Turning big data into smart data: two examples based on the analysis of the Mappa dei Rischi dei Comuni Italiani

Trasformare i big data *in* smart data*: due esempi di analisi della Mappa dei Rischi dei Comuni Italiani*

Oleksandr Didkovskyi, Alessandra Menafoglio, Piercesare Secchi, Giovanni Azzone

Abstract The recently presented *Mappa dei Rischi dei Comuni Italiani* is a freely accessible web portal, implemented by ISTAT, which provides integrated data on different natural risks in Italian municipalities together with socio-economic and demographic data. We here illustrate two paradigmatic examples where the big data of the *Mappa* are transformed into smart data using advanced methods for descriptive statistics, thus providing interesting insights on local patterns and regional trends in terms of building stock vulnerability and social and material vulnerability.

Abstract *La Mappa dei Rischi dei Comuni Italiani è un portale web pubblico, di recente reso operativo da ISTAT; per ogni comune italiano esso fornisce in modo integrato dati relativi ai rischi naturali insieme ad indicatori socio-economici e demografici. In questo lavoro illustriamo due esempi di analisi nelle quali i* big data *della Mappa sono trasformati in* smart data *utilizzando approcci avanzati di statistica descrittiva. Gli esempi forniscono interessanti visioni, sia a livello locale che nazionale, sulla vulnerabilità del patrimonio edilizio italiano e sulla vulnerabilità sociale e materiale del paese.*

Key words: Object oriented data analysis, compositional data, high-dimensional descriptive statistics

---

Oleksandr Didkovskyi
MOX, Department of Mathematics, Politecnico di Milano *and*
Center for Analysis Decision and Society, Human Technopole, Milano.
e-mail: oleksandr.didkovskyi@polimi.it

Alessandra Menafoglio
MOX, Department of Mathematics, Politecnico di Milano.
e-mail: alessandra.menafoglio@polimi.it

Piercesare Secchi
MOX, Department of Mathematics, Politecnico di Milano *and*
Center for Analysis Decision and Society, Human Technopole, Milano.
e-mail: piercesare.secchi@polimi.it

Giovanni Azzone
Department of Management, Economics and Industrial Engineering, Politecnico di Milano.
e-mail: giovanni.azzone@polimi.it

1 Introduction

On the 18th February 2019 the Department Casa Italia of the Italian Government presented to the public the *Mappa dei Rischi dei Comuni Italiani*[1] (MRCI). The freely accessible web portal, implemented by ISTAT, provides integrated information on different natural risks in Italian municipalities - such as earthquake, flooding, landslide, volcano eruption - in conjunction with socio-economic and demographic data. It offers the possibility of viewing and downloading indicators, charts and maps, together with guided interactive features for data searching and filtering.

The Casa Italia task force[2] was established in 2016 to develop a plan for housing and land care aimed at better protection of citizens, and public and private goods. The goal was to define the constituent elements of a national policy for the promotion of housing safety. Quality of living was identified as of primary importance for the mission of Casa Italia, with a particular emphasis on policies for the promotion of security of residential buildings against natural risks. The key idea was that of a multi-hazard approach to risk, focusing on the security of places where people live which are instrumental to their individual safety. As part of the proposed action plans, the MRCI aims to create a widespread awareness of the fragility of the Italian territory. Despite fragmentation, dispersion and diversity of the available databases, an information platform was built to allow a homogeneous and integrated view of the natural risks within the Italian territory.

In the first stage of the project, Casa Italia aimed at integrating and enhancing the rich information on natural risks already available as the result of numerous and intense research activities carried on by several local and national institutions (ISTAT, INGV, ISPRA, ENEA, CNR, MIBACT). Part of this action was devoted to identify the sources of information and the data bases allowing for a unified and integrated vision of the natural risks insisting on the Italian territory, with particular reference to the three factors that compose risk, i.e., hazard, vulnerability, exposure. Due to the mission of Casa Italia, the survey was limited to databases that (a) were elaborated by official and national research institutes, (b) had coverage of the entire national territory, and (c) had a spatial resolution sufficient to allow identification and comparison of local specificities. With reference to the latter point (c), the municipality (i.e., *Comune*) was identified as the smallest spatial statistical unit for the actual analyses. This choice is motivated by the need of integrating and fusing data from different sources, with different spatial resolutions, and originally generated for different aims. For instance, at the national level coverage, the municipality represents

[1] http://www4.istat.it/it/mappa-rischi

[2] The task force 'Casa Italia' of the Italian Presidency of the Council of Ministers was established on September 23, 2016. Its members were Giovanni Azzone (Project Manager and Scientific Director), Massimo Alvisi, Michela Arnaboldi, Alessandro Balducci, Marco Cammelli, Guido Corso, Francesco Curci, Daniela De Leo, Carlo Doglioni, Andrea Flori, Manuela Grecchi, Massimo Livi Bacci, Maurizio Milan, Alessandra Menafoglio, Pietro Petraroia, Fabio Pammolli, Davide Rampello, Piercesare Secchi. The 3rd of July 2017, the Italian Presidency of the Council of Ministers estabilished 'Casa Italia' as one of its departments, committed to the prevention against natural risks (http://www.casaitalia.governo.it/it/). The authors acknowledge the task force 'Casa Italia' for the scientific discussions that were inspirational to the present work.

the smallest administrative unit for which aggregated data on the vulnerability of residential buildings are today available, and these data come from the last national census of 2011. The dataset consists of observations taken at 7983 Municipalities, updated at 2018. Additionally, information about the same indices aggregated by provinces and regions is also provided.

To illustrate the richness of MRCI, in this paper we report on two analyses, which we consider as paradigmatic examples of the application of non trivial statistical descriptive methods and algorithms aimed at transforming big data into smart data, which are then made openly available to policy makers and the citizens.

The first analysis was developed as part of the Casa Italia project [3] and regards the age of the Italian building stock. For each Italian municipality, MRCI reports the distribution of the age of the buildings, grouped in $p = 9$ non overlapping time intervals. These object data are compositional [1]; for their analysis we should embed them in a proper space, which we take to be the simplex in R^8 endowed with the Aitchison geometry [1]. For the sake of this illustration and ease of visualization, we shall represent these data when the original classes are aggregated into 3 classes of ages, and thus represent the data in the simplex embedded in R^2. Within this space we explore data variability by performing a suitable PCA [1]; the projection of the age distributions along the main directions of their variability as object data, offers a powerful representation for the understanding of age of the Italian building stock over the entire national territory.

The second example considers the *Indice di Vulnerabilità Sociale e Materiale*[3] (IVSM) which is a quality of life index computed by ISTAT for each Italian municipality and reported in MRCI. Fort this analysis we consider the Italian provinces as statistical units. For each province we look at the distribution of IVSM among the municipalities it is composed of. Location and scale of these IVSM distributions – not to mention their quantiles – are of course of great interest to the policy maker, indicating respectively the degree of social and material vulnerability of the province and its heterogeneity among the municipalities. Hence an exploration of the IVSM distributions which treats each of them as a whole data object seems appropriate, as opposed to a more naive approach which would work separately on their mean, their standard deviations or their quantiles. The final goal of the analysis is to cluster provinces according to similarity of the distribution of IVSM among their municipalities. This will offer to the policy maker, at the national level, a picture of the country in terms of homogeneous macroregions, characterized by similar issues in terms of social and material vulnerability. In the next two sections we illustrate the two examples. Final comments will close the paper.

2 Compositional analysis of the age of the building stock

The age of a building is a key element to determine its vulnerability to seismic events, as it is directly associated with the seismic regulations in force at the time of the construction, as well as with the advance of building technologies. For the pur-

[3] http://ottomilacensus.istat.it/documentazione/

pose of this illustration, we shall consider the distribution – within municipalities – of the age of the building stock in the following three classes: before 1919, between 1919 and 1980, after 1980. These classes are representative, respectively, of extremely old buildings, buildings of medium age, and buildings erected under the more recent regulations in terms of seismic risk (for further details, see [3]). These data are compositional in their nature, they belong to the simplex embedded in R^2 which we endow with the Aitchison geometry [1]. Accordingly, we perform a compositional principal component analysis (PCA), and we then explore the directions of the simplex along which the dataset displays its maximum variability.

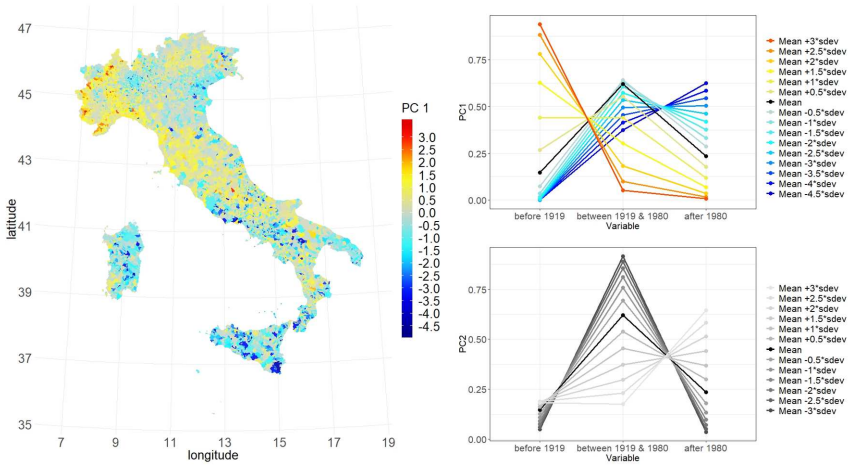

Fig. 1 Standardized scores along first compositional PC of the distribution of building stock in terms of age (left) and variation of the compositions along the first and second compositional PCs.

The left panel of Fig. 1 displays the scores along the main mode of variability of the dataset, which captures 86% of the total variability. The right panel of Fig. 1 displays the variation of the composition of the building stock when moving along the first PC (top panel) and the second PC (bottom panel), to support interpretations. High-scores along the first PC are representative of municipalities with a prevalence of very old buildings over the most recent ones and viceversa. Spatial patterns of high scores are particularly visible along the Apennines in central Italy, in Liguria and in the outer parts of Piemonte, as well as in some parts of Friuli Venezia Giulia. The score values along the first PC thus provide an indication of the overall degree of vulnerability of the building stock with respect to the Italian mean, and suggest to the policy maker the areas that may deserve specific interventions for the promotion of security of buildings against seismic risk.

3 Symbolic analysis of IVSM

MRCI reports different indicators on the vulnerability and resilience of Italian municipalities, taken from the socio-economic perspective. We here focus on IVSM,

which is given at the scale of municipality, and ranges in [90,120], higher values being associated with higher vulnerability. Although data in MRCI are at the scale of municipality, we consider their aggregation at a lower level of resolution – the province – to provide views of the local social vulnerability, as well as of the homogeneity of the indicators on the administrative unit; clearly, other aggregations would be possible (e.g., regional). These multi-scale views are relevant and are made possible by the MRCI. They allow the citizen, as well as the decision maker, to perform analysis and evaluations at a micro-scale or at an aggregated scale. Having grouped the micro-data according to the province, we consider as data object the probability density function (PDF) of IVSM within the province, as estimated by kernel smoothing. The location of the PDF provides indications of the overall vulnerability of the area, whereas the scale of the PDF indicates the regional homogeneity. We here aim to identify cluster of provinces characterized by a similar degree of vulnerability and homogeneity. The analysis and clustering of PDF data requires to properly define a notion of similarity between the PDFs of IVSM estimated at different provinces. As in the case discussed in Section 2, a Euclidean metric (or L^2 metric) would not correctly represent the data constraints; instead, metrics for distributional data are more appropriate and should be preferred. The Wasserstein metric is widely used in Symbolic Data Analysis (SDA, [2]) as a measure of dissimilarity among distributional data, and has insightful interpretation in terms of optimal transport. As an alternative, a functional Aitchison geometry (a.k.a., Bayes Hilbert geometry, [4]) may be used instead. For the sake of brevity, in this illustration we shall focus on the former metric only.

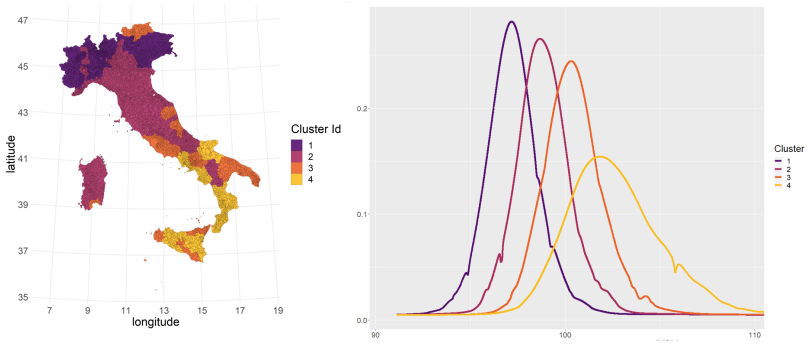

Fig. 2 Map of Italy where provinces are colored according the associated cluster (left) and centroids of the clusters of the IVSM PDFs (right).

Fig.2 reports the results of hierarchical clustering based on a Wasserstein metric with Ward linkage, having set to 4 the number of clusters. The left panel of Fig.2 reports a map of the identified clusters, whereas the right panel reports the centroids of the cluster, computed as Frechét mean of the PDFs within the clusters. These results show that the four clusters differ for both location and scale. The first cluster – associated with some of the Northern provinces – is characterized by an overall

low social and material vulnerability and higher regional homogeneity. Moving to the following clusters, one may observe increasing vulnerability and regional heterogeneity; a North-South trend is clearly visible. The province of Bolzano stands in clear contrast with the neighboring provinces, suggesting a local outlyingness. These analysis can support national and local administrators to evaluate the social and material vulnerability of the population and its possible resilience in response to a natural event.

4 Conclusion and further discussion

MRCI provides a rich and integrated framework to allow for multi-scale analyses on the natural and social risks insisting on the Italian territory. We here embrace the viewpoint of object oriented data analysis ([5]), and recognize as key elements of the analysis the identification of the *data object*, the choice of its geographical scale of reference and of the most appropriate geometry for its mathematical representation – choices that should indeed be guided by the goal of the analysis. In fact, turning *big data* into *smart data* inherently requires a strong interplay of advanced statistical method and experts' knowledge, to provide meaningful summaries and insights, and particularly to identify the *object* and the *objective* of the analysis, which still remain of primary importance even in the data deluge era. As paradigmatic examples of this methodological approach, we illustrated two object oriented analyses, where the use of advanced statistical method for aggregated data served the purpose of getting insights on local patterns and regional trends. These analyses may be extended to provide further views and predictive models, with the final aim of increasing the understanding and the awareness of the complex and multi-scale character of the social and natural risks insisting over the country, and, ultimately, to support the decision making of citizens, local administrators and policy makers.

References

1. Pawlowsky-Glahn, V., Buccianti, A., Compositional Data Analysis: Theory and Applications. John Wiley & Sons, Chichester (2011).
2. Billard, L., Diday, E., Symbolic Data Analysis: Conceptual Statistics and Data Mining. John Wiley & Sons, Chichester (2007).
3. Presidenza del Consiglio dei Ministri, Struttura di Missione Casa Italia, Rapporto sulla Promozione della sicurezza dai Rischi naturali del Patrimonio abitativo, http://www.casaitalia.governo.it/it/approfondimenti/rapporto-sulla-promozione-della-sicurezza/ (2017).
4. Van den Boogaart, K. G., Egozcue, J., Pawlowsky-Glahn, V. Bayes Hilbert spaces. Aust. N. Z. J. Stat., Vol. 56, No. 2, 171-–194 (2014).
5. Marron, J. S., Alonso, A. M.: Overview of object oriented data analysis. Biometrical Journal, Vol. 56, No. 5, 732—753 (2014).