AutoSynth Index: A Synthetic Indicator for Socio-Economic development based on Autoencoders

Giulio Grossi [1], Emilia Rocco[1]

[1] Department of Statistics, Informatics and Computer Science, University of Florence
(e-mail: giulio.grossi@unifi.it, emilia.rocco@unifi.it)

Abstract: In this work, we propose a novel use for neural networks to build socioeconomic indicators, encoding a possible large information set, within single or multiple synthetic indexes, we call this proposal AutoSynth. In particular, we encode such information using an autoencoder, a neural network method to represent in a lower dimensionality space a matrix of features. We apply such a method to the evaluation of socio-economic developments of suburban areas in Florence, and we test the performance of our model against some golden standard methods using a stress test.

Keywords: synthetic indicators, composite indicators, autoencoders, neural network, unsupervised learning

1 Introduction

Composite indicators are statistical measures that combine a set of elementary (or individual) indicators into a single measure of a complex phenomenon, such as the Human Development Index(HDI) or the Environmental performance index (EPI). See Commission *et al.*, 2008 for an account on the construction of synthetic indicators. Recently, Greco *et al.*, 2019 presents a review of the existent literature, focusing on the main goal of indicators construction and on the open challenges. The primary goal of a synthetic indicator should be the transmission of the information contained into each elementary indicator, with the lowest possible loss of such data. Moreover, such indicators rely on making transparent ranking that allows for spatial and temporal comparison between units and therefore are particularly suited to keep track of improvements in complex phenomena. With wider and more detailed sources of information, larger datasets are employed and feature extraction techniques are needed for accounting the amount of information that is considered. Golden standard approaches employ weighted averages or geometric averages to extract a single index from a matrix. An example is the Adjusted Mazziotta-Pareto index (AMPI) Mazziotta & Pareto, 2018, a novel synthetic indicator for measuring well-being. These methods are very transparent, yet it is not completely clear what should be the weights accounted for, and often strong theoretical knowledge is required. This task could become very difficult in presence of large datasets, where describing the relationships between variables could be cumbersome. Unsupervised learning approaches for constructing composite indicators have been deployed during the last 30 years, such as Principal Component Analysis and Factor Analysis, see Greco *et al.*, 2019 and Commission *et al.*, 2008 for some review and comments. In this work, we propose a novel unsupervised framework for developing synthetic indicators. Exploiting modern methods for data analysis, we perform a data compression within a single index, with the minimum loss of information compared to previous approaches. We employ autoencoders based on neural networks that are able to grasp the relevant information in the dataset, even in presence of large datasets and without a backing theory. We apply this estimator to the evaluation of wellbeing in the suburban areas of Florence, and compare results from our methods with ones coming from previous approaches.

2 Methodology

Let $\mathbf{X}$ be a $N \times K$ normalized matrix of covariates, describing socioeconomic phenomena , observed for N units and K covariates. An autoencoder (Hinton & Zemel, 1993) is a type of neural network that consists of an encoder and a decoder, where the encoder maps the input data to a lower-dimensional latent space and the decoder maps the latent representation back to the original data space. The encoder can be seen as a probabilistic mapping function that generates a probability distribution over the latent variables, given the input data, Kramer, 1991 call it as nonlinear PCA, which is a quite familiar method into syntetic indexes literature. Therefore, let be ϕ an encoder function that maps $\mathbf{X}_{N\times K} \rightarrow \mathbf{R}_{N\times 1}$, and similarly let ψ be a decoder function mapping $\mathbf{R}_{N\times 1} \rightarrow \mathbf{X}_{N\times K}$. Thus the autoencoder is trained to minimize

$$\underset{\phi,\psi}{\operatorname{argmin}} ||\mathbf{X} - \psi(\phi(\mathbf{X}))||^2$$

In our application, as we wish to summarize covariates into a single vector, we are interested in calculating the code $\mathbf{R} = \phi(\mathbf{X})$. See figure 1 for a graphical representation. To assess AutoSynth performances we study stress values. Let $\theta = \sqrt{\frac{\sum_{i<j}(d_{ij}-\delta_{ij})^2}{\sum_{i<j} d_{ij}^2}}$ be a stress measure of the discrepancy between the distances in the original high-dimensional space ($d_{i,j}$) and the distances in the lower-dimensional space ($\delta_{i,j}$). Thus, The lower the value of θ, the higher the ability of the low-dimensional variables in representing the original data.

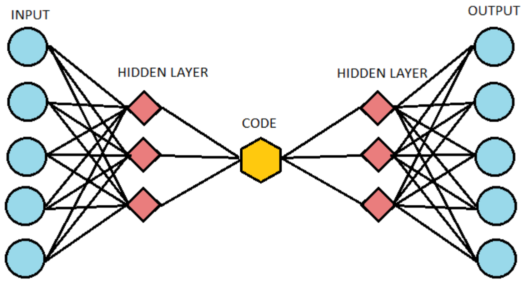

Figure 1: Basic scheme of autoencoders. In this application, inputs will be elementary indicators of socio-economic development, while the code will be the synthetic indicator

Table 1: Fragility dimensions of Florence - year 2021

Demographic	**Economic**	**Social**
% of elders in the population	% of inhabitants in poverty	% of minors in single-parent families
Natural balance	% of families in poverty	% of elders living alone
5-yr variation of inhabitants	% of rented residents	% of foreigners minors
	Median family income	% of graduates
		Permanent residents

3 Measuring Florentine fragilities

We applied our proposed method to the evaluation of fragilities into Florentine suburbs. Fragility can be represented into a composite indicator of three main dimensions: demographic fragility, economic fragility and social fragility. Moreover, we can identify some elementary indicators, previously used in this literature, to represent each of these dimensions. Table 1 shows the indicators used in the analysis, referred to 2021. In total, we collect information over the 74 suburbs that make up Florence. Using the elementary indicators in table 1, we first normalize the variables, as in Mazziotta & Pareto, 2018, and later we apply on the same dataset, AMPI, PCA and AutoSynth transformations, rescaling the compressed variables to the same ”goalposts”, as in Mazziotta & Pareto, 2018. Figure 2 and table 2 reports the fragility maps and the stress value for the three methods considered. From these results, we notice that our model has very noticeable performances in representing the input covariates, and thus is able to reproduce better the original dimensions into a single feature space.

4 Conclusion

Concluding, In this work, we propose to use Autoencoders to construct a synthetic indicator for socio-economic development and apply it to the evaluation

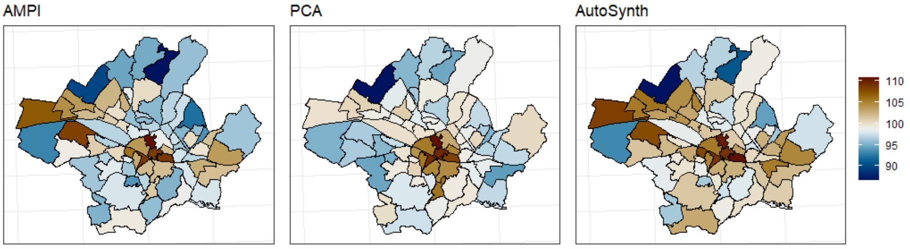

Figure 2: AMPI, PCA and AutoSynth Fragility Index for Florence

Table 2: Stress absolute values for each method considered and as fraction of the AMPI stress test

	AMPI	PCA	AutoSynth
θ	0.03497	0.00657	0.00447
	1	0.188	0.128

of fragility in the Florence suburbs. Results obtained from the stress values suggest an improved ability in dimension reduction, nevertheless, the maps comparison shows similar results with respect to the AMPI. Considering the wide flexibility of autoencoders, their application to the construction of synthetic indicators could become a promising area of study.

References

COMMISSION, JOINT RESEARCH CENTRE-EUROPEAN, *et al.* 2008. *Handbook on constructing composite indicators: methodology and user guide*. OECD publishing.

GRECO, SALVATORE, ISHIZAKA, ALESSIO, TASIOU, MENELAOS, & TORRISI, GIANPIERO. 2019. On the methodological framework of composite indices: A review of the issues of weighting, aggregation, and robustness. *Social indicators research*, **141**, 61–94.

HINTON, GEOFFREY E, & ZEMEL, RICHARD. 1993. Autoencoders, minimum description length and Helmholtz free energy. *Advances in neural information processing systems*, **6**.

KRAMER, MARK A. 1991. Nonlinear principal component analysis using autoassociative neural networks. *AIChE journal*, **37**(2), 233–243.

MAZZIOTTA, MATTEO, & PARETO, ADRIANO. 2018. Measuring well-being over time: The adjusted Mazziotta–Pareto index versus other non-compensatory indices. *Social Indicators Research*, **136**, 967–976.