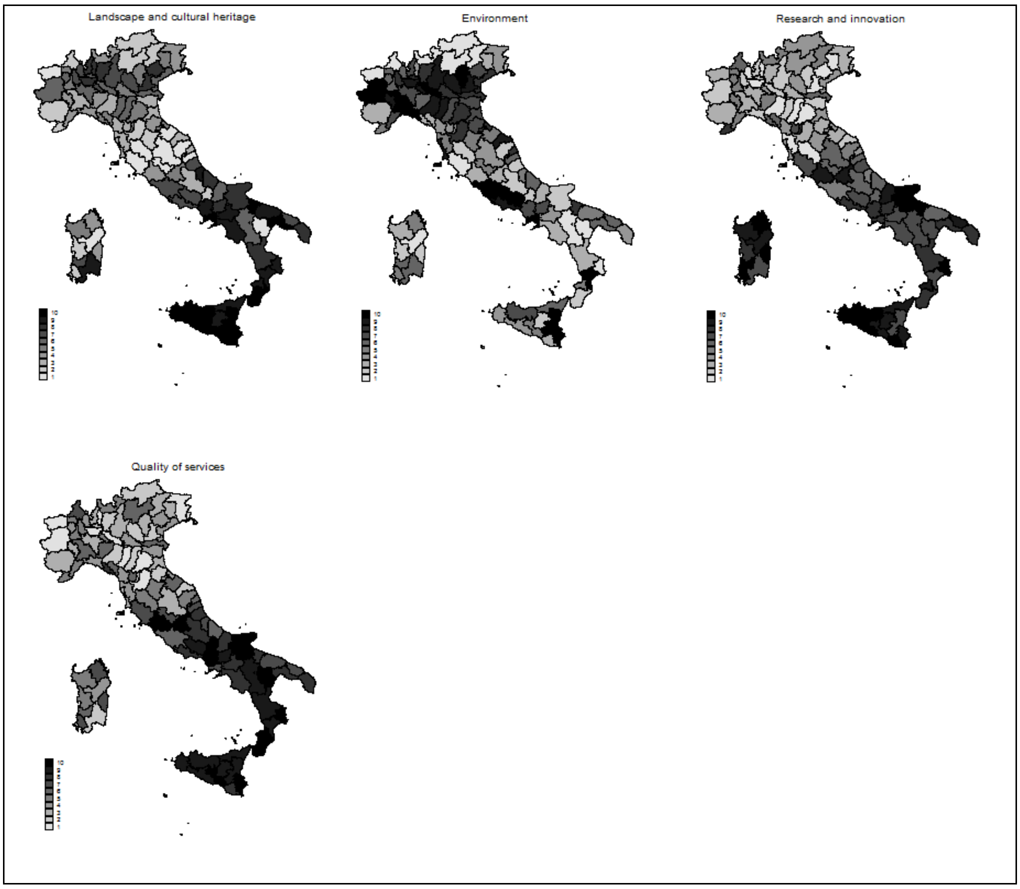
There is an increasing interest on the aggregation of elementary indicators. Beside the simple arithmetic mean, there are several methods to aggregate information. However, there is no better method; each synthetic method reflects a different nature of the composite indicator that emphasizes a different priority in defining the well-being (Wilson et al. 2007).

Here, we explore different methods of aggregation for the elementary indicators constituting the Italian Equitable and Sustainable Well-being Project for Italian Provinces. And we propose a new method, called *Gini-based adjusted Mazziotta-Pareto Index* (GAMPI) that accounts both for vertical and horizontal variability.

Further researches will be conducted to analyze and compare additional aggregation methods.

**Figure 1 –** *Maps of Italian provinces (ranking by deciles) for GAMPI in each domain*.





## References

BECCHETTI L., MASSARI R., NATICCHIONI P. 2014. The drivers of happiness inequality: suggestions for promoting social cohesion. Oxford Economic Papers, Vol. 66, No. 2, pp. 419–442

MAZZIOTTA M., PARETO A. 2016. On a Generalized Non-compensatory Composite Index for Measuring Socio-economic Phenomena, *Social Indicators Research*, Vol. 127, No. 3, pp. 983-1003.

MAZZIOTTA M., PARETO A. 2013a. A Non-Compensatory Composite Index for Measuring Well-Being over Time. *Cogito. Multidisciplinary Research Journal*, Vol.5, No. 4, pp 93-104.

MAZZIOTTA M., PARETO A. 2013b. Methods for constructing composite indices: one for all or all for one. *Rivista Italiana di Economia Demografia e Statistica*, Vol. 67, No.2, pp. 67-80.

MASSOLI P., MAZZIOTTA M., PARETO A., RINALDELLI C. 2014. Indici compositi per il BES. *Giornate delle ricerca in Istat*. Istat.

ISTAT (2015), Il benessere equo e sostenibile in Italia, Istat, Roma.

STIGLITZ J. E., SEN A., FITOUSSI J. P. 2009. Report by the commission on the measurement of economic performance and social progress. Commission on the Measurement of Economic Performance and Social Progress, Paris.

CHELLI F. M., CIOMMI M., EMILI A., GIGLIARANO C., TARALLI S. 2015. Comparing Equitable And Sustainable Well-Being (Bes) Across The Italian Provinces. A Factor Analysis-Based Approach. *Rivista Italiana di Economia Demografia e Statistica*, Vol. 69, No. 3, pp. 61-72.

WILSON J., TYEDMERS P., PELOT R. 2007. Contrasting and comparing sustainable development indicator metrics. *Ecological indicators*, Vol.7, No. 2, pp.299-314.

## SUMMARY

### Measuring local well-being: a comparison among aggregative methods for the equitable and sustainable well-being

Within the "BES' Provinces" Project this work aims to compare different synthesis techniques of elementary indicators for each BES domain. Firstly, 41 elementary indicators are selected from the original dataset of 88 indicators, available for the year 2014, which guarantee robustness, reliability and relevance in accordance with the BES meaning.

Motivated by the debate on the need to summarize the information arising from a large set of variables, in this paper we discuss three different aggregation methods: the Adjusted Mazziotta-Pareto Index, the arithmetic mean weighted by the Gini coefficients of the elementary indicators and a mixed approach based on the two.

Francesco M. CHELLI, Università Politecnica delle Marche, f.chelli@univpm.it
Mariateresa CIOMMI, Università Politecnica delle Marche, m.ciommi@univpm.it
Alessandra EMILI, Università Politecnica delle Marche, a.emili@staff.univpm.it
Chiara GIGLIARANO, Università dell'Insubria, chiara.gigliarano@uninsubria.it
Stefania TARALLI, Istat, taralli@istat.it

# UN'ANALISI DELLA CORRELAZIONE DI INDICI COMPOSITI DI BENESSERE IN ITALIA[1]

Matteo Mazziotta e Adriano Pareto

## 1. Introduzione

Da diversi anni la discussione in merito al ruolo del PIL rispetto alla misurazione del benessere e della qualità della vita dei cittadini è ampia, continua e coinvolge studiosi di discipline diverse a livello internazionale. Mentre nel passato il dibattito era focalizzato principalmente sui paesi in via di sviluppo e, comunque, relegato al mondo accademico, negli ultimi anni l'attenzione si è spostata verso i paesi ad alto reddito e ha coinvolto istituzioni nazionali e internazionali. Molti istituti di statistica, così come organizzazioni non governative, gruppi di riflessione e centri di ricerca hanno proposto nuovi indicatori che superano la visione tradizionale economicista di benessere. Si abbandona l'assunto, discusso in letteratura, che PIL e benessere siano correlati positivamente e, invece, aumenta la convinzione che all'aumentare dell'uno può corrispondere una contrazione dell'altro. O che, ancora meglio, i due spieghino aspetti diversi della realtà socio-economica di un'area geografica o di una specifica sotto-popolazione.

La pubblicazione da parte dell'Istat, nel dicembre del 2015, del terzo rapporto sul Benessere Equo e Sostenibile (BES) ha segnato un precedente unico nella statistica ufficiale internazionale poiché sono state utilizzate metodologie, note in letteratura come indici compositi (OECD, 2008), per sintetizzare gli indicatori elementari di ciascun dominio di *outcome* (Mazziotta e Pareto, 2013; 2015). Si è cercato, in sostanza, di misurare il benessere come fenomeno multidimensionale e come interpretazione di un fattore latente rendendolo, quindi, unidimensionale e visibile. La risoluzione della complessità costituita da un "classico" *dashboard* apre la strada a nuove analisi statistiche per confrontare tra loro gli indici compositi di ciascun dominio comprendendone i reciproci legami nonché le relazioni tra tali indici compositi e il PIL stesso. L'obiettivo del *paper* è proprio analizzare i legami d'influenza reciproca tra gli indici compositi dei domini e, soprattutto, analizzare

[1] Il lavoro è frutto dell'opera di entrambi gli autori. In particolare, i paragrafi 1 e 4 vanno attribuiti a Matteo Mazziotta e i paragrafi 2 e 3 vanno attribuiti ad Adriano Pareto.

quanto il PIL non riesca a spiegare del fattore latente benessere, basandosi non sulle teorie economiche ma utilizzando modelli statistici: in tal modo è possibile misurare e quantificare proprio la percentuale di tale divergenza.

Nel paragrafo 2, viene presentata la matrice delle correlazioni tra gli indici compositi dei domini del BES; mentre, nel paragrafo 3, si evidenzia che i fattori generati dall'Analisi in Componenti Principali, applicata agli indici compositi del BES, una volta messi in relazione con PIL, mostrano la sua parziale capacità informativa del benessere.

## 2. Gli indici compositi

Gli indici compositi di benessere utilizzati in questo studio sono tratti dal Rapporto BES 2015 (Istat, 2015). In particolare, si tratta degli indici compositi delle 9 dimensioni del BES (Salute, Istruzione e formazione, Lavoro e conciliazione dei tempi di vita, Benessere economico, Relazioni sociali, Sicurezza, Benessere soggettivo, Paesaggio e patrimonio culturale, Ambiente), rilevati in Italia a livello regionale, a cui sono stati aggiunti alcuni indicatori complementari.

La lista degli indici considerati, con le rispettive sigle e gli anni di riferimento, è illustrata in tabella 1. Per una descrizione dettagliata degli indicatori, si rimanda al volume dell'Istat (2015).

**Tabella 1 –** *Indici compositi di benessere in Italia (fonte: Istat, 2015)*

| Nome | Descrizione | Anno |
|---|---|---|
| SAL | Indice composito di Salute | 2013 |
| IST | Indice composito di Istruzione e formazione | 2014 |
| LAV | Indice composito di Qualità e soddisfazione del lavoro | 2014 |
| OCC | Tasso di Occupazione standardizzato | 2014 |
| RED | Indice composito di Reddito e disuguaglianza | 2014 |
| DIS | Indice composito di Disagio economico | 2014 |
| REL | Indice composito di Relazioni sociali | 2014 |
| SIC | Indice composito di Sicurezza | 2014 |
| OMI | Tasso standardizzato di Omicidi | 2014 |
| SOG | Indice di Soddisfazione per la propria vita | 2014 |
| PAE | Indice composito di Paesaggio | 2011 |
| AMB | Indice composito di Ambiente | 2012 |

Nella tabella 2 è riportata la matrice di correlazione dei 12 indici compositi e la correlazione di questi ultimi con il PIL del 2014.

Come si può notare, la maggior parte degli indici compositi sono correlati positivamente tra loro (SAL, IST, LAV, OCC, RED, DIS, REL, SOG e PAE), con valori particolarmente elevati (r ≥ 0,700). Anche l'indice composito ambientale (AMB) e il tasso standardizzato di omicidi (OMI) sono correlati positivamente con questo insieme di indicatori, ma il legame risulta più debole per AMP (0,700 ≥ r ≥ 0,450) e piuttosto basso per OMI (0,450 ≥ r ≥ 0,200).

L'indice composito di sicurezza (SIC), invece, mostra una lieve correlazione negativa con gli altri indici compositi (0,200 ≥ r ≥ -0,250).

**Tabella 2 –** *Correlazione tra indici compositi di benessere e PIL*

| Indice composito | SAL | IST | LAV | OCC | RED | DIS | REL | SIC | OMI | SOG | PAE | AMB |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| SAL | 1,000 | 0,842 | 0,911 | 0,917 | 0,906 | 0,876 | 0,902 | -0,232 | 0,457 | 0,871 | 0,803 | 0,559 |
| IST | 0,842 | 1,000 | 0,807 | 0,850 | 0,841 | 0,832 | 0,829 | -0,119 | 0,190 | 0,826 | 0,763 | 0,540 |
| LAV | 0,911 | 0,807 | 1,000 | 0,963 | 0,963 | 0,913 | 0,842 | -0,201 | 0,452 | 0,791 | 0,784 | 0,538 |
| OCC | 0,917 | 0,850 | 0,963 | 1,000 | 0,969 | 0,908 | 0,884 | -0,229 | 0,413 | 0,821 | 0,809 | 0,494 |
| RED | 0,906 | 0,841 | 0,963 | 0,969 | 1,000 | 0,916 | 0,887 | -0,172 | 0,452 | 0,858 | 0,799 | 0,555 |
| DIS | 0,876 | 0,832 | 0,913 | 0,908 | 0,916 | 1,000 | 0,845 | -0,185 | 0,404 | 0,785 | 0,700 | 0,478 |
| REL | 0,902 | 0,829 | 0,842 | 0,884 | 0,887 | 0,845 | 1,000 | -0,084 | 0,427 | 0,927 | 0,865 | 0,639 |
| SIC | -0,232 | -0,119 | -0,201 | -0,229 | -0,172 | -0,185 | -0,084 | 1,000 | -0,048 | 0,020 | -0,129 | 0,178 |
| OMI | 0,457 | 0,190 | 0,452 | 0,413 | 0,452 | 0,404 | 0,427 | -0,048 | 1,000 | 0,418 | 0,228 | 0,428 |
| SOG | 0,871 | 0,826 | 0,791 | 0,821 | 0,858 | 0,785 | 0,927 | 0,020 | 0,418 | 1,000 | 0,775 | 0,696 |
| PAE | 0,803 | 0,763 | 0,784 | 0,809 | 0,799 | 0,700 | 0,865 | -0,129 | 0,228 | 0,775 | 1,000 | 0,532 |
| AMB | 0,559 | 0,540 | 0,538 | 0,494 | 0,555 | 0,478 | 0,639 | 0,178 | 0,428 | 0,696 | 0,532 | 1,000 |
| **PIL** | **0,889** | **0,748** | **0,889** | **0,928** | **0,899** | **0,834** | **0,873** | **-0,221** | **0,554** | **0,847** | **0,733** | **0,577** |

Per quanto riguarda le correlazioni dei 12 indici sintetici con il PIL, la concordanza maggiore si osserva con il tasso di occupazione standardizzato (OCC), seguito dall'indice composito di reddito e disuguaglianza (RED), l'indice composito di qualità e soddisfazione del lavoro (LAV) e l'indice composito di salute (SAL).

Gli indicatori meno concordanti con il PIL sono il tasso standardizzato di omicidi (OMI), con r = 0,554, e l'indice composito ambientale (AMB), con r = 0,577; mentre l'indice più discordante è il composito di sicurezza (SIC), che presenta una correlazione negativa con il PIL (r = -0,221).

Tali risultati confermano che se, da una parte, i principali indici sintetici di benessere possono essere ‘spiegati’ dal PIL, alcuni di essi, come quelli relativi alla sicurezza e all’ambiente, sono quasi del tutto ‘slegati’ da tale indicatore.

## 3. L’Analisi in Componenti Principali

I risultati del paragrafo precedente suggeriscono l’applicazione di un’Analisi in Componenti Principali (ACP) della matrice dei 12 indici compositi considerati[2].

Com’è noto, l’ACP è una tecnica di analisi multivariata che, partendo da un insieme di indicatori originari, consente di ottenere dei nuovi indicatori (le componenti principali o fattori), di importanza decrescente e ortogonali tra loro, combinazioni lineari degli indicatori di partenza. Ciò consente di descrivere le unità statistiche con un minor numero di nuovi indicatori, massimizzando la proporzione di ‘varianza spiegata’ (Dunteman, 1989).

Nella figura 1 sono illustrati lo *scree-plot* e il cerchio delle correlazioni dell’ACP.

Dall’esame dello *scree-plot*, è evidente un ‘gomito’ in corrispondenza del secondo fattore; ciò significa che la maggior parte della variabilità delle regioni italiane (80,77%) può essere spiegata dai primi due fattori. Il terzo fattore spiega il 7,63% della restante varianza, ma avendo un autovalore inferiore a 1 ($\lambda = 0,914$) potrebbe risultare non significativo.

Proiettando le variabili originarie nel piano delle prime due componenti principali si ottiene il cerchio delle correlazioni, dove ciascun indice composito è rappresentato da un punto di coordinate uguali ai due coefficienti di correlazione con il primo e il secondo fattore. Si osservi che il primo fattore risulta fortemente correlato con 9 indici compositi su 12 (SAL, IST, LAV, OCC, RED, DIS, REL, SOG e PAE), mentre il secondo rappresenta soltanto l’indice composito di sicurezza (SIC). Infine, il tasso standardizzato di omicidi (OMI) e l’indice composito ambientale (AMB), vanno a collocarsi in una posizione intermedia tra i due assi, risultando parzialmente correlati con entrambi i fattori.

Nella figura 2 sono riportate le rappresentazioni grafiche delle relazioni tra il PIL e primi due fattori dell’ACP.

La correlazione tra il PIL e il primo fattore risulta elevatissima ($r = 0,9213$), a conferma del fatto che la maggior parte dell’informazione relativa al benessere delle regioni può essere desunta dal PIL. E’ interessante notare, tuttavia, che il primo fattore spiega circa il 70% della varianza complessiva. Ne consegue che il

[2] L’analisi è stata condotta su variabili standardizzate (media nulla e varianza unitaria). In tal modo, la matrice di varianze e covarianze coincide con la matrice di correlazione.

PIL non riesce a ‘catturare’ il restante 30% di informazione. Infatti, il secondo fattore dell’ACP, che rappresenta la sicurezza (SIC) e, in parte, l’ambiente (AMB) risulta del tutto incorrelato con il PIL (r = 0,0446).

**Figura 1 –** *Scree-plot e cerchio delle correlazioni dell’ACP*

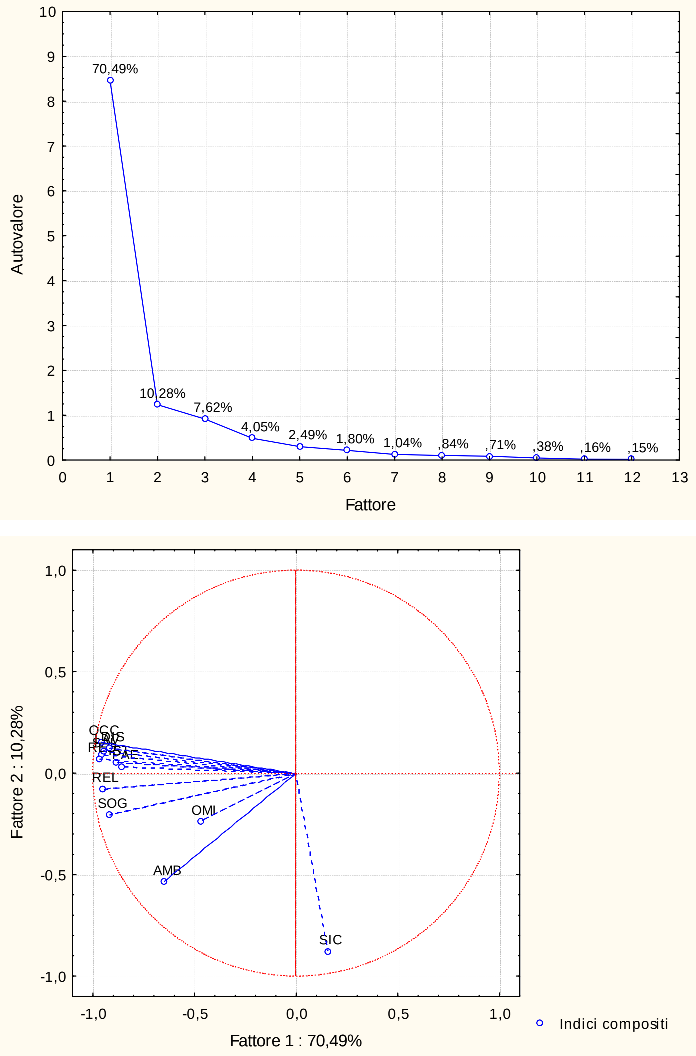
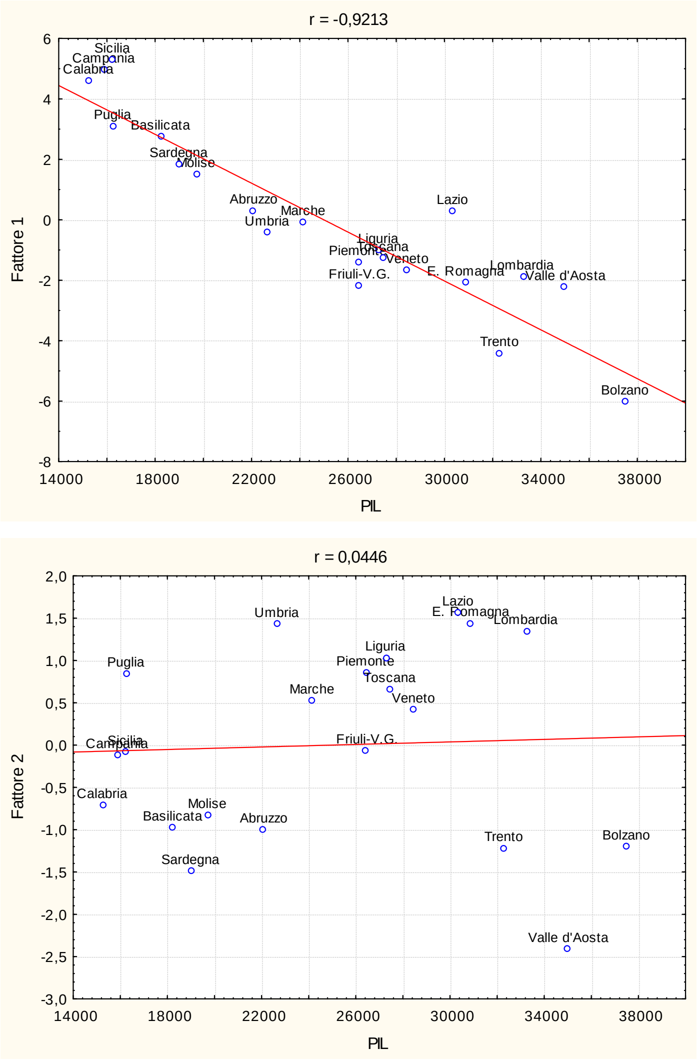


**Figura 2 –** *Relazioni tra i primi due fattori dell'ACP e il PIL*



## 4. Conclusioni

Diversi approcci socio-economici, nel corso degli anni, hanno sostenuto l'inefficacia del PIL come misura del benessere trovando nella multidimensionalità la risposta più convincente dal punto di vista teorico (Rinaldi e Zelli, 2014). La pubblicazione degli indici sintetici del BES a livello regionale è sembrata un'occasione istituzionalmente e metodologicamente valida per provare a quantificare quanto (dal punto di vista numerico) il PIL non possa spiegare il benessere come fattore latente. L'analisi in componenti principali ha dimostrato che, a livello regionale, tale quota è pari a circa il 30%; sembra necessario sottolineare che, dal punto di vista strettamente metodologico, passando a un dettaglio territoriale maggiore (per es., a livello provinciale), la varianza non 'spiegata' dal PIL potrebbe essere ancora più alta del 30%: a tal riguardo, sono in corso studi degli autori in cui è stato considerato il PIL calcolato dall'Istituto Tagliacarne che confermano tale tendenza. All'approccio teorico si accompagna quello metodologico in cui, grazie a modelli statistici, è possibile quantificare il disallineamento (non completo) tra l'approccio multidimensionale del benessere e quello unidimensionale del PIL.

Ovviamente, tale convinzione non deve essere un punto di arrivo, ma un punto di partenza per proseguire l'attività di definizione del benessere dal punto di vista teorico e di misurazione delle diverse componenti (dimensioni) che meglio possano rappresentarlo. I recenti studi fatti in via sperimentale per la costruzione di indicatori di benessere a livello comunale, partendo da archivi amministrativi raccolti in sistemi informativi, stanno aprendo la strada a un filone di ricerca che sembra essere particolarmente apprezzato dalle istituzioni politiche (nazionali e locali); la possibilità di misurare performance sociali ed economiche ad un livello di dettaglio così disaggregato costituisce un ausilio fondamentale per il *policy maker* che vuole indirizzare in modo più efficace gli interventi/azioni sul territorio.

La metodologia statistica e la statistica ufficiale, più che in altre circostanze, si mettono al servizio delle comunità per misurare al fine di agire e, quindi, migliorare il benessere dei cittadini.

## Riferimenti bibliografici

DUNTEMAN G. H., 1989. Principal Components Analysis. Newbury Park: Sage Publications.

ISTAT, 2015. Rapporto Bes 2015: il benessere equo e sostenibile in Italia, http://www.istat.it/it/files/2015/12/Rapporto_BES_2015.pdf

MAZZIOTTA M., PARETO A., 2013. Methods for Constructing Composite Indices: One for All or All for One?, Rivista italiana di economia, demografia e statistica, LXVII, 2, pp. 67-80.

MAZZIOTTA M., PARETO A., 2015. On a Generalized Non-compensatory Composite Index for Measuring Socio-economic Phenomena. Social Indicators Research, doi: DOI 10.1007/s11205-015-0998-2, Springer.

OECD, 2008. Handbook on Constructing Composite Indicators. Methodology and user guide. Paris: OECD Publications.

RINALDI A., ZELLI R., 2014. Misurare il benessere. La sfida degli indicatori alternativi al Pil. Roma: Donzelli Editore.

## SUMMARY

### A Correlation Analysis of Italian Well-being Composite Indicators

In December 2015, the Italian National Institute of Statistics (Istat) published the third edition of the Italian Equitable and Sustainable Well-being (BES) report. The big innovation, compared to past editions, is the calculation of composite indicators for each of the eight outcomes domains (health, education, labour, economic well-being, social relations, security, landscape and cultural heritage, environment) at regional level and over time. This result was obtained after 3 years of experimentation in which many methods have been tested in order to respect the fundamental constraints for a publication of official statistics and to obtain the most robust results from a statistical point of view. Substantially, the adopted approach has no equal in the world literature of official statistics and puts Italy at the forefront on the study of these topics.

The aim of this paper is to 'compare' the composite indicators of each well-being domain with individual indicators well-known in literature and created to represent the phenomenon as a unidimensional measure. This comparison is done through a correlation analysis and graphical representations. One of the main objectives is to understand whether the latent structure represented by the composite indicators is different (and how) than the GDP, the classic one-dimensional measure used for socio-economic analysis.

Matteo MAZZIOTTA, Istat, mazziott@istat.it
Adriano PARETO, Istat, pareto@istat.it

# INDAGINE SPERIMENTALE SULLA MOBILITA' TRANSNAZIONALE PER APPRENDIMENTO: UNA PRIMA VALUTAZIONE DEI RISULTATI[1]

Raffaella Cascioli, Liana Verzicco

## 1. Introduzione

L'Unione Europea investe da tempo sulla mobilità transnazionale con finalità di apprendimento, intesa come elemento utile alla crescita dei paesi dell'Unione e come strumento chiave per dare ai giovani maggiori opportunità di studio e di formazione all'estero volte a sviluppare l'acquisizione o il miglioramento di competenze professionali, linguistiche, interpersonali e interculturali. L'importanza di questo obiettivo è dimostrata dai crescenti finanziamenti destinati alla mobilità nell'ambito del programma comunitario Socrates (che comprende, tra gli altri, Erasmus e Comenius). Nel quadro della Strategia Europea relativo all'Istruzione e alla Formazione (Education and Training 2020) tra i parametri di riferimento (*benchmark*) da raggiungere entro il 2020 sono stati inseriti due indicatori relativi i) alla mobilità per apprendimento nei percorsi di istruzione e formazione professionale iniziale (*Learning mobility in initial vocational education and training (IVET)*) e ii) alla mobilità effettuata nei corsi dell'istruzione terziaria (*Learning mobility in Higher education*). Per quanto concerne il primo *benchmark*, l'obiettivo da raggiungere è che almeno il 6% dei 18-34enni che ha conseguito un diploma o una qualifica professionale deve aver trascorso, durante gli studi, un periodo di studio o formazione di almeno due settimane all'estero (legato al programma di istruzione o di formazione professionale) Per quanto riguarda il secondo indicatore, entro il 2020 almeno il 20% dei 18-34enni con un titolo terziario deve aver svolto un periodo di studio o formazione all'estero (legato al programma di livello terziario frequentato) di almeno tre mesi.

Ad oggi, Eurostat non è in grado di produrre questi indicatori utilizzando le fonti amministrative che alimentano la raccolta dati UOE. Le uniche statistiche disponibili, infatti, sono relative alla mobilità degli studenti dell'istruzione terziaria nell'Unione europea (Ue) che studiano in uno Stato membro dell'Unione ma

[1] R. Cascioli è autrice dei paragrafi 2,3,4; L.Verzicco è autrice dei paragrafi 1 e 4.
Alla costruzione del modello logit hanno collaborato entrambe le due autrici.

hanno conseguito il diploma secondario in un altro paese (indipendentemente dal fatto che questo paese sia o no membro dell'Unione europea).

Al fine di acquisire un primo set comparabile di dati utili al monitoraggio e all'analisi dei progressi verso gli obiettivi prefissati, Eurostat ha promosso, in accordo con diversi Paesi dell'Unione europea, una raccolta dati pilota sulle esperienze di *learning mobility*. Il principale obiettivo dell'indagine era quello di ottenere informazioni quantitative e qualitative sulla mobilità per apprendimento nei percorsi di istruzione/formazione professionale iniziale (*IVET mobility*), sulla mobilità a livello di istruzione terziaria e sulla mobilità più in generale che include ogni forma di apprendimento, sia in un contesto formale sia in altro contesto. Questa sperimentazione ha consentito, inoltre, di valutare gli indicatori utili al monitoraggio della mobilità per apprendimento.

L'Italia è tra i 16 paesi europei che hanno partecipato a questa Indagine Pilota e le informazioni sul fenomeno sono state raccolte dall'ISTAT attraverso un modulo aggiuntivo inserito nella Rilevazione continua sulle forze di lavoro.

Il presente lavoro si propone di descrivere l'indagine svolta da ISTAT e presentarne i principali risultati. L'obiettivo è quello di fornire informazioni di dettaglio su un fenomeno in crescita e presentare un primo quadro generale del fenomeno, della sua diffusione e delle sue caratteristiche.

## 2. Metodologia dell'indagine

L'indagine Sperimentale sulla mobilità transnazionale per apprendimento si è svolta nel quarto trimestre del 2014 all'interno della Rilevazione sulle forze di lavoro. Il target di riferimento sono stati i giovani di età compresa tra i 18 ed i 34 anni. La numerosità campionaria, pari a circa 23mila unità, ha rappresentato una popolazione totale (18-34 anni) pari a 11milioni e 128mila giovani. La rilevazione è stata condotta con tecnica mista: le prime interviste sono state realizzate con tecnica CAPI (*computer assisted personal interviews*) e le successive con CATI (*computer assisted telephone interviews*); rispettivamente il 51,6% e il 48,4% del campione.

Nel lavoro vengono presentati i principali risultati ottenuti, sia in termini di descrizione del fenomeno in oggetto (le diverse tipologie di percorsi di studio all'interno dei quali è avvenuta l'esperienza della mobilità), sia descrivendo le principali caratteristiche dei partecipanti ( genere, territorio, ecc). Verranno infine presentati i risultati dell'applicazione di modelli multivariati volti ad approfondire quanto osservato nell'analisi univariata.

## 3. Principali risultati

Diversi sono stati gli indicatori di *learning mobility* ricavabili dall'indagine che sono stati oggetto di analisi. Nel dettaglio:

1) Mobilità di chi ha un titolo di livello ISCED 3: la quota di 18-34enni in possesso di un titolo secondario superiore che sono stati all'estero per almeno due settimane per attività di studio o formazione legate al percorso di studio.

2) Mobilità di chi ha un titolo di livello ISCED 3 ad indirizzo "*vocational*" (*IVET benchmark*): la quota di 18-34enni in possesso di un titolo secondario superiore ad indirizzo "tecnico/professionale" che sono stati all'estero per almeno due settimane per attività di studio o formazione legate al percorso di studio.

3) Mobilità di chi ha un titolo di livello ISCED 3 di tipo "*general*": la quota di 18-34enni in possesso di un titolo secondario superiore ad indirizzo "generale" che sono stati all'estero per almeno due settimane per attività di studio o formazione legate al percorso di studio.

4) Mobilità di chi ha conseguito un titolo terziario (livello ISCED 5-8) (*Tertiary benchmark*): la quota di 18-34enni in possesso di un titolo terziario che sono stati all'estero per attività di studio o formazione per almeno tre mesi come parte del percorso di istruzione terziario.

5) Mobilità per apprendimento "non-formale": la quota di 18-34enni che hanno avuto un'esperienza all'estero - indipendente dalla durata - per motivi di apprendimento fuori dai percorsi di istruzione "formale"[2].

6) Almeno un tipo di mobilità: la quota di 18-34enni che hanno sperimentato almeno un tipo di mobilità transnazionale per apprendimento, sia all'interno dei percorsi di istruzione/formazione sia in altri ambiti e per altre ragioni.

### *3.1. Le caratteristiche socio-demografiche dei giovani con esperienze di learning mobility*

La quota di 18-34enni in possesso di un titolo secondario superiore che sono stati all'estero per almeno due settimane per attività di studio o formazione legate al percorso scolastico è risultata pari al 6,2% (Tabella 1). Le più "mobili" durante gli studi secondari superiori di II grado sono le donne (7,4% vs 5,1% dei ragazzi).

[2] L'istruzione "formale", individuata in base ai parametri della Classificazione ISCED 2011, comprende l'istruzione scolastica, universitaria e accademica, e anche i percorsi IFP, IFTS, ITS o di formazione professionale regionale iniziale di I e II livello. L'istruzione "non formale" comprende le attività formative organizzate da vari soggetti al di fuori del sistema di istruzione e formazione.

Si rileva una minore diffusione della mobilità nel Mezzogiorno, dove l’incidenza del fenomeno è pari al 4,8% contro il 6,2% nel Nord e il 7,4% nel Centro. La *learning mobility* nei percorsi di scuola secondaria superiore è un fenomeno recente: è infatti più diffuso tra le coorti più giovani (18-24enni) dove l’incidenza raggiunge l’8,8%. Queste esperienze risultano meno frequenti, invece, nelle coorti che hanno conseguito il diploma da più di 5 anni (circa la metà rispetto a chi ha concluso gli studi secondari superiori da meno di 5 anni).

**Tabella 1** – *Giovani 18-34 anni con un'esperienza di mobilità transnazionale per apprendimento per tipo di mobilità, sesso, ripartizione geografica, classe di età, distanza dal conseguimento del titolo di studio, classe di laurea, indirizzo di laurea, livello di istruzione, permanenza in istruzione e titolo di studio dei genitori (a)- Anno 2014 (incidenze percentuali)*

| | **Mobilità dei diplomati e laureati durante gli studi (b)** | | | | **Mobilità per apprendimento "non formale"** | **Almeno un tipo di mobilità** |
|---|---|---|---|---|---|---|
| | Secondari superiori | Secondari superiori "vocational" (*IVET Benchmark* ) | Secondari superiori "general" | Terziari (*Tertiary Benchmark* ) | | |
| SESSO | | | | | | |
| Maschi | 5,1 | 3,5 | 10,2 | 8,7 | 4,5 | 9,6 |
| Femmine | 7,4 | 5,1 | 10,4 | 9,0 | 5,6 | 13,8 |
| RIPARTIZIONE GEOGRAFICA | | | | | | |
| Nord | 6,2 | 4,7 | 11,3 | 9,4 | 6,3 | 14,1 |
| Centro | 7,4 | 4,5 | 13,0 | 8,9 | 6,3 | 14,0 |
| Mezzogiorno | 4,8 | 3,3 | 8,2 | 8,0 | 3,1 | 7,8 |
| CLASSE DI ETA' | | | | | | |
| 18-24 | 8,8 | 5,6 | 12,7 | 10,3 | 5,6 | 13,6 |
| 25-34 | 4,3 | 3,5 | 7,1 | 8,7 | 4,7 | 10,5 |
| DISTANZA DAL CONSEGUIMENTO DEL TITOLO DI STUDIO | | | | | | |
| Minore o uguale a cinque anni | 8,5 | 5,7 | 12,7 | 9,4 | 7,9 | 18,4 |
| Oltre 5 anni | 4,2 | 3,2 | 7,3 | 7,7 | 3,0 | 6,8 |
| CLASSE DI LAUREA | | | | | | |
| Laurea magistrale | _ | _ | _ | 10,7 | _ | _ |
| Laurea di primo livello | _ | _ | _ | 6,5 | _ | _ |
| INDIRIZZO DI LAUREA | | | | | | |
| Lauree MST | _ | _ | _ | 9,5 | _ | _ |
| Laure non MST | _ | _ | _ | 8,7 | _ | _ |
| LIVELLO DI ISTRUZIONE | | | | | | |
| Teriario | _ | _ | _ | _ | 11,3 | 26,2 |
| Secondario superiore | _ | _ | _ | _ | 4,7 | 10,9 |
| Al più secondario inferiore | _ | _ | _ | _ | 2,0 | 4,1 |
| ANCORA/NON PIU' IN ISTRUZIONE | | | | | | |
| Ancora in istruzione | 12,0 | 8,8 | 13,1 | 11,1 | 8,5 | 21,3 |
| Non più in istruzione | 4,1 | 3,5 | 6,5 | 8,3 | 3,9 | 8,4 |
| TITOLO DI STUDIO DEI GENITORI (c) | | | | | | |
| Teriario | 14,2 | 9,0 | 16,2 | 13,5 | 11,8 | 29,6 |
| Secondario superiore | 7,4 | 5,1 | 10,8 | 9,5 | 6,4 | 15,2 |
| Al più secondario inferiore | 4,1 | 3,1 | 7,8 | 8,8 | 2,6 | 6,0 |
| **Totale** | **6,2** | **4,2** | **10,4** | **8,9** | **5,1** | **11,6** |
| **Totale popolazione target** *in migliaia (d)* | **6.633** | **4.629** | **2.083** | **1.916** | **11.083** | **11.128** |

*Fonte: ISTAT - Rilevazione sulle forze di lavoro*

*(a) titolo di studio più elevato tra quello del padre e della madre*

*(b) mobilità avvenuta all'interno del percorso di studio relativo al titolo*

*(c) variabile raccolta sul solo sottoinsieme dei giovani che vivono ancora nella famiglia di origine: il 65% circa dell'insieme dei 18-34enni*

*(d) esclusi i missing*

L’incidenza dei percorsi di mobilità appare decisamente più elevata tra coloro che hanno completato un ciclo di studi di tipo liceale (10,4%) rispetto a coloro che hanno seguito un indirizzo tecnico/professionale (4,2%). Anche nei collettivi differenziati per indirizzo di studio trova conferma quanto rilevato per l’insieme dei giovani diplomati, relativamente al divario territoriale e alla crescente diffusione del fenomeno nel tempo. La differenza tra i generi resta invece evidente solo per chi ha fatto studi ad indirizzo tecnico/professionale, dove la mobilità risulta pari al 3,5% tra i maschi e al 5,1% tra le femmine.

La correlazione tra l’aver avuto esperienze di mobilità all’estero per ragioni di apprendimento durante la scuola secondaria superiore e il proseguimento degli studi all’università evidenzia come *l’attachment* agli studi sia uno dei fattori che influenzano maggiormente la scelta di fare esperienze di *learning mobility*.. L’incidenza di mobilità tra i diplomati che si sono poi iscritti all’università è infatti pari al 12,0% mentre è solo del 4,1% tra chi dopo il conseguimento del diploma ha interrotto gli studi.

Rispetto a quanto rilevato negli studi secondari superiori, le esperienze di mobilità durante il percorso di studi terziari sono più numerose e frequenti. Anche prendendo in considerazione le sole esperienze di studio o formazione di durata pari ad almeno tre mesi (il *benchmark* europeo), la quota di 18-34enni che sono stati all’estero durante gli studi di livello terziario risulta pari all’8,9%. Non sembra esserci una significativa differenza di genere e le distanze tra Centro-Nord e Mezzogiorno appaiono più attenuate, mentre resta evidente l’incidenza più alta tra chi ha conseguito il titolo più recentemente. Nei percorsi di istruzione terziaria le esperienze di *learning mobility* vengono effettuate prevalentemente durante la frequenza di un corso di laurea magistrale (o post laurea): 10,7% rispetto alla mobilità per motivi di studio fatta durante i corsi di laurea di primo livello (6,5%).

Non si registrano significative differenze rispetto agli indirizzi disciplinari dei corsi di studio. Distinguendo tra i corsi di laurea MTS (o STEM), vale a dire le lauree in materie scientifiche, tecnologiche e matematiche, e il resto delle altre discipline, le esperienze di *learning mobility* appaiono solo leggermente più diffuse nel primo gruppo. Infine, anche a livello terziario, la variabile “ancora/non più in istruzione” mette in evidenza il legame tra l’interesse ad investire in istruzione/formazione e l’interesse verso le esperienze di *learning mobility*. Tuttavia, nel caso della mobilità a livello terziario, la variabile “ancora/non più in istruzione” deve essere interpretata anche considerando il fatto che, le esperienze di *learning mobility* sono in generale poco diffuse durante i corsi di primo livello mentre sono situazioni molto frequenti durante i corsi magistrali di secondo livello. Ne consegue che la maggiore incidenza di *lerning mobility* che si registra tra i giovani in possesso di un titolo terziario che sono ancora in istruzione deriva anche dal fatto che questi giovani sono prevalentemente coloro che dopo il conseguimento della laurea breve seguono un corso di laurea magistrale.

La quota di 18-34enni che hanno avuto un'esperienza all'estero per ragioni di apprendimento al di fuori dei percorsi di istruzione "formale" è pari al 5,1%. L'incidenza è maggiore tra le ragazze, tra i più giovani e nel Centro-Nord. Anche questo tipo di *learning mobility* è più diffuso tra coloro che hanno conseguito un titolo da meno di cinque anni, ad indicare che recarsi all'estero con l'obiettivo di acquisire formazione aggiuntiva a quella ricevuta nei programmi scolastici o universitari è un fenomeno che si sta diffondendo in questi ultimi anni e dunque tra le coorti più giovani. L'incidenza di tali esperienze è massima tra coloro che possiedono un titolo di studio terziario, raggiunge infatti l'11,3% contro il 4,7% di chi è in possesso di un titolo secondario superiore. Infine, anche la mobilità per apprendimento "non formale" appare maggiormente diffusa tra coloro che sono ancora inseriti in un percorso di istruzione formale.

Per quanto riguarda l'ultimo indicatore, quello di sintesi delle diverse tipologie di mobilità, si osserva che la quota di 18-34enni che hanno sperimentato almeno un tipo di mobilità transnazionale per apprendimento è pari all'11,6%. Risulta poco meno di uno su sette tra le femmine, nelle regioni del Centro-Nord e tra i più giovani, mentre è di poco inferiore al 20 per cento tra coloro che hanno conseguito un titolo più recentemente. E' interessante notare che, se si considera il complesso di esperienze di mobilità all'estero per ragioni di apprendimento - avvenute sia durante il percorso di istruzione formale sia fuori dal contesto scolastico/universitario - oltre un giovane su quattro, tra quelli in possesso di titolo terziario, risulta aver avuto almeno una esperienza all'estero, mentre sono più contenute le analoghe incidenze nei giovani in possesso di titoli di studio di livello inferiore.

Prendendo in esame le caratteristiche socioculturali delle famiglie di origine[3], emerge che le incidenze maggiori di esperienze di mobilità transnazionale per apprendimento si riscontrano laddove il livello d'istruzione dei genitori è più alto.

Le differenze maggiori si registrano nel ciclo di studi di scuola secondaria superiore, dove le esperienze di mobilità riguardano il 14,2% dei giovani con almeno un genitore in possesso di un titolo di studio terziario, il 7,4% di quelli con almeno un genitore in possesso del diploma e solo il 4% di quelli i cui genitori posseggono al più la licenza media.

Anche la mobilità per apprendimento al di fuori dal contesto dell'istruzione formale appare influenzata dal livello socioculturale della famiglia di origine: il

[3] In particolare si utilizza come *proxy* dello status socioculturale della famiglia l'informazione sul titolo di studio più alto tra quello del padre e della madre. Questa informazione è rilevata dall'Indagine sulle forze di lavoro solo per i giovani che vivono ancora nella famiglia di origine.

12% circa dei giovani che provengono da ambienti culturalmente più elevati ha maturato una esperienza all'estero a fini formativi "non formali", contro il 6,4% di coloro i cui genitori hanno un livello di istruzione medio ed appena il 2,6% per coloro i cui genitori hanno un basso livello di istruzione.

### *3.2. Un approfondimento sulle determinanti della "probabilità" di fare un'esperienza di learning mobility: un'analisi multivariata*

Per approfondire i risultati emersi dall'analisi descrittiva del fenomeno della *learning mobility* - al fine di definire meglio le caratteristiche individuali che maggiormente influenzano la probabilità di effettuare un'esperienza di mobilità transnazionale per apprendimento nei diversi contesti educativi - si è condotta un'analisi multivariata attraverso l'applicazione di modelli logistici.

Il modello è stato stimato per le cinque tipologie di *learning mobility* oggetto del presente studio, con la sola esclusione dell'indicatore di sintesi delle diverse forme di mobilità. In particolare, si è analizzata la mobilità a livello di studi secondari, distinguendo i giovani in possesso di un titolo di studio di tipo *vocational* da coloro in possesso di un titolo di studio *general*. Si è poi analizzata la mobilità dei giovani con un titolo di studio terziario, ed infine è stato sviluppato un modello relativo alla mobilità transnazionale con finalità di apprendimento "non formale". In ciascuno dei modelli la variabile risposta è rappresentata dalla probabilità di aver effettuato una esperienza transnazionale di *learning mobility*. La tabella 2 riporta gli *odds ratios* relativamente a tutte le variabili inserite nei modelli.

L'analisi multivariata conferma l'effetto discriminante del genere sulla probabilità che un 18-34enne in possesso di un titolo secondario superiore abbia trascorso un periodo all'estero di almeno due settimane per attività di studio o formazione legate al programma di studio seguito: la probabilità di avere tale tipo di esperienza si riduce, infatti, di circa un terzo nei giovani uomini rispetto alle donne. Anche l'area geografica di residenza ha un effetto significativo: nelle regioni del Centro-nord la probabilità di un'esperienza all'estero durante il percorso di studi secondari è superiore di circa il 50% rispetto al Mezzogiorno. Si conferma quanto evidenziato con l'analisi descrittiva per quanto riguarda la maggiore diffusione di tali esperienze tra chi ha conseguito il diploma negli anni più recenti. Infine, si dimostra significativo l'essere ancora in istruzione "formale" (in prevalenza la frequenza di percorsi universitari/accademici).

I risultati dei modelli multivariati applicati distintamente per i due percorsi di scuola secondaria, quello tecnico-professionale e quello liceale, non mostrano sostanziali differenze rispetto al modello applicato all'insieme dei giovani che hanno studiato nel complesso dei corsi secondari di II grado. L'unica differenza si registra tra i giovani che hanno avuto esperienze di mobilità durante gli studi liceali, dove la variabile "genere" non appare discriminante.

Nel ciclo di studi terziari, la scelta di intraprendere un'esperienza di studio all'estero è più frequente tra chi frequenta i corsi di laurea magistrale rispetto agli studenti dei corsi di laurea di primo livello. L'analisi multivariata conferma che la probabilità di *Learning mobility* tra chi ha conseguito una laurea magistrale rispetto ai laureati di primo livello è del 70% più elevata (*odds ratio* pari a 1,736). La probabilità è comunque maggiore per chi ha conseguito il titolo universitario negli anni più recenti, quando la partecipazione ad esperienze di studio all'estero ha iniziato a crescere.

Il risultato del modello logistico applicato alla mobilità per apprendimento di tipo "non formale" indica effetti significativi per diverse variabili esplicative. L'analisi multivariata non conferma il vantaggio, osservato nell'analisi descrittiva, delle donne sugli uomini, mentre avvalora l'effetto discriminante dell'area geografica di residenza. Per i giovani residenti nel Centro-nord la probabilità di maturare esperienze di mobilità all'estero fuori dai percorsi di studio formali è più che doppia rispetto ai giovani del Mezzogiorno. L'analisi conferma come anche questa forma di mobilità si sia diffusa negli anni più recenti. Resta discriminante il livello di istruzione raggiunto dal giovane, anche controllando per le altre caratteristiche, in particolare la classe di età. Infatti, rispetto a chi ha un titolo di studio secondario superiore la probabilità di recarsi all'estero per attività di apprendimento fuori dai percorsi scolastici/universitari è oltre due volte maggiore tra chi ha conseguito un titolo terziario e risulta dimezzata in chi ha al massimo un titolo di studio dell'obbligo. Infine, essere ancora inserito in un percorso di istruzione/formazione si conferma un fattore correlato positivamente alla mobilità all'estero svolta al di fuori dei percorsi di formazione " formali".

In tabella 3 sono presentati i risultati dell'analisi multivariata applicata sul collettivo di giovani 18-34enni che vivono ancora nella famiglia di origine. Questo approfondimento permette di introdurre nell'analisi la variabile che si riferisce al titolo di studio più alto tra quello del padre e della madre, *proxy* della condizione socioculturale in cui è cresciuto il giovane. I risultati del modello logit[4] non mostrano differenze rilevanti rispetto a quanto emerso dalle precedenti analisi logistiche, ma fanno invece emergere la forte influenza del background familiare sulle possibilità del giovane di maturare esperienze di mobilità all'estero a fini di apprendimento. Questo risultato è indipendente dal contesto in cui queste

[4] L'analisi logistica utilizzata in questo caso è con procedura *stepwise*. Data la ridotta numerosità del collettivo oggetto di questa analisi (solo il 65% circa dei giovani tra i 18 ed i 34 anni vive ancora nella famiglia di origine) si è ritenuto più robusta una presentazione dei soli risultati statisticamente significativi.

esperienze sono maturate e indipendente dalle principali caratteristiche possedute dal giovane, individuate come discriminanti nelle precedenti analisi.

**Tabella 2 –** *Giovani 18-34 anni: risultati dei modelli logit sulla probabilità di svolgere un'esperienza di mobilità transnazionale per apprendimento - Anno 2014 - (Odds ratios)*

| | **Mobilità dei diplomati e laureati durante gli studi (a)** | | | | **Mobilità per apprendimento "non formale"** |
|---|---|---|---|---|---|
| **Variabili esplicative** | Secondari superiori | Secondari superiori "vocational" (*IVET Benchmark* ) | Secondari superiori "general" | Terziari (*Tertiary Benchmark* ) | |
| SESSO | | | | | |
| Maschi | 0,719*** | 0,673*** | 0,917 | 0,962 | 0,959 |
| Femmine *(base)* | | | | | |
| RIPARTIZIONE GEOGRAFICA | | | | | |
| Centro-Nord | 1,495*** | 1,455*** | 1,567*** | 1,203 | 2,000*** |
| Mezzogiorno (base) | | | | | |
| CLASSE DI ETA' | | | | | |
| 18-24 | _ | _ | _ | _ | 0,93 |
| 25-34 *(base)* | | | | | |
| DISTANZA DAL CONSEGUIMENTO DEL TITOLO DI STUDIO | | | | | |
| Minore o uguale a 5 anni | 1,435*** | 1,546*** | 1,371*** | 1,318** | 1,606*** |
| Oltre 5 anni *(base)* | | | | | |
| CLASSE DI LAUREA | | | | | |
| Laurea magistrale | _ | _ | _ | 1,736*** | _ |
| Laurea di primo livello *(base)* | | | | | |
| INDIRIZZO DI LAUREA | | | | | |
| Lauree MST | _ | _ | _ | 1,030 | _ |
| Laure non MST *(base)* | | | | | |
| LIVELLO DI ISTRUZIONE | | | | | |
| Teriario | _ | _ | _ | _ | 2,309*** |
| Secondario superiore *(base)* | | | | | |
| Al più secondario inferiore | _ | _ | _ | _ | 0,467*** |
| ANCORA/NON PIU' IN ISTRUZIONE | | | | | |
| Ancora in istruzione | 2,804*** | 2,305*** | 1,917*** | _ | 2,053*** |
| Non più in istruzione *(base)* | | | | | |

*Fonte: ISTAT - Rilevazione sulle forze di lavoro*
*(a) mobilità avvenuta all'interno del percorso di studio relativo al titolo*
*(***)=p<0.01; (**)=p<0.05;(*)=p<0.1.*

**Tabella 3 –** *Giovani 18-34 anni che vivono ancora in famiglia: risultati dei modelli logit sulla probabilità di svolgere un'esperienza di mobilità transnazionale per apprendimento - Anno 2014 - (Odds ratios – Procedura stepwise)*

| | **Mobilità dei diplomati e laureati durante gli studi (a)** | | | | **Mobilità per apprendimento "non formale"** |
|---|---|---|---|---|---|
| **Variabili esplicative** | Secondari superiori | Secondari superiori "vocational" (*IVET Benchmark* ) | Secondari superiori "general" | Terziari (*Tertiary Benchmark* ) | |
| SESSO | | | | | |
| Maschi | 0,629*** | 0,513*** | n.s. | n.s. | 0,731*** |
| Femmine *(base)* | | | | | |
| RIPARTIZIONE GEOGRAFICA | | | | | |
| Centro-Nord | 1,276*** | n.s. | 1,493*** | n.s. | 1,836*** |
| Mezzogiorno (base) | | | | | |
| CLASSE DI ETA' | | | | | |
| 18-24 | _ | _ | _ | _ | n.s. |
| 25-34 *(base)* | | | | | |
| DISTANZA DAL CONSEGUIMENTO DEL TITOLO DI STUDIO | | | | | |
| Minore o uguale a 5 anni | 1,336*** | 1,477*** | n.s. | n.s. | 1,369*** |
| Oltre 5 anni *(base)* | | | | | |
| CLASSE DI LAUREA | | | | | |
| Laurea magistrale | _ | _ | _ | 1,762*** | _ |
| Laurea di primo livello *(base)* | | | | | |
| INDIRIZZO DI LAUREA | | | | | |
| Lauree MST | _ | _ | _ | n.s. | _ |
| Laure non MST *(base)* | | | | | |
| LIVELLO DI ISTRUZIONE | | | | | |
| Teriario | _ | _ | _ | _ | 1,862*** |
| Secondario superiore *(base)* | | | | | |
| Al più secondario inferiore | _ | _ | _ | _ | 0,629*** |
| ANCORA/NON PIU' IN ISTRUZIONE | | | | | |
| Ancora in istruzione | 2,247*** | 2,066*** | 1,597*** | n.s. | 1,550*** |
| Non più in istruzione | | | | | |
| TITOLO DI STUDIO DEI GENITORI (b) | | | | | |
| Teriario | 1,665*** | 1,626*** | 1,500*** | 1,412*** | 1,580*** |
| Secondario superiore *(base)* | | | | | |
| Al più secondario inferiore | 0,679*** | 0,656*** | 0,790*** | n.s. | 0,569*** |

*Fonte: ISTAT - Rilevazione sulle forze di lavoro*

*(a) mobilità avvenuta all'interno del percorso di studio relativo al titolo*

*(b) titolo di studio più elevato tra quello del padre e della madre*

*n.s.= variabile non presente nel modello stimato perché non soddisfa il livello di significatività dello 0,05.*

*(***)=p<0.01; (**)=p<0.05;*

## 4. Conclusioni

Ancora poco diffusa appare la mobilità transnazionale a fini di apprendimento nei percorsi secondari superiori, in particolare nei corsi ad indirizzo *vocational* rispetto a quelli di tipo *general*. Si tratta di una prima importante indicazione per le politiche europee, data l’attenzione che l'Ue pone proprio sulle esperienze di mobilità transnazionale nei percorsi *labour market oriented*. Le esperienze di mobilità verso i paesi esteri con finalità di apprendimento sono più frequenti durante gli studi terziari ma appaiono evidentemente maggiori tra chi studia nei corsi di laurea magistrale rispetto a chi frequenta i corsi di laurea di primo livello.

La partecipazione dei giovani ad esperienze di mobilità transnazionale a fini di apprendimento mostra un andamento crescente nel tempo. Si registrano differenze marcate tra le diverse aree geografiche del paese e rispetto alle diverse condizioni socioculturali della famiglia di origine.  La partecipazione alla *learning mobility* è meno diffusa tra i giovani residenti nel Mezzogiorno e tra coloro i cui genitori possiedono un livello basso di istruzione. Solo tra i giovani che hanno studiato all'università, le disuguaglianze derivanti dai diversi contesti territoriali di residenza scompaiono e quelle determinate dai diversi background socio-culturali si attenuano. Per quanto riguarda, invece, le esperienze di *learning mobility* svolte al di fuori dei percorsi di istruzione formale, la ripartizione territoriale di residenza sembra incidere fortemente sui livelli di partecipazione osservati.

## Riferimenti bibliografici

CONSIGLIO DELL'UNIONE EUROPEA Conclusioni su un criterio di riferimento nel settore della mobilità per l'apprendimento (2011/C 372/08).

EUROPEAN STATISTICAL SYSTEM, 2012. ESS agreement 2014: Pilot data collection on learning mobility via household surveys (both mobility in initial vocational education and training and general youth mobility).

EHEA Ministerial Conference, 2012, Mobility for Better Learning **-** Mobility strategy 2020 for the European Higher Education Area (EHEA).EUROSTAT, 2015, Learning mobility statistics: http://ec.europa.eu/eurostat/ statistics-explained/index.php?title=Learning_mobility_statistics&oldid=282885**.**

## SUMMARY

## The 2014 ISTAT Learning Mobility Survey: a first assessment of results.

The aim of this paper is to present the main results emerged from the analysis of the data from Learning Mobility Survey, the pilot survey conducted by ISTAT in 2014 on the young people aged 18-34.

The transnational mobility experiences in studying and training made within the learning programmes  increase with the level of education. Rather little in upper secondary school - and in particular in the vocational programmes - are more prevalent during the tertiary level courses, particularly in master degrees. Participation in transnational mobility for learning purposes shows an increasing trend over time and presents remarkable differences among geographical areas and respect to the different educational background of young's origin family. These experiences, in fact, are much less widespread among young people living in the South and among those whose parents have a low education level. The differences in participation in learning mobility experiences, instead, are reduced among young people who have studied at the university. The participation level in "learning mobility" realized outside formal education programmes is greatly different according to the geographical areas.

Raffaella CASCIOLI, ISTAT - Direzione Centrale per le Statistiche Sociali e il Censimento della Popolazione, racascio@istat.it

Liana VERZICCO, ISTAT - Direzione Centrale per le Statistiche Sociali e il Censimento della Popolazione, verzicco@istat.it

# HOW TO ACHIEVE CONSISTENCY BETWEEN MONTHLY AND QUARTERLY SEASONALLY ADJUSTED TIME SERIES?

Cinzia Graziani, Silvia Loriga, Michele Antonio Salvatore, Andrea Spizzichino

## 1. Introduction

In the Italian LFS an innovation has been recently introduced in the seasonal adjustment (SA) procedure, in particular as regards the methodology to guarantee consistency between monthly and quarterly SA time series.

The procedure previously used (until the release of December 2015 data) foresaw that, once seasonally adjusted, monthly time series were reconciled to the corresponding quarterly SA series. In that way the weighted average of monthly data (with weights equal to the number of weeks in each month, 4 or 5) was made equal to the corresponding quarterly figure. To ensure consistency between different aggregates and the total population (contemporary constraints) and between monthly and quarterly SA series (inter-temporal constraints) a reconciliation procedure in two steps was used[1].

This procedure was justified not only by the need to ensure consistency, but also because of the higher quality that quarterly SA series had compared with the monthly, at the moment when monthly data started to be produced by Istat (late 2009). Quarterly SA series were in fact available starting from the fourth quarter 1992, while the monthly since January 2004.

Nowadays, the length of monthly series (13 years) guarantees the production of SA series whose quality is comparable with the corresponding quarterly series.

Moreover, we experienced that monthly SA series are the most suitable to describe the short term dynamic of the indicators and they are more timely than quarterly ones in capturing changes in the cycle, thus producing less and smaller revisions. For these reasons, after having conducted experimental studies, a new procedure has been introduced in occasion of the monthly release referred to January 2016 (1st March 2016). The new procedure essentially consists in producing monthly SA series and obtaining the corresponding quarterly SA series by calculating a weighted average as previously described. The results are monthly

[1] Di Fonzo, T., Marini, M. (2011) "Simultaneous and Two-step Reconciliation of Systems of Time Series: Methodological and Practical Issues", JRSS C, 60, 143–164.

and quarterly SA series consistent among them and at different levels of aggregation.

In this paper the previous and the new procedures to achieve consistency between monthly and quarterly SA time series are described and a comparison of the results obtained through the two methods is shown, considering several quality dimensions of SA series (volatility, revisions, etc.)

## 2. The previous procedure

The procedure previously used to produce the LFS seasonally adjusted (SA) series, henceforth called 'OLD,' foresaw the reconciliation of the monthly SA series to the population totals (contemporary constraints) and to the quarterly SA series (inter-temporal constraints). Below the phases of this process are described.

Quarterly we produced and seasonally adjusted the following series:

- Population by labour status, gender and age class (30 series)[2];
- Population by labour status, economic sector and geographical area[3] (15 series);
- Employment by professional status and by temporary and permanent job[4] (3 series).

Monthly the following:

- Population by labour status, gender and age class (30 series);
- Employment by professional status and by temporary and permanent job (3 series).

Through a reconciliation procedure based on the so called *two steps* method by Di Fonzo and Marini, monthly series were reconciled to the quarterly.

First we reconciled the quarterly series as follows:

A. Reconciliation of the quarterly series by labour status, gender and age classes and by professional status and temporary and permanent job to the total amounts by labour status resulting from the series by labour status, economic sector and geographical area;

[2] Labour status: employment, unemployment, inactive; age classes: 15-24 years, 25-34 years, 35-49 years, 50-64 years, 65 years and over.

[3] Economic sector: agriculture, manufacturing, construction, services; geographical areas: Northern, Center and Southern of Italy.

[4] Temporary employees, permanent employees and self-employed.

B. Reconciliation of the quarterly series to the population totals by gender and age classes.

Afterwards the monthly series were reconciled as follows:

C. Reconciliation of the monthly series by gender and age classes to the corresponding quarterly series and to the population totals by gender and age classes;

D. Reconciliation of the monthly series by professional status and temporary and permanent job to the corresponding quarterly and to the employment level resulting by the C step.

Briefly, the OLD procedure ensured consistency between monthly and quarterly series constraining the first to the latest. That because of the higher quality that quarterly SA series had compared with monthly ones at the moment when monthly data started to be produced by Istat (late 2009). Quarterly SA series were in fact available starting from the fourth quarter 1992, while the monthly since January 2004.

## 3. The new procedure

After seven years, the experience gained in the seasonal adjustment and the availability of long monthly series allowed us to test an alternative strategy, henceforth called “NEW”: obtaining quarterly seasonally adjusted data as weighted average of monthly SA data (with weights 4 or 5 depending on the weeks in each month). Two reasons have led to test this new procedure.

First, the ability of the seasonally adjusted series, monthly and quarterly, to detect more or less quickly changes in trends, turning points in particular.

The analysis of SA series produced by the OLD method puts in evidence that more and stronger revisions occur close to the turning points. This evidence takes place for monthly and quarterly series, but with different time lags. If we analyze the monthly SA series before the reconciliation procedure, a monthly series needs 3-4 months to detect a turning point, a quarterly one needs 2-3 quarters. So, using the OLD method, the reconciliation procedure leads to amplify the period needed to detect a turning point in a monthly series and the revision effect (to 6-9 months).

The second reason concerns the request for more detailed SA time series (by more age classes and, for the employment, by professional status and temporary and permanent employees) which means more constraints to ensure consistency and thus greater complexity and higher bias probability.

According to the NEW procedure, only monthly series, both for the monthly and the quarterly required details, are seasonally adjusted. Every month 33 series5 are seasonally adjusted, every three months, when the quarter is completed, 15 monthly series more[6].

After the seasonal adjustment, some simple re-proportioning steps get monthly series consistent each other and with the population totals. Below the NEW procedure steps:

A. Monthly SA series by labour status, gender and age classes are re-proportionated to population totals by sex and age classes;
B. Re-proportioning of monthly employment SA series by professional status and temporary and permanent job to the employment series obtained at the previous step A.
C. Monthly employment SA series by economic activity and geographical area are re-proportionated to the employment series obtained at step A. Quarterly SA series are computed as weighted averages of the monthly series obtained at the previous steps, this step runs only when quarterly figures have to be released (every three months).

## 4. Comparing old and new procedure: experimentation results

Before being adopted, the NEW procedure has been tested for four months, since September to December 2015, and a comparative analysis has been conducted. A first evidence in favour of the adoption of the NEW procedure, which ensures about the final impact on the series to be released, is that monthly and quarterly SA series produced according to the OLD and to the NEW procedures show similar patterns. Furthermore during these four months both procedures have been evaluated in terms of quality dimensions, as volatility, revisions and residual seasonality.

### *4.1.Volatility and reconciliation effect*

To compare the series produced by the OLD and the NEW procedures with respect to the base series, that is the SA series without any reconciliation or re-proportioning process, we analyzed volatility through the frequency of double large

[5] Population by labour status, gender and five age classes (30 series) and employment by professional status and temporary and permanent job (3 series).

[6] Population by labour status , economic sector (for employment) and geographical area.

inversions and the relative standard deviation for 26 released series[7]. In this way also the reconciliation effect was evaluated.

Looking at the frequency of double large inversions of sign on the entire series (tab.1), considering 0.2% or 0.3% as large MoM change (RICVCSEG02 and RICVCSEG03), there aren’t relevant differences between OLD and NEW series with respect to the base series: if we assume 0.2% as a large MoM change, we observe a reduction only for 9 series over 26 in the NEW procedure and only for 10 series in the OLD. Also considering a threshold of 0.3%, the indicators for OLD and NEW procedures are similar and for both there is an improvement: for half of the series the frequency of large double inversions decreases.

**Table.1** – *Monthly SA released series by double large inversions in OLD and NEW procedure with respect to the base series.*

| | RICVCSEG02 | | RICVCSEG03 | |
|---|---|---|---|---|
| | NEW vs base | OLD vs base | NEW vs base | OLD vs base |
| Double large inversions increase | 15 | 14 | 10 | 12 |
| Double large inversions reduction | 9 | 10 | 12 | 13 |
| No change | 2 | 2 | 4 | 1 |

To point out if the NEW method adds variability to the series, the relative standard deviations over the series are computed for both the procedures and compared. Adopting the new method it was found a reduction in 16 series, however in the other 10 series the increase of the relative standard deviation is less than 2.5%. Taking into account all of the 26 series, the relative standard deviation slightly decreased on average (-0.4%).

Comparing OLD and NEW procedures with respect to the base series it was observed that for 22 series in NEW procedure the standard deviation is closer to the one computed on the base series, the difference increases for the remaining 4 series but is very small (0.04 percentage points on average).

With respect to the base series also mean, absolute and maximum differences and correlation have been evaluated (for the levels and for the changes, for example tab.2 provides correlations for MoM changes): as we expected, differences result to be larger for series in OLD procedure than for the ones in NEW, while correlations are bigger between NEW and base series than between the OLD and the base.

[7] In order to isolate individual components (trend-cycle, seasonality and irregular) the detail in which series are seasonally adjusted is greater than the one in which they are later released.

**Table.2 –** *Employment, unemployment, inactive and unemployment rate - MoM changes correlations between OLD and NEW procedures with respect to base series*

| | Employment | | Unemployment | | Inactive | | Unemployment rate | |
|---|---|---|---|---|---|---|---|---|
| | old-base | new-base | old-base | new-base | old-base | new-base | old-base | new-base |
| Over entire series | 0.71 | 0.75 | 0.89 | 1.00 | 0.89 | 0.96 | 0.90 | 1.00 |
| Over last 25 months | 0.72 | 0.73 | 0.92 | 1.00 | 0.88 | 0.98 | 0.92 | 1.00 |
| Over last 13 months | 0.80 | 0.79 | 0.92 | 1.00 | 0.78 | 0.97 | 0.94 | 1.00 |

It was also noticed, in particular for the main figures (employed, unemployed, inactive 15-64 years), that the number of MoM changes falling outside the confidence interval (considering the raw data sampling error) is similar for both procedures. In the whole series, considering the period January 2004 - December 2015, we observe:

- For the employed: only two MoM changes exceed the confidence interval, both for OLD and NEW series;

- For the unemployed, 21 MoM changes exceeding the confidence interval in the OLD, 25 in the NEW (4 more with NEW than the OLD, but just one more compared to the base series);

- For inactive 15-64 years, 5 MoM changes exceed the confidence interval, both for OLD and NEW series.

Regarding quarterly series, volatility of the NEW series is substantially unchanged with respect to the OLD series: there is an increase in half of the series and a reduction in the other half. On average there is an insignificant increase in volatility in terms of relative standard deviation of 1.1%.

### *4.2.Revisions*

To analyze and evaluate the revisions which occurred adopting the NEW procedure, the results obtained in the two procedures in terms of mean revisions, mean and maximum absolute revisions and sign revisions frequencies of the changes over the previous period (MoM or QoQ changes) are compared.

The tables below present some revision indicators for monthly seasonally adjusted series; in particular here, for the major aggregates, are provided indicators computed over last 37 points[8]. Table 3 refers to indicators computed on series in the OLD procedure, table 4 to the ones computed on series in the NEW procedure.

[8] Revisions between the release of October 2015 and the release of September 2015 are referred to the part of the series from Sep-12 to Sep-15; between November and October 2015 to the part from Oct-12 to Oct-15; for December over November 2015 to the part from Nov-12 to Nov-15.

Comparing the two tables, no substantial differences between OLD and NEW procedure arise in terms of mean and absolute revisions (MR, MAR and RMAX).

The sign revisions in the NEW procedure point out more revisions than in the OLD (SIGREV: absolute frequency of sign revisions) but their width is smaller, except for unemployment (SIGREV>0.1 percentage points means how many sign revision occurrences are bigger than 0.1 percentage points, SIGREV>0.2 percentage points bigger than 0.2 percentage points and SIGREV>0.3 percentage points bigger than 0.3 percentage points).

**Table.3 –** *OLD procedure: Revision indicators for major SA monthly aggregates on MoM changes over last 37 points*

| Rel(t)/Rel(t-1) | MR | MAR | RMAX | SIGREV | SIGREV>0.1 | SIGREV>0.2 | SIGREV>0.3 |
|---|---|---|---|---|---|---|---|
| | | | | Employment | | | |
| Oct/Sep-15 | 0.0 | 0.0 | 0.1 | 2 | 0 | 0 | 0 |
| Nov/Oct-15 | 0.0 | 0.0 | 0.1 | 1 | 0 | 0 | 0 |
| Dec/Nov-15 | 0.0 | 0.0 | 0.0 | 1 | 0 | 0 | 0 |
| | | | | Unemployment | | | |
| Oct/Sep-15 | -0.1 | 0.3 | 1.2 | 1 | 1 | 0 | 0 |
| Nov/Oct-15 | 0.0 | 0.1 | 0.3 | 0 | 0 | 0 | 0 |
| Dec/Nov-15 | 0.0 | 0.1 | 0.5 | 0 | 0 | 0 | 0 |
| | | | | Inactive | | | |
| Oct/Sep-15 | 0.0 | 0.1 | 0.3 | 1 | 0 | 0 | 0 |
| Nov/Oct-15 | 0.0 | 0.0 | 0.1 | 0 | 0 | 0 | 0 |
| Dec/Nov-15 | 0.0 | 0.0 | 0.1 | 0 | 0 | 0 | 0 |
| | | | | Employment rate | | | |
| Oct/Sep-15 | 0.0 | 0.0 | 0.1 | 1 | 0 | 0 | 0 |
| Nov/Oct-15 | 0.0 | 0.0 | 0.0 | 0 | 0 | 0 | 0 |
| Dec/Nov-15 | 0.0 | 0.0 | 0.0 | 0 | 0 | 0 | 0 |
| | | | | Unemployment rate | | | |
| Oct/Sep-15 | 0.0 | 0.0 | 0.1 | 1 | 0 | 0 | 0 |
| Nov/Oct-15 | 0.0 | 0.0 | 0.0 | 0 | 0 | 0 | 0 |
| Dec/Nov-15 | 0.0 | 0.0 | 0.1 | 0 | 0 | 0 | 0 |
| | | | | Inactivity rate | | | |
| Oct/Sep-15 | 0.0 | 0.0 | 0.1 | 2 | 0 | 0 | 0 |
| Nov/Oct-15 | 0.0 | 0.0 | 0.0 | 0 | 0 | 0 | 0 |
| Dec/Nov-15 | 0.0 | 0.0 | 0.0 | 0 | 0 | 0 | 0 |

Also revisions in absolute values have been analyzed: in particular larger revisions occur in OLD series when a new complete quarter is available (for example look at Fig.1, at revisions October 2015 vs. September 2015 in OLD procedure, when releasing October we also released the third quarter 2015), instead in the following months, November vs. October and December vs. November 2015, revisions are smaller.

**Table.4 –** *NEW procedure: Revision indicators for major SA monthly aggregates on MoM changes over last 37 points*

| Rel(t)/Rel(t-1) | MR | MAR | RMAX | SIGREV | SIGREV>0.1 | SIGREV>0.2 | SIGREV>0.3 |
|---|---|---|---|---|---|---|---|
| Employment | | | | | | | |
| Oct/Sep-15 | 0.0 | 0.0 | 0.1 | 1 | 0 | 0 | 0 |
| Nov/Oct-15 | 0.0 | 0.0 | 0.1 | 1 | 0 | 0 | 0 |
| Dec/Nov-15 | 0.0 | 0.1 | 0.2 | 4 | 2 | 0 | 0 |
| Unemployment | | | | | | | |
| Oct/Sep-15 | 0.0 | 0.3 | 0.9 | 1 | 1 | 1 | 0 |
| Nov/Oct-15 | 0.0 | 0.2 | 0.5 | 2 | 2 | 1 | 1 |
| Dec/Nov-15 | 0.0 | 0.6 | 1.6 | 7 | 7 | 7 | 6 |
| Inactive | | | | | | | |
| Oct/Sep-15 | 0.0 | 0.0 | 0.1 | 0 | 0 | 0 | 0 |
| Nov/Oct-15 | 0.0 | 0.0 | 0.1 | 0 | 0 | 0 | 0 |
| Dec/Nov-15 | 0.0 | 0.0 | 0.0 | 2 | 0 | 0 | 0 |
| Employment rate | | | | | | | |
| Oct/Sep-15 | 0.0 | 0.0 | 0.1 | 3 | 0 | 0 | 0 |
| Nov/Oct-15 | 0.0 | 0.0 | 0.1 | 1 | 0 | 0 | 0 |
| Dec/Nov-15 | 0.0 | 0.0 | 0.1 | 6 | 0 | 0 | 0 |
| Unemployment rate | | | | | | | |
| Oct/Sep-15 | 0.0 | 0.0 | 0.1 | 0 | 0 | 0 | 0 |
| Nov/Oct-15 | 0.0 | 0.0 | 0.1 | 0 | 0 | 0 | 0 |
| Dec/Nov-15 | 0.0 | 0.1 | 0.2 | 7 | 5 | 0 | 0 |
| Inactivity rate | | | | | | | |
| Oct/Sep-15 | 0.0 | 0.0 | 0.0 | 0 | 0 | 0 | 0 |
| Nov/Oct-15 | 0.0 | 0.0 | 0.0 | 2 | 0 | 0 | 0 |
| Dec/Nov-15 | 0.0 | 0.0 | 0.0 | 0 | 0 | 0 | 0 |

In the NEW series (for example look on the right of fig.1), however, revisions are lower on average and more uniformly distributed over the months (tab.5)

**Figure 1 –** *Unemployment (SA data, Jan-14/Dec-15, releases Sep-Dec 2015, values in millions)*

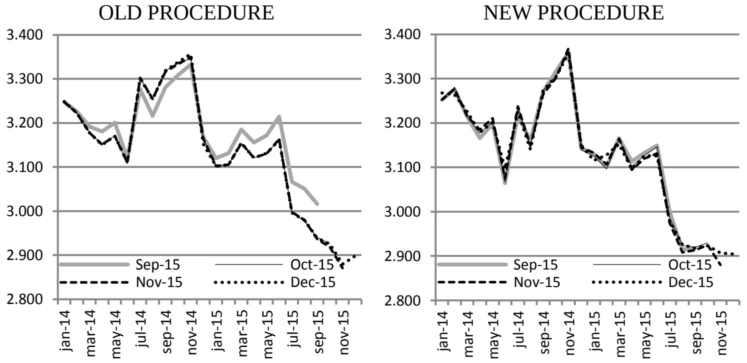


**Table.5 –** *Maximum absolute revisions in levels for major monthly aggregates over last 13 points (SA data).*

| | Release(t)/Release(t-1) | OLD procedure | NEW procedure |
|---|---|---|---|
| Employment | Oct-15/Sep-15 | 63,293 | 25,738 |
| | Nov-15/Oct-15 | 7,513 | 21,241 |
| | Dec-15/Nov-15 | 11,039 | 23,732 |
| Unemployment | Oct-15/Sep-15 | 76,011 | 14,551 |
| | Nov-15/Oct-15 | 8,801 | 16,029 |
| | Dec-15/Nov-15 | 11,836 | 26,866 |
| Inactive | Oct-15/Sep-15 | 139,694 | 15,160 |
| | Nov-15/Oct-15 | 8,782 | 17,194 |
| | Dec-15/Nov-15 | 22,830 | 7,390 |

### *4.3.New procedure impact*

Finally, the evaluation of the impact resulting from the adoption of the NEW procedure have been taken into account. Releasing January 2016, adopting the NEW procedure, we could compute revisions occurred in the MoM changes between the two approaches comparing the series released with January 2016 with the previous released in December 2015 (we considered the last 37 months and the last 13 months of the series, but also the entire series, since January 2004).

**Table.6 –** *Revisions of MoM changes in OLD and NEW procedure - summary statistics-Feb04-Dec15 (values in percentage points)*

| Indicators | Employ-ment | Unemploy-ment | Inactive | Employ-ment rate | Unemploy-ment rate | Inactivity rate |
|---|---|---|---|---|---|---|
| MAR | 0.11 | 0.89 | 0.18 | 0.10 | 0.07 | 0.07 |
| MR | 0.00 | 0.01 | -0.01 | 0.00 | 0.00 | 0.00 |
| St dev | 0.008 | 0.075 | 0.016 | 0.011 | 0.006 | 0.006 |
| T-test for MR | 0.35 | 0.08 | -0.39 | -0.17 | 0.08 | -0.37 |
| MR Statistical significance | NO | NO | NO | NO | NO | NO |
| % positive revisions | 50.3 | 50.3 | 47.6 | 48.3 | 49.7 | 49.0 |
| % negative revisions | 49.7 | 49.7 | 52.4 | 51.7 | 50.3 | 51.0 |
| % revisions equal to zero | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| % sign(new) = sign(old) | 85.3 | 86.7 | 82.5 | 82.5 | 84.6 | 82.5 |

In table 6, referred to the whole series, some evidences are shown up: first it can be noticed that for all the most important figures revisions are not significant[9] and also are fairly equally distributed between positive and negative; second, in the NEW series the sign of the MoM change is the same of the OLD in at least 82% of cases.

Releasing January 2016 also the fourth quarter 2015 has been released; looking at tab.7 it's possible to analyze that impact also on quarterly series.

Revisions on QoQ growth rate are not relevant (MAR and MR), revisions different from zero are equally distributed between positive and negative ones and at least in 92,3% of cases the sign remains the same.

**Table.7 – Revisions of QoQ changes in OLD and NEW procedure - summary statistics -1993Q1-2015Q4 (values in percentage points)**

| Indicators | Employ-ment | Unemploy-ment | Inactive | Employ-ment rate | Unemploy-ment rate | Inactivity rate |
|---|---|---|---|---|---|---|
| MAR | 0.06 | 0.52 | 0.11 | 0.03 | 0.04 | 0.04 |
| MR | 0.00 | -0.02 | 0.00 | 0.00 | 0.00 | 0.00 |
| St dev | 0.007 | 0.045 | 0.013 | 0.004 | 0.004 | 0.005 |
| T-test for MR | 0.58 | -0.35 | -0.36 | 0.59 | -0.40 | -0.34 |
| MR Statistical significance | NO | NO | NO | NO | NO | NO |
| % positive revisions | 25.3 | 24.2 | 28.6 | 45.1 | 25.3 | 48.4 |
| % negative revisions | 26.4 | 27.5 | 23.1 | 53.8 | 26.4 | 50.5 |
| % revisions equal to zero | 48.4 | 48.4 | 48.4 | 1.1 | 48.4 | 1.1 |
| % sign(new) = sign(old) | 92.3 | 94.5 | 93.4 | 95.6 | 93.4 | 95.6 |

Another important issue about quarterly seasonally adjusted series obtained in the NEW procedure concerns residual seasonality: actually there could be a residual seasonality in quarterly SA series obtained as weighted average of SA monthly data. Therefore all 'calculated' quarterly series have been re-seasonally adjusted and it has been checked there was no difference between series before and after, so no residual seasonality has been founded.

## 5. Conclusions

Monthly and quarterly seasonally adjusted data requires to be inter-temporally consistent each other. Until the release of December 2015 that constraint was

[9] Statistical significance of the MR is evaluated considering the following T-statistic critical values (0,1/0,05/0,01): 1.66/1.98/2.61.

satisfied reconciling monthly seasonally adjusted data to the quarterly, with resulting revisions due to the different delay in monthly and quarterly seasonally adjusted series to capture dynamics. More experience and the availability of fairly long monthly series allowed us to adopt from January 2016 a very simplified procedure that guarantees the inter-temporal consistency obtaining the seasonally adjusted quarterly data as a weighted averages of the corresponding monthly data.

In this way a large number of constraints to ensure consistency and thus greater complexity and higher bias probability, also due to the request for more detailed SA monthly series, have been avoided.

Implications and impacts in terms of the quality of the series to be released have been taken into account and the main results are shown in this paper. The analysis of representative indicators of quality dimensions such as the volatility and revisions resulting from trials conducted in parallel for four months, made it possible to compare the two procedures, and to draw some important and comforting conclusions in support of the new method.

In terms of volatility it can be affirmed that the new monthly series are not to be considered more volatile than the old, indeed, in most cases a reduction in the volatility is achieved and also, analyzing the reconciliation effect, it was found that the monthly pattern described by the base series is less revised by the new seasonally adjusted series than by the old. Lastly, monthly and quarterly series obtained using the new approach seems to be not affected by revisions more than the ones in old procedure; moreover revisions are smaller and uniformly distributed over the months, instead of being concentrated in the first month of the quarter as a result of the update when the complete quarterly figure is available.

## Riferimenti bibliografici

DI FONZO T., MARINI M., 2011. Simultaneous and two-step reconciliation of systems of time series: methodological and practical issues, Journal of the Royal Statistical Society C, 60, 2, pp. 143-164

EUROSTAT, 2015. ESS guidelines on seasonal adjustment, *Eurostat Manuals and Guidelines*, http://ec.europa.eu/eurostat/documents/3859598/6830795/KS-GQ-15-001-EN-N.pdf .

EUROSTAT, 2013. ESS guidelines on revision policy for PEEIs, *Eurostat Methodologies and Working papers*,

McKENZIE R., GAMBA M., 2008. Interpreting the results of Revision Analyses: Recommended Summary Statistics. *Contribution to the OECD/Eurostat Task Force on "Performing Revisions Analysis for Sub-Annual Economic Statistics"*, http://www.oecd.org/std/40315546.pdf.

## SUMMARY

### How to achieve consistency between monthly and quarterly seasonally adjusted time series?

An innovation is being introduced in the seasonal adjustment (SA) procedure, in particular as regards the methodology to guarantee consistency between monthly and quarterly SA time series.

The procedure currently used foresees that, once seasonally adjusted, monthly time series are reconciled to the corresponding quarterly SA series. In this way the weighted average of monthly data (with weights equal to the number of weeks in each month, 4 or 5) is made equal to the corresponding quarterly figure.

This procedure was justified not only by the need to ensure consistency between the corresponding monthly and quarterly time series, but also because of the higher quality that quarterly SA series had compared with monthly ones, at the moment when monthly data started to be produced by Istat (late 2009). Quarterly SA series were in fact available starting from the fourth quarter 1992, while the monthly since January 2004.

Today, the length of monthly series (13 years) guarantees the production of SA series whose quality is comparable with the corresponding quarterly. Moreover, we experienced that SA monthly series are the most suitable to describe the short term dynamic of the indicators and they are more timely than quarterly ones in capturing changes in the dynamic, thus producing less and smaller revisions.

For these reasons, a new procedure is being introduced in occasion of the monthly release referred to January 2016. The new procedure essentially consists in producing monthly SA series and obtaining the corresponding quarterly SA series by calculating a weighted average as previously described. The results are monthly and quarterly SA series consistent among them and at different levels of aggregation.

In this paper the previous and the new procedures to achieve consistency between monthly and quarterly SA time series are described and a comparison of the results obtained through the two methods is shown, considering several quality dimensions of SA series (volatility, revisions, etc.) recommended by Eurostat.

Cinzia GRAZIANI, ISTAT, cingraziani@istat.it
Silvia LORIGA, ISTAT, siloriga@istat.it
Michele Antonio SALVATORE, ISTAT, salvatore@istat.it
Andrea SPIZZICHINO, ISTAT, spizzich@istat.it

# SCOMPOSIZIONE DI INDICI E TASSI DI VARIAZIONE: FOCUS SU VARIABILI RAPPORTO E LEGGE GENERALE DI SCOMPOSIZIONE DEI TASSI DI VARIAZIONE (LSV)

Marco Lattanzio

## Premessa

Solitamente, in particolare nell'ambito delle statistiche ufficiali, risulta utile conoscere la relazione che lega un indice calcolato su un aggregato con gli indici relativi a insiemi che partizionano l'aggregato, definiti componenti oppure la relazione che lega gli indici di una serie di variabili con l'indice di una variabile che è funzione delle precedenti. Nel lavoro si focalizza l'attenzione sulle variabili rapporto ovvero ottenute dal rapporto di altre due variabili. Le formule saranno ricavate sia per gli indici, risultato che sarà ottenuto con un approccio *top-down*, sia per i tassi di variazione per i quali è possibile utilizzare sia un approccio *top-down* sia un approccio analitico. Questo secondo approccio sfrutta un risultato generale, descritto in appendice, che permette di ricavare la relazione di scomposizione delle variazioni una volta nota la relazione di scomposizione degli indici.

## 1. Indici semplici e variabili rapporto

### 1.1. *Definizione di indici e variazioni*

Sia $X_t$ una variabile osservata in serie storica ad intervalli regolari (mensili, trimestrali, etc.) dove $t$ indica l'istante di tempo relativo all'osservazione. Definiremo il suo indice semplice a base fissa come:

$$I_t^X = \frac{X_t}{X_b} \tag{1}$$

dove $b$ rappresenta l'anno base, rispetto al quale vengono rapportati i valori osservati ad ogni istante $t$ [1].

Sia ora la variabile $X_t$, ottenuta dal rapporto di due variabili, definite rispettivamente $N_t$ e $D_t$.

L' indice semplice della variabile rapporto $X_t$, sarà definito come segue:

[1] Di norma nel calcolo di indici a base fissa, quale valore osservato in un istante temporale b, si considera la media delle osservazioni di X in un determinato anno, per cui b si definisce anno base.

$$I_t^X = \frac{X_t}{X_b} = \frac{N_t/D_t}{N_b/D_b} \qquad (2)$$

Definiamo inoltre la variazione (o tasso di variazione) di passo *h*[2], indicata con il simbolo $V_h(\cdot)$, la differenza relativa tra l'indice della variabile *X* al tempo *t+h*, e l'indice rispetto al tempo *t,* in valore percentuale, definita dalla relazione seguente:

$$V_h(I_t^X) = \left(\frac{I_{t+h}^X}{I_t^X} - 1\right) 100 \qquad (3)$$

Si osserva che la relazione che esprime la variazione della variabile *X* calcolata sull'indice restituisce lo stesso risultato se calcolata sulla variabile stessa, per via della semplificazione dei termini riferiti all'anno base[3].

$$V_h(I_t^X) = \left(\frac{I_{t+h}^X}{I_t^X} - 1\right) 100 = \left(\frac{\frac{X_{t+h}}{X_b}}{\frac{X_t}{X_b}} - 1\right) 100 = \left(\frac{X_{t+h}}{X_t} - 1\right) 100 = V_h(X_t) \qquad (4)$$

### *1.2. Scomposizioni*

Ricaveremo le formule di scomposizione degli indici e dei tassi di variazione sia per la variabile rapporto come funzione di numeratore e denominatore, sia per la stessa variabile come aggregazione di variabili rapporto calcolate su partizioni dell'insieme aggregato. Come detto in precedenza le relazioni per i tassi di variazione saranno ottenuti diversamente, in particolare utilizzando una formula generale, definita Legge di Scomposizione dei tassi di Variazione (LSV) di seguito descritta:
sia $Y_t = f\,(Z_t^1, Z_t^2, \dots, Z_t^k)$ , dove $Y_t$ rappresenta l'indice (o la variabile) aggregata e i singoli $Z_t^j$rappresentano gli indici (o le variabili) componenti allora, in generale, si ha:

$$V_h(Y_t) = V_h\left(f\,(Z_t^1, Z_t^2, \dots, Z_t^k)\right) = g(\,V_h(Z_t^1), V_h(Z_t^2), \dots, V_h(Z_t^k)\,) =$$
$$= \frac{1}{f} < \nabla f, V_h\left(\underline{Z}_t\right)\underline{Z}_t > + \frac{1}{200\,f} < V_h\left(\underline{Z}_t\right)\underline{Z}_t^T, \nabla^2 f\; V_h\left(\underline{Z}_t^T\right)\underline{Z}_t > + R\left(\underline{Z}_t, h\right) \qquad (4)$$

Dove con $V_h\left(\underline{Z}_t\right)$ si indica il vettore dei tassi di variazione di passo *h* delle componenti e $\underline{Z}_t$ il vettore degli indici o delle variabili componenti, $\nabla f$ e $\nabla^2 f$ rispettivamente il

[2] h=1 per le variazioni congiunturali, h=4 per le variazioni tendenziali di osservazioni trimestrali e h=12 per le variazioni tendenziali di osservazioni mensili

[3] Occorre precisare che l'uguaglianza descritta può non essere verificata in presenza di arrotondamenti intermedi effettuati sull'indice prima del calcolo della variazione. Traendo spunto dalla teoria sui metodi numerici (cfr. [2]), si verifica che l'errore originario sugli indici, causato dall'operazione di arrotondamento , si propaga tramite il calcolo della variazione tendenziale in misura inversamente proporzionale alla differenza degli indici (più la differenza è piccola più l'errore si amplifica).

vettore gradiente e la matrice Hessiana della funzione *f*, *R* il termine residuo e $<.,.>$ il prodotto scalare tra vettori.

Nel caso in cui la funzione *f* è di grado al più 2 come funzione delle $Z_t^j$ il residuo sarà pari a 0 e pertanto la stima sarà precisa. Negli altri casi il residuo non sarà trascurabile.

## *1.3. Scomposizione numeratore-denominatore*

Risulta di facile verifica che l'indice semplice di una variabile rapporto è uguale al rapporto degli indici rispettivamente della variabile a numeratore e a denominatore:

$$I_t^X = \frac{N_t/D_t}{N_b/D_b} = \frac{N_t/N_b}{D_t/D_b} = \frac{I_t^N}{I_t^D} \tag{5}$$

***Proposizione 1.3.1***

Sia $Y_t = f(Z_t^1, Z_t^2) = \frac{Z_t^1}{Z_t^2} = \frac{N_t}{D_t}$, con $D_t \neq 0\ \forall t$, allora, applicando la formula (4) si ottiene la seguente relazione di scomposizione:

$$V_h(I_t^X) = \left(V_h(I_t^N) - V_h(I_t^D)\right)\left(1 - \frac{V_h(I_t^D)}{100}\right) + R\left(\underline{Z}_t, h\right) \tag{6}$$

Dimostrazione:

$$V_h(Y_t) = \frac{1}{f}\sum_i \frac{\partial f}{\partial Z_t^i}\ V_h(Z_t^i)Z_t^i + \frac{1}{200\,f}\sum_{i,j}\frac{\partial f}{\partial Z_t^i}\frac{\partial f}{\partial Z_t^j}V_h(Z_t^i)\,V_h(Z_t^j)\,Z_t^iZ_t^j$$

$$\nabla f = \left(\frac{\partial f}{\partial N_t}, \frac{\partial f}{\partial D_t}\right) = \left(\frac{1}{D_t}, \frac{-N_t}{(D_t)^2}\right)\ e\ \nabla^2 f = \begin{pmatrix} 0 & -\frac{1}{(D_t)^2} \\ -\frac{1}{(D_t)^2} & \frac{2N_t}{(D_t)^3} \end{pmatrix}$$

$$\begin{aligned} V_h(Y_t) = \frac{D_t}{N_t}&\left(\frac{1}{D_t}V_h(N_t)N_t - \frac{N_t}{(D_t)^2}V_h(D_t)D_t\right) \\ &- \frac{1}{200}\frac{D_t}{N_t}\left(\frac{2}{(D_t)^2}V_h(N_t)N_tV_h(D_t)D_t + \frac{2N_t}{(D_t)^3}V_h(D_t)^2(D_t)^2\right) = \end{aligned}$$

$$= V_h(N_t) - V_h(D_t) - \frac{1}{100}V_h(N_t)V_h(D_t) + \frac{1}{100}V_h(D_t)^2 =$$

$$\left(V_h(N_t) - V_h(D_t)\right)\left(1 - \frac{V_h(D_t)}{100}\right) \qquad \square$$

Se osserviamo che $\frac{V_h(I_t^D)}{100} << 1$, possiamo considerare la seguente approssimazione:

$$V_h(I_t^X)\ \approx\ V_h(I_t^N) - V_h(I_t^D) \tag{7}$$

La relazione (5) non è né lineare né quadratica come funzione delle $Z_t^j$ pertanto il residuo non è trascurabile. Si può dimostrare con approccio *top-down* che vale la seguente stima precisa:

$$V_h(I_t^X) = \frac{V_h(I_t^N) - V_h(I_t^D)}{1 + \frac{V_h(I_t^D)}{100}}$$

### 1.4. Scomposizione di un aggregato in sotto indici

Supponiamo di avere una partizione delle unità statistiche in *n* sottoinsiemi disgiunti di $S$, indicati con $S_1, \dots, S_n$. Può essere un esempio la classificazione di un insieme di unità statistiche rispetto all'attività economica svolta (attualmente ATECO 2007). Ad esempio pensiamo ad $S$ come il macro settore dell'industria (aggregato B-F) e gli *n* sottoinsiemi come le sezioni Ateco che lo compongono (sezioni B, C, D, E ed F). Altre scomposizioni possono essere pensate e vedremo che la scelta della scomposizione può portare ad avere risultati differenti.

Definiamo l'indice relativo ad un sotto insieme $S_i$ di $S$, come l'indice calcolato sulla variabile rapporto osservata sul settore $S_i$ all'istante *t* rispetto all'anno base. Quest'ultima variabile si definisce come il rapporto tra l'ammontare della variabile *N* misurato sulle unità *j* appartenenti al sotto-insieme *i*-esimo e l'ammontare relativo alla variabile *D*[4]. Indicheremo con $m_{i,t}$ la numerosità dell'insieme $S_i$[5].

$$I_{S_i,t}^X = \frac{X_{S_i,t}}{X_{S_i,b}} = \frac{\frac{\sum_{i \in S_j} N_{j,t}}{\sum_{i \in S_j} D_{j,t}}}{\frac{\sum_{i \in S_j} N_{j,b}}{\sum_{i \in S_j} D_{j,b}}}$$

Dalla definizione precedente si ricava l'indice dell'aggregato $S = \bigcup_{i=1}^{n} S_i$ che sarà uguale a

$$I_{S,t}^X = \frac{X_{S,t}}{X_{S,b}} = \frac{\frac{\sum_{i=1}^{n} \sum_{i \in S_j} N_{j,t}}{\sum_{i=1}^{n} \sum_{i \in S_j} D_{j,t}}}{\frac{\sum_{i=1}^{n} \sum_{i \in S_j} N_{j,b}}{\sum_{i=1}^{n} \sum_{i \in S_j} D_{j,b}}}$$

[4] La formulazione si adatta al caso di indagini censuarie. Nel caso di indagini campionarie i valori $N_{j,t}$ e $D_{j,t}$ devono essere sostituite dalle rispettive stime campionarie.

[5] La numerosità dell'insieme $S_i$ al tempo t non è necessariamente la dell' insieme in un diverso istante temporale: in questa formulazione rientrano entrambi i casi di indagine panel e non panel in cui si considera la demografia di impresa.

Per semplicità scriveremo $N_{s_i,t}$ per indicare $\sum_{i \in s_j} N_{j,t}$. La stessa notazione è usata in maniera speculare per la variabile *D*. L'indice generale del settore *S*, $I^X_{S,t}$ , può essere scomposto per ricavare una relazione del tipo:

$$I^X_{S,t} = f(I^X_{s_1,t}, I^X_{s_2,t}, \dots, I^X_{s_n,t}) \qquad (8)$$

Saremo interessati inoltre a verificare la proprietà di additività dell'indice. In particolare l'indice aggregato (oppure la variazione) verificherà questa proprietà se è possibile ricondurre il suo calcolo ad una media ponderata di sotto-indici, ricavando un sistema di pesi, $\pi_1, \dots, \pi_n$ ,che sommano ad 1

Tale proprietà ha come implicazione la proprietà di internalità (o coerenza interna), per cui l'indice generale sarà contenuto per costruzione in un intervallo avente come estremi il minimo e il massimo degli indici delle componenti.

***Proposizione 1.4.1***

L'indice aggregato di variabili rapporto non verifica la proprietà di additività

$$I^X_{S,t} \neq \sum_i I^X_{s_i,t}\, \pi_{i,b}$$

$$\textstyle\sum_i \pi_{i,b} = 1.$$

Dimostrazione:

$$I^X_{S,t} = \frac{X_{S,t}}{X_{S,b}} = \frac{\frac{N_{S,t}}{D_{S,t}}}{\frac{N_{S,b}}{D_{S,b}}} = \frac{\frac{\sum_i N_{s_i,t}}{\sum_i D_{s_i,t}}}{\frac{\sum_i N_{s_i,b}}{\sum_i D_{s_i,b}}} = \frac{\sum_i \frac{N_{s_i,t}}{D_{s_i,t}} \frac{D_{s_i,t}}{\sum_i D_{s_i,t}}}{\frac{\sum_i N_{s_i,b}}{\sum_i D_{s_i,b}}} =$$

$$= \frac{\sum_i \frac{\frac{N_{s_i,t}}{D_{s_i,t}}}{\frac{N_{s_i,b}}{D_{s_i,b}}} \frac{N_{s_i,b}}{D_{s_i,b}} \frac{D_{s_i,t}}{\sum_i D_{s_i,t}}}{\frac{N_{s_i,b}}{\sum_i D_{s_i,b}}} = \sum_i I^X_{s_i,t} \frac{N_{s_i,b}}{\sum_i N_{s_i,b}} \frac{\frac{D_{s_i,t}}{\sum_i D_{s_i,t}}}{\frac{D_{s_i,b}}{\sum_i D_{s_i,b}}}$$

Nell'ultima relazione si può definire:

$$I^{\pi_D}_{s_i\,,t} = \frac{\frac{D_{s_i,t}}{\sum_i D_{s_i,t}}}{\frac{D_{s_i,b}}{\sum_i D_{s_i,b}}}$$

che rappresenta l'indice della composizione della variabile denominatore *D* misurato sulle unità appartenenti al sotto settore *i*-esimo $S_i$ rispetto all'aggregato *S*.

Abbiamo quindi:

$$I^X_{S,t} = \sum_i I^X_{s_i,t}\, I^{\pi_D}_{s_i,t}\, \pi^N_{s_i,b} \neq \sum_i I^X_{s_i,t}\, \pi^N_{s_i,b} \qquad (9)$$

$$\sum_i \pi^N_{s_i,b} = 1 \qquad \square$$

L'indice si può esprimere come media ponderata di prodotti di indici. Questo indice non dipende solo dagli indici della variabile rapporto dei sotto-settori ma anche da altrettanti indici (variabili rispetto a *t*) che misurano la dinamica della variabile a denominatore tramite una sua funzione, la composizione rispetto ai macro settori. La relazione è di tipo additivo in quanto verifica le relazioni in (9) ma lo è rispetto al prodotto tra l'indice della variabile rapporto e della composizione. Per cui non risulta verificata la proprietà di additività dei sotto-indici per la singola variabile rapporto $X$.

***Proposizione 1.4.2***

*Sia* $I^X_{S,t} = f\left(I^X_{1,t}, I^X_{2,t}, \dots, I^X_{n,t}, I^{\pi_D}_{1,t}, I^{\pi_D}_{2,t}, \dots, I^{\pi_D}_{n,t}\right) = \sum_i I^X_{s_i,t}\, I^{\pi_D}_{s_i,t}\, \pi^N_{s_i,b}$ allora applicando la formula (4) si ottiene la seguente relazione di scomposizione per le variazioni:

$$V_h\left(I^X_{S,t}\right) = \sum_i \left(V_h\left(I^X_{s_i,t}\right) \pi^N_{s_i,t} + V_h\left(I^{\pi_D}_{s_i\,,t}\right)\right) \pi^N_{s_i,t} + \sum_i \frac{V_h\left(I^X_{s_i,t}\right) V_h\left(I^{\pi_D}_{s_i\,,t}\right)}{100}\, \pi^N_{s_i,t} \qquad (10)$$

dove si è posto $\left(\frac{\frac{D_{s_i,t+h}}{\sum_i D_{s_i,t+h}}}{\frac{D_{s_i,t}}{\sum_i D_{s_i,t}}} - 1\right) 100 = V_h\left(\pi^D_{s_i,t}\right) = V_h\left(I^{\pi_D}_{s_i\,,t}\right)$, termine che rappresenta la variazione di passo *h* della composizione della variabile denominatore *D* (o del suo indice) .

Dimostrazione:

Scriveremo per semplicità di notazione $i$ al posto di $s_i$e $\pi^i_b$ al posto di $\pi^N_{i,b}$.

$$Y_t = I^X_{S,t} = f\left(\underline{Z}_{\,t}\right) = \sum_{i=1}^n Z^i_t\, Z^{n+i}_t\, \pi^i_b = \sum_{i=1}^n I^X_{i,t}\, I^{\pi_D}_{i,t}\, \pi^i_b$$

Utilizziamo la formula (4) :

$$\nabla f = \left(\frac{\partial f}{\partial I^X_{1,t}}, \frac{\partial f}{\partial I^X_{2,t}}, \dots, \frac{\partial f}{\partial I^X_{n,t}}, \frac{\partial f}{\partial I^{\pi_D}_{1,t}}, \frac{\partial f}{\partial I^{\pi_D}_{2,t}}, \dots, \frac{\partial f}{\partial I^{\pi_D}_{n,t}}\right) = \left(I^{\pi_D}_{1,t}\pi^1_b, I^{\pi_D}_{2,t}\pi^2_b, \dots, I^{\pi_D}_{n,t}\pi^n_b, I^X_{1,t}\pi^1_b, I^X_{2,t}\pi^2_b, \dots, I^X_{n,t}\pi^n_b\right)$$

$$\nabla^2 f = \begin{pmatrix} \frac{\partial f}{\partial^2 I_{1,t}^X} & \cdots & \frac{\partial f}{\partial I_{1,t}^X \partial I_{n,t}^X} & \frac{\partial f}{\partial I_{1,t}^X \partial I_{1,t}^{\pi_D}} & \cdots & \frac{\partial f}{\partial I_{1,t}^X \partial I_{n,t}^{\pi_D}} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f}{\partial I_{n,t}^X \partial I_{1,t}^X} & \cdots & \frac{\partial f}{\partial^2 I_{n,t}^X} & \frac{\partial f}{\partial I_{n,t}^X \partial I_{1,t}^{\pi_D}} & \cdots & \frac{\partial f}{\partial I_{n,t}^X \partial I_{n,t}^{\pi_D}} \\ \frac{\partial f}{\partial I_{1,t}^{\pi_D} \partial I_{1,t}^X} & \cdots & \frac{\partial f}{\partial I_{1,t}^{\pi_D} \partial I_{n,t}^X} & \frac{\partial f}{\partial^2 I_{1,t}^{\pi_D}} & \cdots & \frac{\partial f}{\partial I_{1,t}^{\pi_D} \partial I_{n,t}^{\pi_D}} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f}{\partial I_{n,t}^{\pi_D} \partial I_{1,t}^X} & \cdots & \frac{\partial f}{\partial I_{n,t}^{\pi_D} \partial I_{n,t}^X} & \frac{\partial f}{\partial I_{n,t}^{\pi_D} \partial I_{1,t}^{\pi_D}} & \cdots & \frac{\partial f}{\partial^2 I_{n,t}^{\pi_D}} \end{pmatrix} = \begin{pmatrix} 0 & \cdots & 0 & \pi_b^1 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & 0 & 0 & \dots & \pi_b^n \\ \pi_b^1 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & \pi_b^n & 0 & \dots & 0 \end{pmatrix} =$$

$$= \begin{pmatrix} 0 & diag(\pi_b^1, \dots, \pi_b^n) \\ diag(\pi_b^1, \dots, \pi_b^n) & 0 \end{pmatrix}$$

$$\nabla^2 f \cdot V_h(Z_t^i) Z_t^i = \left(V_h(I_{1,t}^{\pi_D}) I_{1,t}^{\pi_D} \pi_b^1, \dots, V_h(I_{1,t}^{\pi_D}) I_{n,t}^{\pi_D} \pi_b^n, V_h(I_{1,t}^X) I_{1,t}^X \pi_b^1, \dots, V_h(I_{n,t}^X) I_{n,t}^X \pi_b^n\right)$$

$$< V_h(Z_t^i) Z_t^i, \nabla^2 f \cdot V_h(Z_t^i) Z_t^i >= \sum_{i=1}^n V_h(I_{i,t}^{\pi_D}) I_{i,t}^{\pi_D} V_h(I_{i,t}^X) I_{i,t}^X \pi_b^i$$

$$V_h(Y_t) = \frac{1}{\sum_{i=1}^n I_{i,t}^X \, I_{i,t}^{\pi_D} \, \pi_b^i} \left( \sum_{i=1}^n I_{i,t}^{\pi_D} \pi_b^i V_h(I_{1,t}^X) I_{1,t}^X + \sum_{i=1}^n I_{i,t}^X \pi_b^i V_h(I_{i,t}^{\pi_D}) I_{i,t}^{\pi_D} \right) + \frac{1}{200 \sum_{i=1}^n I_{i,t}^X \, I_{i,t}^{\pi_D} \, \pi_b^i} \sum_{i=1}^n V_h(I_{i,t}^{\pi_D}) I_{i,t}^{\pi_D} V_h(I_{i,t}^X) I_{i,t}^X \pi_b^i =$$

$$= \frac{1}{\sum_{i=1}^n I_{i,t}^X \, I_{i,t}^{\pi_D} \, \pi_b^i} \sum_{i=1}^n I_{i,t}^X I_{i,t}^{\pi_D} \pi_b^i \left( V_h(I_{i,t}^X) + V_h(I_{i,t}^{\pi_D}) \right) + \frac{1}{200 \sum_{i=1}^n I_{i,t}^X \, I_{i,t}^{\pi_D} \, \pi_b^i} \sum_{\substack{i,j=1 \\ i \neq j}}^n I_{i,t}^{\pi_D} I_{i,t}^X \pi_b^i \left( V_h(I_{i,t}^{\pi_D}) \cdot V_h(I_{j,t}^X) \right)$$

Effettuando alcune semplificazioni otteniamo:

$$\frac{I_{i,t}^{\pi_D} I_{i,t}^X \pi_b^i}{\sum_{i=1}^n I_{i,t}^X \, I_{i,t}^{\pi_D} \, \pi_b^i} = \frac{N_{i,t}}{\sum_{i=1}^n N_{i,t}} = \pi_t^i$$

Ottenendo la seguente relazione:

$$V_h(Y_t) = \sum_{i=1}^n \left( V_h(I_{s_i,t}^X) + V_h(I_{s_i,t}^{\pi_D}) \right) \pi_{i,t}^N + \frac{1}{200} \sum_{i=1}^n \left( V_h(I_{s_i,t}^{\pi_D}) \cdot V_h(I_{s_i,t}^X) \right) \pi_{i,t}^N \quad (11)$$

In questo caso $R\left(\underline{Z}_t, h\right) = 0$ in quanto *f* è di secondo grado come funzione delle $\underline{Z}_t$

Le formule sulla scomposizione di indici e variazioni sviluppate precedentemente fanno emergere inoltre alcune interessanti caratteristiche dell'indice semplice di variabili rapporto:

- la variazione dell'indice non è una media ponderata di variazioni, caratteristica comune degli indici semplici di variabili livello, ma è pari alla somma di tre medie ponderate: una prima media delle variazioni della variabile rapporto, una seconda media delle variazioni della composizione della variabile denominatore e, infine, una terza media ponderata del prodotto tra le due variazioni precedentemente descritte, diviso 100. Quest'ultimo termine $\sum_i \frac{V_h\left(X_{s_i,t}\right)V_h\left(\pi^D_{s_i,t}\right)}{100}\,\pi^N_{s_i,t}$ , il più delle volte risulta essere trascurabile, rispetto alle prime due componenti della variazione[6];

- guardando alla scomposizione in sotto-indici la dinamica generale è funzione della dinamica di una funzione della variabile denominatore osservata sui sotto-settori definita. Essa tiene conto della dinamica del peso della variabile denominatore calcolata sulle unità dell'insieme *i*-esimo rispetto al totale dell'aggregato su cui si calcola l'indice generale, ha generalmente una variabilità minore rispetto alla variabile livello e, affinché ci sia un effetto sulla variazione finale, definito *effetto composizione*, la ricomposizione strutturale tra i settori $S_i$ deve essere rilevante. Inoltre questo effetto è diverso a seconda della scomposizione dell'aggregato *S* che si considera;
- dalla scomposizione ne consegue che l'indice non gode, in generale, della proprietà di additività e, dunque, della proprietà di internalità rispetto i sotto-indici, per cui si può verificare che il valore dell'indice generale sia inferiore al minimo oppure superiore al massimo degli indici dei sotto settori (casi definiti *fuori range*);
- a queste relazioni seguono delle implicazioni sull'utilizzo di un approccio indiretto alla destagionalizzazione. L'eventuale destagionalizzazione delle serie dei sotto-indici e una successiva loro aggregazione per ottenere l'indice aggregato destagionalizzato non sarebbe esaustiva e corretta, a meno di destagionalizzare anche l'indice della composizione della variabile *D*.;
- la scomposizione appena mostrata è molto utile per la ricerca delle cause di valori anomali all'interno dei sotto-insiemi di classificazione, in quanto consente di evidenziare il contributo delle singole componenti sulla variazione complessiva.

### *1.5. La legge di scomposizione delle variazioni (LSV) di un indice in funzione delle componenti: un approccio di stima con polinomi di Taylor*

Sia $Y_t = f\ (Z_t^1, Z_t^2, \dots, Z_t^k)$ , i cui elementi sono definiti come nel paragrafo 1.2. Cercheremo di ottenere la seguente relazione:

[6] In particolare i valori del termine indicato risultano spesso inferiori a 0,05 per cui trascurabili a seguito di arrotondamento.

$$V_h(Y_t) = V_h\left(f\left(Z_t^1, Z_t^2, \dots, Z_t^k\right)\right) = g\left(V_h(Z_t^1), V_h(Z_t^2), \dots, V_h(Z_t^k)\right) = \frac{1}{f} < \nabla f, V_h(\underline{Z}_t)\underline{Z}_t >$$
$$+ \frac{1}{200\,f} < V_h(\underline{Z}_t)\underline{Z}_t^T, \nabla^2 f\, V_h(\underline{Z}_t^T)\underline{Z}_t > \; + o(h^2) + o\left(\|\underline{Z}_{t+h} - \underline{Z}_t\|^2\right) \qquad (12)$$

Consideriamo le $k$ funzioni $Z_1(t), Z_2(t), \dots, Z_k(t)$, ognuna ottenuta tramite una qualsiasi interpolazione delle osservazioni in serie storica che sia almeno due volte derivabile rispetto a *t*. Supponiamo inoltre che la funzione $f$ sia per ipotesi differenziabile due volte come funzione delle $Z_i(t)$. Definiamo con $Y(t)$ la funzione ottenuta applicando la $f$ alle interpolazioni delle osservazioni:

$$Y(t) = f\left(Z_1(t), Z_2(t), \dots, Z_k(t)\right) \qquad (13)$$

Avremo allora per il teorema sulla derivabilità delle funzioni composte (cfr. Analisi Matematica) che $Y(t)$ è derivabile rispetto a $t$ in quanto composizione di una funzione differenziabile applicata ad un vettore di funzioni derivabili.

Ritornando alla notazione $Z_t^i$ per indicare $Z_i(t)$, sia $V_h(Z_t^i)$ la variazione della variabile $Z_t^i$ di passo *h*, come definita in (3), e $\dot{Z}_t^i = \frac{dZ_i(t)}{dt}$ la derivata della variabile $Z_i(t)$ rispetto al tempo *t*, grazie alla derivabilità è utilizzare le stime conpolinomi di Taylor nel modo seguente.

Data la *f* saremo interessati a ricavare una funzione *g* tale che:

$$V_h(Y_t) = V_h\left(f\left(Z_t^1, Z_t^2, \dots, Z_t^k\right)\right) = g\left(V_h(Z_t^1), V_h(Z_t^2), \dots, V_h(Z_t^k)\right) \qquad (14)$$

Sappiamo che per la derivabilità di $Z_t^i$ vale la seguente stima al primo ordine

$$Z_{t+h}^i - Z_t^i = h\,\dot{Z}_t^i + o_i(h) \qquad (15)$$

E da questa stima possiamo ricavare la variazione di passo *h* in funzione della derivata

$$V_h(Z_t^i) = \frac{h\,\dot{Z}_t^i + o_i(h)}{Z_t^i}\,100 \qquad (16)$$

Saremo interessati alla relazione inversa ossia:

$$\dot{Z}_t^i = \frac{V_h(Z_t^i)\,Z_t^i}{100\,h} - \frac{o_i(h)}{h} \qquad (17)$$

Nell'ipotesi in cui la variabile $Z_t^i$ sia due volte derivabile, utilizzeremo anche la stima al second'ordine:

$$Z^i(t+h) - Z^i(t) = h\,\dot{Z}_t^i + \frac{1}{2}h^2\ddot{Z}_t^i + o_i\,(h^2) \qquad (18)$$

Ora, sappiamo dalla doppia derivabilità della funzione *f* come funzione di *t* che vale la seguente relazione:

$$f(\underline{Z}_{t+h})-f(\underline{Z}_t)=h\ <\nabla f,\dot{\underline{Z}}_t>+\tfrac{1}{2}h^2<\dot{\underline{Z}}_t,\nabla^2 f\cdot\dot{\underline{Z}}_t>+\tfrac{1}{2}h^2<\nabla f,\ddot{\underline{Z}}_t>+o\left(\|\underline{Z}_{t+h}-\underline{Z}_t\|^2\right)$$

dove gli elementi dell'equazione sono stati definiti nel paragrafo 1.2.
Mettendo a fattore il gradiente e semplificando otteniamo

$$f(\underline{Z}_{t+h})-f(\underline{Z}_t)=<\nabla f,h\dot{\underline{Z}}_t+\tfrac{1}{2}h^2\ddot{\underline{Z}}_t>+\tfrac{1}{2}h^2<\dot{\underline{Z}}_t,\nabla^2 f\cdot\dot{\underline{Z}}_t>\ +o\left(\|\underline{Z}_{t+h}-\underline{Z}_t\|^2\right)\qquad(19)$$

Dalla (17) e dalla (18) avremo :

$$\frac{V_h(Z_t^i)Z_t^i}{100}=Z_{t+h}^i-Z_t^i=h\,\dot{Z}_t^\imath+\frac{1}{2}h^2\ddot{Z}_t^\imath+o_i(h^2)$$

$$\frac{V_h(Z_t^i)Z_t^i}{100}=Z_{t+h}^i-Z_t^i=h\,\dot{Z}_t^\imath+o_i(h)\ \Rightarrow\ \dot{Z}_t^\imath=\frac{V_h(Z_t^i)Z_t^i}{100\,h}-\frac{o_i(h)}{h}$$

Sostituendo nella (19) otteniamo:

$$\frac{V_h(Y_t)Y_t}{100}=Y_{t+h}-Y_t==\ f(\underline{Z}_{t+h})-f(\underline{Z}_t)=<\nabla f,h\dot{\underline{Z}}_t+\tfrac{1}{2}h^2\ddot{\underline{Z}}_t>$$

$$+\frac{1}{2}h^2<\dot{\underline{Z}}_t,\nabla^2 f\cdot\dot{\underline{Z}}_t>\ +o\left(\|\underline{Z}_{t+h}-\underline{Z}_t\|^2\right)=\sum_i\frac{\partial f}{\partial Z_t^i}\left(\frac{V_h(Z_t^i)Z_t^i}{100}-o_i\,(h^2)\right)+$$

$$=\ \tfrac{1}{2}h^2\textstyle\sum_{i,j}\frac{\partial f}{\partial Z_t^i}\frac{\partial f}{\partial Z_t^j}\left(\frac{V_h(Z_t^i)Z_t^i}{100\,h}-\frac{o_i(h)}{h}\right)\left(\frac{V_h(Z_t^j)Z_t^j}{100\,h}-\frac{o_i(h)}{h}\right)+o\left(\|\underline{Z}_{t+h}-\underline{Z}_t\|^2\right)\qquad(20)$$

Divideremo l'equazione in due parti, la prima parte che rappresenta l'equazione che stavamo cercando e la seconda costituita da tutti i residui, che vengono accorpati in un'unica funzione *R*. Per semplicità utilizzeremo la notazione $f$ per indicare $Y_t$:

$$V_h(Y_t)=\tfrac{Y_{t+h}^i-Y_t^i}{Y_t^i}100=\text{per la (27)}=$$

$$=\frac{1}{f}\sum_i\frac{\partial f}{\partial Z_t^i}\ V_h(Z_t^i)Z_t^i+\frac{1}{200\,f}\sum_{i,j}\frac{\partial f}{\partial Z_t^i}\frac{\partial f}{\partial Z_t^j}V_h(Z_t^i)\,V_h(Z_t^j)\,Z_t^iZ_t^j\ +\ R(\nabla f,\nabla^2 f,\underline{Z}_t,h)=$$

$$=\frac{1}{f}<\nabla f,V_h(\underline{Z})\underline{Z}>\ +\ \frac{1}{200\,f}<V_h(\underline{Z})\underline{Z}^T,\nabla^2 f\,V_h(\underline{Z})\underline{Z}>\ +R(\nabla f,\nabla^2 f,\underline{Z}_t,h)=$$

$$=g(\,V_h(Z_t^1),V_h(Z_t^2),\dots,V_h(Z_t^k)\,)+R(\nabla f,\nabla^2 f,\underline{Z}_t,h)\qquad(21)$$

Si può notare come la (21) ha la forma dell'equazione che stavamo cercando a meno di un residuo $R(\cdot)$ da stimare. Tramite una semplice applicazione dell'algebra degli *o*-piccolo si può verificare che

$$R = o(h^2) + o\left(\|\underline{Z}_{t+h} - \underline{Z}_t\|^2\right). \qquad \square$$

## Riferimenti bibliografici

FUSCO N., MARCELLINI P., SBORDONE C. 1996. *Analisi Matematica 2*. Liguori.

BINI D., CAPOVANI M., MENCHI O. 1988. *Metodi numerici per l'algebra lineare*. Zanichelli.

BENEDETTI C. 1962. *Teorie e tecniche dei numeri indice*. Metron, No. 22, pp. 3-97.

## SUMMARY

### Decomposition of indexes and growth rates: focus on simple index numbers for ratio variables and a general formula.

This work shows the results of a study for defining a theoretical and analytical framework on the fixed base index numbers calculated on ratio variables, focusing on some of their properties to be considered when used. With the aim of providing hints to explain its dynamic, the general index is broken down into individual components with respect to any possible classification of units, and analyzed in relation to the dynamics of its numerator and denominator. The same breakdowns are spelled out for the calculation of the related growth rates. Furthermore, a general formula for decomposing growth rates based on the relation between indexes is presented.

Marco LATTANZIO, ISTAT, lattanzio@istat.it

# DECISION MAKING AND SAINT PETERSBURG PARADOX: FOCUSING ON HEURISTIC PARAMETERS, CONSIDERING THE NON-ERGODIC CONTEXT AND THE GAMBLING RISKS

Antonio Cappiello

## 1. Introduction

In 1713 Nicolaus I Bernoulli, in a correspondence with Pierre Rémond de Montmort, first identified a particular lottery game producing infinite expected gain[1]. However, the appellation of this mechanism as *Saint Petersburg Paradox* is coming from Daniel Bernoulli's contribution, entitled *Specimen theoriae novae de mensura sortis* (1737) in the *Commentaries of the Imperial Academy of Science of Saint Petersburg*. In this publication, Daniel Bernoulli clearly shows the intent of measuring the risk (*mensura sortis*) starting from some case studies and proposes a solution of the paradox stated by Nicolaus. Actually, the scope was to challenge the theoretical predominant paradigm of the expected value used for the risk assessment. The risk assessment analysis connected to the paradox is often forgotten because the attention is mainly focused on finding some convergence of the expected value.

## 2. Scope of the formalisation and decision making

Since theories are built to help us with the practical decisions, most readers may be not necessarily interested in pure mathematical demonstrations dealing with the problem of infinity which, furthermore, would have no impact on the limited human life cycle. In fact, it would be worthwhile to focus on some basic parameters in order to understand the risks of decision biases coming from some theoretical results. Moreover, the "tendency to infinity" of the expected value can be "balanced" with a counter-formal heuristic approach, which can underline the illusory possibility of  the infinite gain emerging of the paradox. The following

[1] "*Peter tosses a coin and continues to do so until it should land 'heads' when it comes to the ground. He agrees to give Paul one ducat if he gets 'heads' on the very first throw, two ducats if he gets it on the second, four if on the third, eight if on the fourth, and so on, so that with each additional throw the number of ducats he must pay is doubled. Suppose we seek to determine the value of Paul's expectation*".  Letter dated 9 September 1713 to P. de Montmort (correspondence of N. Bernoulli on St. Petersburg Game - Translated by R. J. Pulskamp, Xavier Univ, Cincinnati, OH. Jan 1, 2013).

quote by A. E. Newton could be useful to make some reflections on the infinity and the semantic of it:

"*Who was it who said, "I hold the buying of more books than one can peradventure read, as nothing less than the soul's reaching towards infinity; which is the only thing that raises us above the beasts that perish?" Whoever it was, I agree with him".* (A. E. Newton, 1921. A magnificent farce and other diversions of a book collector)

When we talk about infinity, sometimes we simply intend a huge number of something (*more books that anyone can read;* because only in this way we could let our *soul reach the infinity*). This may imply many reflections, for instance: provided that we have a *soul* (and this would already be in fact our intrinsic infinite quality), why we should employ large amount of finite things in order to reach the infinity that we already possess? Why possess for a personal scope something that you could never read? Anyway, how do we use these books? Do we simply look and admire the covers? Do we leave them to future generations? And, in the end, if we just perish like the *beasts* he cited, what sense could all this have? Therefore, this quote also contains many paradoxes and, I would say, as much confusion as the St. Petersburg paradox does. In the case of A. E. Newton's quote, a rational observer would say that the most common reason for being a collector is a compulsive self-satisfying behaviour driven by the emotions connected with the ever new objects coming into his possessions. Now, we should try to detach a little bit from the pure theory and ask ourselves some core questions: why do we use formalisation? Is it always useful to reason with infinite perspective? What is the context of the study? Does it help us to make useful choices? And, finally, are mathematical axioms a dogma?

## 3. Saint Petersburg paradox

What is the paradox about? The game consists in tossing a coin. The consecutive occurrences of tail events will produce a gain which value will exponentially increase as long as tails continue to be consecutively generated by each toss [2]. The game ends when head occurs. If we formalise these procedures and make some calculations, we realize that the game generates an infinite expected gain [3]. According to the original formulation of the paradox (note 1) the EV would be:

$$E(X) = \sum_{k=1}^{n} 2^{k-1} \cdot 2^{-k} \qquad (1)$$

However, if we suppose that the gain after the first tail occurrence is *2* Euro (instead of 1) and that the gain after *n* tails in a row will be $2^n$ Euro (instead of $2^{n-1}$), the substance of the paradox would not change much and we would follow and execute more smoothly some calculations. The expected value of the game will therefore be:

[2] $E(X) = \sum_{k=1}^{n} 2^k \cdot 2^{-k}$ for $k \to \infty$ it will produce an infinite expected gain (2)

[3] $E(X) = \sum_{k=1}^{\infty} 2^k \cdot 2^{-k} = \infty$ (3)

## 4. A very practical answer to Saint Petersburg paradox

If we want to "monetarise" at a certain point, we should interrupt the game. Therefore, the number of tosses should became finite and this would already undermine the expected infinite value which, as empirically observed, is logarithmically diverging for a large number of repeated games (see a simulation in fig. 3). Moreover, the expected value is also a questionable parameter for forecasting the most probable outcome, especially when the most remunerative events are associated to lower and lower probabilities[2]. Moreover, even if the number of tries becomes finite, the expected value - as we will see hereafter - could represent a suboptimal reference for the decision making process. Actually, there is not a real paradox but only a fallacy in the choice of the model to describe the empirical case. In fact, the biases generating the paradox seem connected on how we calculate the expected value (EV). A practical example of how a decision can be biased if it is taken only on information coming from EV is the game show *Deal or No Deal*. At a certain point of the game the player receives a money offer (usually inferior to the EV calculated on the remaining prizes) for ending the game and renounce gambling for further higher available prizes. Since the choice is only based on the EV, the player should always renounce because of the unfair proposal. However, it could be sometimes wise to accept the offer despite its being inferior to EV because of the high dispersion of the value of the prizes and the great incertitude connected with their probability to occur[3]. Actually, the expected value formula may be the key of the paradox because it fails to give fair practical information on how to assess a very uncertain stochastic context. In fact, the distribution of probabilities connected to the payoff in the St. Petersburg games is very asymmetric and therefore it could not be sufficiently described by the mean

[2] In other words, the EV is only a theoretical reference but may not always be a good parameter for making the best choice especially if the game is a stochastic variable with infinite possible results, it presents a high risk of low payoff and it is repeated only a few times.

[3] For example, if the remaining prizes are 1000, 1500, 2000, 2500, 3000, 99000, and therefore the EV=18166, it could be wise to accept an offer of 9000 because, in this game without repetition, it would be very risky to continue gambling to reach the highest available price (the probability of getting the best prize would only be 16,6 % while the chances of ending with less than the offered amount is 83,4).

(in this case the EV would be undefined) as in the normal distributions. If instead of the mean (EV) we consider the median value[4] (which could more fairly reduce the noise of highly skewed distributions), a fair price to enter the game should be of very few Euro.

No reasonable gain could be expected (at least in a finite number of games) considering the high magnitude of uncertainty about the occurrences of highest gains and their capacity to cover (in case of repeated games) all the previous loss in order to obtain a reasonable positive payoff.

Bernoulli himself proposed a solution for the paradox, but his attempt to resolve it with a utility function was not a "real solution" to the paradox in itself, even if it is a valuable effort to approach a theoretical concept in an empirical context. In fact, Bernoulli proposed a utility function[5] that considers the player's expected utility as a natural logarithmic function of the expected payoff. In other words, utility does not scale linearly with the payoff value but is logarithmically decreasing.

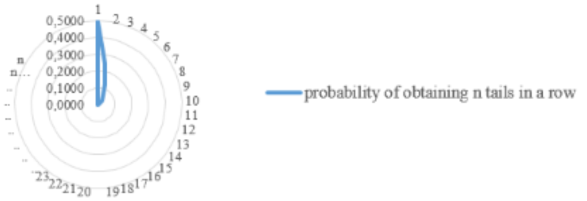
$$EU(X) = \sum_{k=1}^{\infty} \ln 2^k \cdot 2^{-k} \tag{4}$$

The problem with this function, beside the choice of its characteristics based on the psychological factors, is that we could always conceive another paradox that would not be explained by the ad hoc built function. If we suppose the payoff is $e^{2^k}$, the function is not resolving the paradox anymore[6].

$$EU(X) = \sum_{k=1}^{\infty} \ln e^{2^k} \cdot 2^{-k} = \infty \tag{5}$$

If we develop the matrix of probability connected with the increasing win associated with the consecutive tails occurrences, we would obtain a rapidly decreasing plot associated with increasing consecutive winning events (Fig.1).

**Fig.1 -** Probability connected to increasing win for consecutive tails occurrences



[4] In the case of the previous note example about the show *Deal or No deal*, the median (2250) could represent a better parameter for evaluating the convenience of the offer.

[5] V*alue must not be based on the price, but on the utility it yields. A gain of one thousand ducats is more significant to the pauper than to a rich man though both gain the same"* (D. Bernoulli 1788).

[6] Maybe, since the economic theory mainly aims at teaching us how to behave in uncertain situations in order to make the best possible choices, we could be more interested in a finite version of the game.

## 5. Some heuristic hints

If we consider the sum of the probabilities connected to the payoff emerging from all the possible series of tails in a row[7], we would have:

$$F(G) = \sum_{k=1}^{\infty} 2^{-k} \qquad (6)$$

Obviously, this is converging to 1. If we consider the cumulated probability of '*not obtaining tails in a row*', this can be described -with some approximations- by *F(L)*:

$$F(L) = \sum_{k=1}^{\infty} 1 - 2^{-k} \qquad (7)$$

This sum is diverging. With k going from 1 to *n*, we could rewrite it as:

$$F(L) = (1 - 2^{-1}) + .. + (1 - 2^{-n}) \ \textit{if } k \rightarrow \infty, n{=}\infty \ [9] \ F(L) = n - (2^{-1} + .. + 2^{-n}) = \infty$$

Since $(2^{-1}+\ldots+2^{-n})$ is equal to [6] and converging to 1 and *n* is diverging, the series F($L$) [7] is diverging. This is another way to say that our less desirable events are indeed most likely to occur than the desired ones[8]. Quite fortunate payoff will instead occur very rarely with a probability converging to 0 (see also Fig. 1). The expected value is an "average" of the gain we could expect; however, many other elements should be taken into consideration. In fact, the expected value is giving us the mean coming from all the possible occurrences with different gain and related probability. Since we normally do not trust on average value concerning any kind of distribution but we want to get additional information (e.g. standard deviation, dispersion index, skewness, etc.), there would be similar reasons to also consider the variability of the gains and their connected probabilities. Actually, if we consider the original formulation of the St. Petersburg lottery, we can deduce that the EV is not infinite but undefined. In fact, since the variance is not finite anymore[9], the strong law of large number will not apply. In the game, the EV is influenced by the outstanding gain connected to very rare occurrences. Why should we therefore only focus on the EV of this variable and undermine that there would be a high gain only with sufficient high repetitions of the game? What are we neglecting to consider?

[7] Starting from a payoff $2^1$ with probability $2^{-1}$.

[8] More and more fortunate events will occur less and less frequently while games that would generate a less desirable payoff will asymptotically stabilize to, at least, half of the total plays. The remaining n/2 plays would generate modest gains that would probably only balance the fees for entering the game

[9] If our variables are independent but not identically distributed then the average should converge to the EV $\bar{X}_n - E(\bar{X}_n) \rightarrow 0$ only if $X_k$ has a finite variance: $\sum_{K=1}^{\infty} Var(X_k)/K^2 < \infty$ See *Kolmogorov's strong law of large numbers* and Sen; Singer (1993). *Large sample methods in statistics.* Ch. & Hall

## 6. Some behavioural remarks: could we better perceive negative or positive consequences? Are our choices based on a time context?

One important factor to consider in the decision-making is the animal nature and its reactions to external stimuli (space and time context). A recent neuroscience experiment[10] reveals that there are many factors pushing animals to take suboptimal choice for a large reward even if this is very rarely delivered. Among these factors, there are two important elements to consider: (1) their insensitivity to risk because they are not able to evaluate the uncertainty of the prize; (2) the perception of a "loss" as a punishment (they consider the loss as frequently omitted reward instead of a very probable occurrence). Another important element is the rigidity induced by the habit to conduct an apparently good strategy seeking the maximum reward (underestimating the risk), which biases the correct formation of risk aversion in case of very unsure reward. This factor is also correlated with the pure temptation to gamble which may be predominant over any other factors. The attraction to rewards can generate positive reinforcements dominating the risk of punishment. As concerns human gambling, an important key is "the risk of losing" (losing in a negative game session what was gained in a previous favourable game session[11]) that is different from the "failure to win" representing also the frustration caused by the lack of the expected gain. Focusing the attention on the "failure to win" only considers the frustration for missing an expected reward but does not take into account the risk of the negative events (in other words, the session game was considered misfortunate but not risky). The choice mechanism is therefore affected by many factors but obviously all these decisions are based on a finite segment of time in which the subjects reinforce a habit in order to reach his perceived optimal choice (we cannot therefore consider the context as an infinite space).

## 7. Are we neglecting to consider the non-ergodicity?

Recently, interesting observations were made on the expected value and its lack of capacity to determine the price of the Saint Petersburg game due to non-ergodic property of the time averages. The expected value formula implies that the time averages of the considered games are equal to the average of the entire system (ergodicity). Evaluating the ergodicity of a system is a very crucial element especially when conducting physics tests where the sample results of the experiments should generate reliable universal implications. In other words, ergodicity supposes that a system (probabilistic ensemble) observed for a sufficient

[10] Paglieri et al. 2014, *Nonhuman gamblers,* Frontiers in Behavioural Neuroscience (8:33).

[11] Zeeb et al., 2009.

period of time *t* is representative of all the possible states of the universe (sample space) in a way that the relative frequencies of selected sample coincide with the postulated predicted probabilities associated with the possible occurrences of the considered universe [11]. This means that we have to verify along with time our theoretical assumptions contained in the expected value formula [2]. In case the time averages coincide with ensemble average [10] the system is ergodic, otherwise it is a non-ergodic system.

$A = \langle A \rangle \quad (10)$ A is the time average and ⟨A⟩ the ensemble average. The consequence of ergodicity is that the variable of interest do not change overtime and even if very small fluctuations are observed, in a sufficiently long period, they do not influence the variables (this means that has not relevant effect on the ensemble system).

More generally, in continuous context, the condition of ergodicity can be defined as:

$$\lim_{t\to\infty} F\,(x,t) = P(x) \tag{11}$$

That is to say, that the relative frequencies *F* observed over time tends to the postulated probabilities *P* governing the ensemble system. In our case, the ensemble average can be described by the expected value defined by the Bernoulli game [2] while the time average is the average payoff coming from the experiments conducted over time [fig. 3]. The issue of ergodicity has been analysed in detail especially by Peters[12], and an immediate visualisation (proposed by Koelman[13] 2012) of the differences in averages can be useful to smoothly understand the context of this issue. If we consider a mechanism similar to the *Steinhaus Sequence*, we can build a recursive sequence based on the powers of 2 which follows a consequent recursive deterministic pattern. In the matrix, starting from the first row, each alternate sequence of empty cells is filled with the number corresponding to 2 powered to a number equal to the previous row number[14] (i.e. the empty cells of the 4th row are filled with $2^{4\text{-}1}$ and so on, see numbers in bold in Fig 2).

**Fig. 2 -** Building a matrix of the power of 2 using a *Steinhaus Sequence principle*

| | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | .. |
| *2* | . | *2* | . | *2* | . | *2* | .. | . | *2* | *2* | . | *2* | . | *2* | | .. |
| *2* | ***4*** | *2* | . | *2* | ***4*** | *2* | .. | *2* | ***4*** | *2* | . | *2* | ***4*** | *2* | | .. |
| *2* | *4* | *2* | ***8*** | *2* | *4* | *2* | .. | *2* | *4* | *2* | ***8*** | *2* | *4* | *2* | | .. |
| *2* | *4* | *2* | *8* | *2* | *4* | *2* | ***16*** | *2* | *4* | *2* | *8* | *2* | *4* | *2* | | .. |

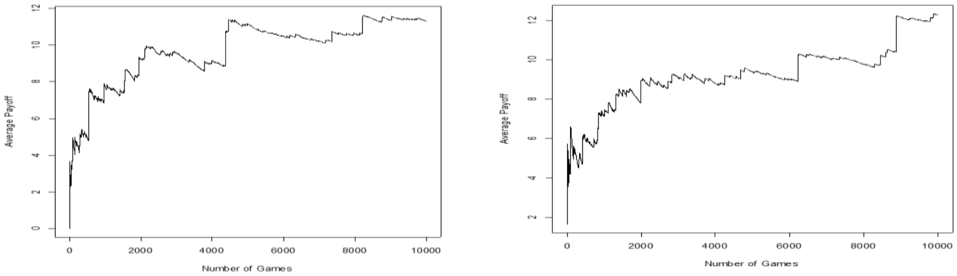
[12] Peters O. 2011, *The time resolution of the St. Petersburg Paradox*, Ph. Trans. Royal Society, n. 369

[13] This very useful hint was proposed by Johannes Koelman, *Statistical Physics Attacks St. Petersburg: Paradox Resolved*, on Science 2.0 (Scientific Blogging) 18th November 2012.

[14] This sequence is similar to the distribution of probability of the St. Petersburg game. In fact, if you casually choose a number from this matrix, you would have probability ½ to extract a 2, ¼ to extract a 4 and *1/n* to extract *n*. We could easily calculate the averages for the first *2, 3, 4..n* numbers located in the corresponding rows (the first 2 numbers in the 2nd row, the first 3 numbers in the 3rd row, etc.).

We can observe that the average is not independent (at least for a considerably large number of games) from the number of games played. These averages are finite and fluctuate considerably for each set of experimental games, although the pattern shows an infinite logarithmic increment. Since the average of all *time averages* does not converge to a finite value in the long run, the assumption of an ergodic context is confuted and time averages can not substitute the entire *space average* (as calculated with the expected value formula). Therefore, the non-ergodicity implies that time is a creative factor influencing the positive trend of time averages.

**Fig. 3** – Two simulated sets of average payoff for 10,000 games



Source: author's simulation using R (for source code, see the references mentioned at the end of the article)

In order to establish a fair price, we could repeat the game n times and observe the simulated pattern of payoff; therefore, we could use the average payoff as an indicator of the general tendency and propose a price balanced on it (naturally, for each different set of repetition of the game the price would be different). The debate on non-ergodicity therefore becomes interesting if we consider the finite available time and if we are supposed to play only a limited number of games.

## 8. Probabilities to obtain a significant payoff

A significant gain should happen if a certain number of tails occur in a row. The main problem at decision level is to be aware of the probability that this event will happen and the respective probability that this would not happen (this is relevant because every time our suitable gain is not compensating our initial investment, we are facing a loss). However, more generally, if we consider that the probability to obtain $n$ tails in a row is $P(n) = (2^{-1})^n,$ it is immediate to recognise that a significant gain is based on a very low probability while the opposite and undesired events implying low gain or loss are likely to occur with a higher probability. The combination of expected gain in an infinite repetition of the game is biasing our decision because it lacks two main elements: the empirical limited duration of the game and its concrete management structure (time life resources that we are likely going to dedicate to this game and financial resources of the gambler and the bookmakers). Even if the probability of obtaining a significant number of tails in a row is only a partial consideration of the overall context of the game, this may be the main evidence that could lead to a wise decision in the short run, while the theoretical paradox is based on a supposed infinite context. Computer based simulation can easily confirm this kind of reasoning. Nowadays we dispose of free available tools to perceive the biases of the human expectations tending to over valuate fortunate events with low occurrence probability (see *references and further reading* section). The gamble booming in our society seems based, besides the psychological and sociological factors, on the incorrect way to calculate the occurrence of favourable events and the mechanism that generates them[15]. At the end of the XX century Ambrose Bierce said that lottery is a tax on people who are bad at mathematics. Nowadays this could be the case of only some gamblers, but it seems that most of them are quite aware of their systematic position of inferiority to the bookmakers even because of their empirical findings after some repeated bets. However, the complete scientific awareness would maybe reduce in a more aware and radical way the bias of the choices and not only as concerns gambling. Actually, in less than a century, we have suddenly awakened in a world where every tool we use is completely parametrised and probabilistically set. Setting a price for a plane ticket or booking a hotel room, sending customised commercial messages on internet tracking the users' behaviour, etc. These models certainly do not set infinite and asymptotic parameters when evaluating our daily life behaviours. In the next paragraph, we make some reflections on how this affects our lives.

## 9. Assessing the risk of everyday life decisions: learning about failure (ex-post adjustments) or log-frame evaluation (ex-ante and instant adjustments)?

Compared to the beginning of the XVIII century many things profoundly changed. The widespread awareness about general concepts related to probability and the computer literacy should switch the debate to a completely different level than a pure academic debate among scientists. Nevertheless, despite the application of the probability models and the computerisation of almost every structured human life management or decision process, we still face a society in which some decisions (extreme gambling, compulsive buying, etc.) are biased because of the lack of consideration about the risk assessment. This is part of the human nature which perceives the risk very differently according many different contexts, time

[15] A risk assessment literacy would not only be an awareness of the formal description of the phenomenon but especially a reasonable way to make some decisions.

and emotional reactions[16] (par. 6). The *Saint Petersburg paradox*, to a certain extent recalls the martingale mechanism but the inner process of the game (theoretical infinite gain) makes us focus more on the infinite expected value than all the other factors that in the medium-short run could affect our desired perspective of gain. However, also in this case we could demonstrate some inconsistency limiting our analysis to the available time for the gambler and the possible occurrences he would face with their consequences. Empirically and theoretically this would not be worth much in term of human resources dedicated to it. Nevertheless, if we consider other fields of application such as physics or computational models including patterns with almost infinite cycles, this analysis then assumes a different nature and it remains very useful because of the non-ergodic implications and their consequences. In this case, we consider a completely different structure and context than the initial framework connected to human gambling repeated to infinity. Recently the concept of *learning about failure* became very fashionable with a huge literature applying it to different fields of the social sciences; the empirical counterproof and the awareness of the positive expectation fallacy could bring some consciousness of the differences among theoretical and real effects of infinite expected value applied to different contexts. Nevertheless, in the context of extreme gambling, the only *learning about failure* could lead to very negative results because your failure would likely assume the form of a bankruptcy from which you could hardly learn and start over again avoiding your past errors. If someone should still have some scepticism, he would not to certainly put trust on famous empirical analogical experiments anymore (e.g. Buffon[17]). Nowadays you may want to empirically experience the consequences of playing a consistent number of times with simulated random variables games which are broadly available also on line (some references and example are indicated in the references and further reading section) or generate computed results with simple programming (e.g. java script or even setting parameters with simple spread sheets). Anyone experimenting the simulations would be happy about avoiding losing time and other resources on the research of empirical proofs.

[16] How would you feel about someone saying: '*in a martingale game, at some point, you will surely win whatever amount you desire, provided that you continue betting in order to reach a positive payoff (compensating all previous losses)'.* This is theoretically true only if you have infinite time and money.

[17] In 1777, Buffon conducted an experiment repeating it for 2048 times and found out that the number of tails in a row corresponding to 1, 2, 3, 4, 5, 6, 7, 8, and 9 had a frequency of 1061, 494, 232, 137, 56, 29, 25, 8, and 6 respectively. The average payoff was 4.91. *Essais d'arithmétique morale* in *Supplément* à l'*Histoire Naturelle*, V. 4, Imprimerie Royale, Paris. At page 394 Buffon describes the experiment he made with the help of a kid: *J'ai donc fait deux mille quarante-huit expériences sur cette question, c'est-à-dire j'ai joué deux mille quarantehuit fois ce jeu, en faisant jeter la pièce par un enfant.*

Indeed, since compulsive gambling is becoming a social problem, it could become a legal requirement for bookmakers to put free simulation machines at the disposals of the clients (foreseeing some symbolic form of incentive to test it before playing with real money), this would be more effective than a simple disclaimer only mentioning the risks leading to pathological gambling and the list of the probabilities connected to the potential gains. Finally, why is the study of this paradox still so important after two centuries? Because it has to do with the decision-making and risk assessment[18]. The considerations on the time factor and the non-ergodic context of the experiment reveal new useful elements for assessing social and scientific phenomena. This approach considers the pattern of payoff in finite plays, their related *time average* (par. 7) and the behavioural components of the choices (par. 6). The non-ergodic property of the St. Petersburg system underlines the correlation of the average payoff with the time factor and the misperception of the probability of possible negative events computed and attenuated in the *ensemble system.* However, all these factors have to be inserted in a variegated framework connected to human nature and all the elements that may affect the context. Therefore, an apparently purely mathematical problem is instead involving a very horizontal multidisciplinary approach. For that reason I acknowledge T. Parisi (Engineer), for the clues on IT programming and the important reflections on the capillary use of technology in our daily life. As concerns the psychological factors of choices, A. Carstoiu (medical doctor at *Clinic de Psihiatrie* in Bucharest), showed me the mechanism affecting human choice while S. Bosnic (researcher at Croatian Veterinary Institute) made me interesting comments on animal behaviours. I am also grateful to G. Celeste, E. Morandi and P. Pasqualis (directive members of the Italian Notariat) with whom we proposed very challenging research projects that allowed the knowledge exchange with Academics at the University of Moscow (HSE) and Saint Petersburg[19]. It seems therefore not a chance that in this last city the *Imperial Academy of Science* first published D. Bernoulli's work and his first proposed solution to solve the paradox. Indeed, in Russia, there was the opportunity to present some of our results at multidisciplinary conferences for the analysis of the economy and the society and we became keener on philosophical and conceptual aspects of the economic analysis. The paradox is one of the most challenging exercises because it involves

[18] Which are key factors of financial and insurance markets but, to some extent, with all kind of choices.

[19] Among them: the Scientific C. of Higher School of Economics of Moscow (V. Mkhitarian,V. Sirotin, M. Arkhipova and L. Rodionova), D. Raskov (St. Petersburg University), O. Ozerova (Sociological Inst., Russian Academy of Sciences, St. Petersburg), A. Nemtsov (Moscow Research Inst. of Psychiatry), M. Markov (St. Petersburg Univ.), J. Nye (George Mason Univ.), E. Poelmans (Univ. Leuven), K. Storchmann (NY Univ.), R. White (Univ. of Alabama) and J. Leitzel (Univ. of Chicago).

the capacity of rethinking about all the standards given for guaranteed in the scientific assessment. These arguments show that it is important to consider the irreversibility of the choices (in a given finite context), the behavioural factors, and the reasonable expected gain assessed from different point of view (considering the not-ergodic context and the EV capacity to properly describe the empirical outcome).

## References and further readings

BERNOULLI D. 1954 [1738]. *Exposition of a New Theory on the Measurement of Risk*, Econometrica, 22; pp. 23–36.

CAPPIELLO A. 2014. *World Bank Doing Business Project and the statistical methods based on ranks: the paradox of the time indicator,* RIEDS, vol. 1, January-March, pp. 79-86.

PAGLIERI F. et al. 2014, *Nonhuman gamblers*, Behavioural Neuroscience (8:33).

PETERS O. 2011, *The time resolution of the St. Petersburg Paradox*, Philosophical Transactions of the Royal Society, n. 369, pp. 4913-4931.

VAN DEN ENDE K. 2016, *Durf Falen*. Lannoo.

BERN UNIVERSITY OF TEACHER EDUCATION. GameGrid, Java (on line simulation of St. Petersburg game and related Java code) *java-online.ch/gamegrid*

KNILL O., Matematik, *mathematik.com/Petersburg* (Harvard on line simulation).

KOELMAN J. 2012, *Statistical Physics Attacks St. Petersburg: Paradox Resolved*, on Science 2.0 (Scientific Blogging) 18th November.

## SUMMARY

The *Saint Petersburg Paradox* is still a contemporary issue because of the great impact on the probabilistic theory and decision-making. This article proposes some hints on avoiding the trap of the infinite EV. The highly stochastic mechanism and its EV have always to be contextualized in the limited period where we take our choices taking into account all possible limitations deriving from the theory (including the non-ergodic features and some inappropriate consequences we may attribute to the EV). This *contextualisation* is one of the most important factors to consider especially when we deal with infinite quantity coming from models that may misrepresent our field of application and therefore generate paradoxes.

Antonio CAPPIELLO, Economist, National Council of the Civil Law Notaries (Consiglio Nazionale del Notariato) and advisor at International Cooperation Center for Statistics ICstat. *cnn.acappiello@notariato.it; economics@live.co.uk*

**SOCIETÀ E RIVISTA ADERENTI AL SISTEMA ISDS**
**ISSN ASSEGNATO: 0035-6832**

*Direttore Responsabile:* Dott. CLAUDIO CECCARELLI

Iscrizione della Rivista al Tribunale di Roma del 5 dicembre 1950 N. 1864

Associazione all'Unione Stampa Periodica Italiana

TRIMESTRALE

*La copertina è stata ideata e realizzata da Pardini, Apostoli, Maggi* p.a.m.@tin.it – *Roma*

Stampato da CLEUP sc
“Coop. Libraria Editrice Università di Padova”
Via G. Belzoni, 118/3 – Padova (Tel. 049/650261)
www.cleup.it

# ATTIVITÀ DELLA SOCIETÀ

## A) RIUNIONI SCIENTIFICHE

XXXVII La mobilità dei fattori produttivi nell'area del Mediterraneo (Palermo, 15-17 giugno 2000).

XXXVIII Qualità dell'informazione statistica e strategie di programmazione a livello locale (Arcavacata di Rende, 10-12 maggio 2001).

XXXIX L'Europa in trasformazione (Siena, 20-22 maggio 2002).

XL Implicazioni demografiche, economiche e sociali dello sviluppo sostenibile (Bari, 15-17 maggio 2003).

XLI Sviluppo economico e sociale e ulteriori ampliamenti dell'Unione Europea (Torino, 20-22 maggio 2004).

XLII Sistemi urbani e riorganizzazione del territorio (Lucca, 19-21 maggio 2005).

XLIII Mobilità delle risorse nel bacino del Mediterraneo e globalizzazione (Palermo, 25-27 maggio 2006).

XLIV Impresa, lavoro e territorio nel quadro dei processi di localizzazione e trasformazione economica (Teramo 24-26 maggio 2007).

XLV Geopolitica del Mediterraneo (Bari, 29-31 maggio 2008).

XLVI Povertà ed esclusione sociale (Firenze 28-30 maggio 2009).

XLVII Un mondo in movimento: approccio multidisciplinare ai fenomeni migratori (Milano 27-29 maggio 2010).

XLVIII 150 anni di Statistica per lo sviluppo del territorio: 1861-2011. (Roma 26-28 maggio 2011).

XLIX Mobilità e sviluppo: il ruolo del turismo. (San Benedetto del Tronto, 24-26 maggio 2012).

50esima Trasformazioni economiche e sociali agli inizi del terzo millennio: analisi e prospettive (Università Europea di Roma, 29-31 maggio 2013).

LI Popolazione, sviluppo e ambiente: il caso del Mediterraneo (Università Federico II di Napoli, 29-31 maggio 2014).

LII Le dinamiche economiche e sociali in tempo di crisi (Università Politecnica delle Marche, 28-30 maggio 2015).

LIII Mutamento economico e tendenze socio-demografiche tra sfide e opportunità (Università degli Studi Internazionali di Roma, 26-28 maggio 2016).