Assessing Partial Triadic Analysis with MaximumEntropy Bootstrap: an application to BES Italianeducation indicators

Caputi Marco (caputi@istat.it)
ISTAT
Livio Fenga
ISTAT https://orcid.org/0000-0002-8185-2680

Research Article

Keywords: BES, Education, Equitable and sustainable well-being, Maximum Entropy Bootstrap, Partial Triadic Analysis, STATIS approach

Posted Date: March 22nd, 2021

DOI: https://doi.org/10.21203/rs.3.rs-347489/v1

Noname manuscript No.
(will be inserted by the editor)

Assessing Partial Triadic Analysis with Maximum Entropy Bootstrap: an application to BES Italian education indicators

Marco Caputi · Livio Fenga

Received: date / Accepted: date

Abstract In this paper we analyse a set of Italian equitable and sustainable well-being indicators, related to the education and training domain, considered at a regional level over different years. To this end, the Partial Triadic Analysis – a multiway approach which allows to properly deal with time-dependent data structures – will be used in conjunction with a bootstrap method to carry out the assessment of the outcomes obtained. The adopted resampling scheme – called Maximum Entropy Bootstrap – proved to be an appropriate analytical tool, given its ability to preserve the longitudinal data structure without incurring in the usual restrictive assumptions of more traditional bootstrap methods. To the best of the authors' knowledge, this is the first time a bootstrap method for longitudinal data has been employed in PTA.

Keywords BES · Education · Equitable and sustainable well-being · Maximum Entropy Bootstrap · Partial Triadic Analysis · STATIS approach

1 Introduction

The Partial Triadic Analysis (PTA) has been proposed at the end of the 1970s Jaffrenou (1978) as a multi-table analysis technique, aimed at efficiently extracting valuable information from complex structures. Derived from the triadic analysis, the PTA – proposed at first in the field of ecology by Thioulouse et al. (1987) and subsequently renamed by Kroonenberg (1989) – follows the same analytical approach as the STATIS method Lavit (1994). In essence, it is a powerful statistical tool fruitfully employed in both theoretical and applied

Marco Caputi
Italian National Institute of Statistics, via Cesare Balbo 16, 00184, Rome Italy E-mail: caputi@istat.it

Livio Fenga
Italian National Institute of Statistics, via Cesare Balbo 16, 00184, Rome Italy E-mail: fenga@istat.it

research, characterised by the presence of three–way information-sets which can be represented as a same-structure sequence of data matrices (arrays). In particular, in the case of repeated measurements on the same units, PTA allows the capture of both the structural relationships as well as the dynamics of the observed units with respect to a set of variables.

The PTA, as well as other STATIS-related methods, has been successfully employed in many fields of research, including social and economic sciences (see, among others, Bolasco (1986), Bolasco (1992), Bolasco et al. (1992)).

A careful evaluation of the outcomes generated by this technique has been conducted using a suitable resampling method (illustrated in Section 5). To the best of the authors' knowledge, this is the first time a bootstrap method able to take into account the longitudinal structure of the data has been employed for this purpose. In more details, we will show how bootstrap-based confidence intervals can greatly enhance the amount of valuable information that can be extracted from a longitudinal data set, even if the length of the time series is of moderate size.

2 Description of the data

On yearly basis, the Italian National of Statistical issues a report focusing on equitable and sustainable well-being (BES) (Istat (2020)). The driving force behind that is to offer an integrated picture of the main economic, social and environmental interactions, with particular emphasis on the territorial aspects. In more details, those published are a set of 130 basic indicators related to 12 domains, i.e.

1. Health;
2. Education and training;
3. Work and life balance;
4. Economic well-being;
5. Social relationships;
6. Politics and Institutions;
7. Security;
8. Subjective well-being;
9. Landscape and cultural heritage;
10. Environment;
11. Innovation, research and creativity;
12. Quality of services.

The structure of the BES indicators tend to evolve according to the growing attention paid, at a European level, to innovative measurement systems and state-of-the-art projects devoted to deepening the relationships between economic policies and i) the objectives of well-being, equity and sustainability

as well as ii) the analysis of the determinants leading to a sustainable and inclusive economic growth.

In accordance with the Italian applicable law, the BES indicators must always be taken into consideration whenever a new economic policy, impacting the quality of citizens' life in various ways, is implemented. This fact is consistent with the goal pursued by these indicators, mainly related to the evaluation of the progress achieved by a society as a whole. In fact, they have been specifically designed to capture vital information about the quality of life of a generic citizen in his lifespan. Out of the group of the BES basic indicators, in this paper, we restrict our attention to those belonging to the domain pertaining education and training, considered at NUTS 2 (regional) level, for the period 2008 – 2017. In particular, the following indicators will be considered:

- **Participation in the school system of children aged 4-5 (*kinder*)**: percentage of children, aged 4-5 years, participating in pre-primary education or in primary education on total children aged 4-5 years.
- **People with at least an upper secondary education level aged 25-64 years (*hi_sc*)**: percentage of people, aged 25-64 years, having completed at least the upper secondary education (according to the International Standard Classification of Education level not below 3) over the total number of people belonging to the same age group.
- **People having completed tertiary education aged 30-34 years (*degree*)**: percentage of people, aged 30-34 years, having completed tertiary education (according to the International Standard Classification of Education level equal to 5, 6, 7 or 8) over the total number of people belonging to the same age group.
- **People not in education, employment, or training (*neet*)**: percentage of people, aged 15-29 years, which are not in education, employment, or training over the total number of people of the same age group.
- **Participation in life-long learning (*cont_tr*)**: percentage of people, aged 25-64 years, participating in formal or non-formal education over the total number of people belonging to the same age group.

For the sake of brevity, in what follows, each of the above listed variables will be referred to using the abbreviation reported in parenthesis.

3 Basic notation and tools

The dataset considered in this paper can be seen as a three-way array of continuous real variables, say $\{X_{ijk}; i = 1, \ldots, I, j = 1, \ldots, J, k = 1, \ldots, K\}$ representing, respectively, the units, the variables and the different occasions to which the data relates. For brevity, with the symbol $\{X_k; k = 1, \ldots, K\}$, the data matrix observed at time k is denoted. Furthermore, to each occasion, it is associated a triplet (X_k, S_k, D_k), with D_k and S_k being positive definite

matrices of dimension $I \times I$ and $J \times J$, respectively. In particular, these two matrices are associated to the following inner products:

$$\langle u, v \rangle = u^{'} D_k v \qquad \forall u, v \in \mathbb{R}^{\mathrm{I}}, \tag{1}$$

$$\langle w, z \rangle = w^{'} S_k z \qquad \forall w, z \in \mathbb{R}^{\mathrm{J}}, \tag{2}$$

where D_k is a diagonal matrix of positive values containing the weights of the units (tipically $\frac{1}{I} I$) whereas S_k (not necessarily diagonal) represents the metric associated to the variables.

Without loss of generality, in what follows, the columns of X_k will be centered with respect to the matrix D for an easier mathematical representation, that is: $1_I D X_k = 0_J$, where 1_I and 0_J are, respectively, the I-vector of ones and the J-vector of zeroes. Furthermore, from now on, we will consider $S_k = I_J$ since, for each occasion, the size effect has been removed by dividing each column by its standard deviation.

Since we deal with a three-way array, made up of data matrices related to different occasions, the comparison among these matrices can be of particular interest. To this end, we consider the Frobenius matrix product defined on the vector space $\mathcal{M}_{I \times J}(\mathbb{R})$ of real matrices, i.e.

$$\langle A, B \rangle_F = Tr\left(A^{'} B\right) \qquad \forall A, B \in \mathcal{M}_{I \times J}(\mathbb{R}), \tag{3}$$

where Tr is the usual trace operator. Since the Frobenius matrix product is known to possess the scalar product properties, it is always possible to extend, to each of the matrices belonging to $\mathcal{M}_{I \times J}$, the variance, covariance and correlation coefficient statistics by denoting them, respectively, as follows:

$$VarV\,(A) = Tr\left(A^{'} A\right), \tag{4}$$

$$CovV\,(A, B) = Tr\left(A^{'} B\right), \tag{5}$$

$$RV\,(A, B) = \frac{Tr\left(A^{'} B\right)}{\sqrt{Tr\,(A^{'} A)\, Tr\,(B^{'} B)}}. \tag{6}$$

In particular, it holds that $RV \in [0, 1]$, taking the value 1 if and only if $A = \gamma B$, for some scalar γ, with $\gamma \neq 0$. It is also noted that the RV coefficient can be seen as a scalar product between matrices scaled with respect to their Frobenius norms (represented by the denominator of RHS in Eqn. 6).

Finally, we observe that the Frobenius matrix product can also be expressed as

$$\langle A, B \rangle_F = \left\langle vec(A), vec(B)^{'} \right\rangle, \tag{7}$$

with $vec()$ being the so called matrix stack operator, which – by stacking all the columns vectors of the matrix each one below one another – creates a single column vector.

4 Partial Triadic Analysis

The PTA (Kroonenberg (1989), Thioulouse et al. (1987)) is a statistical method specifically designed to perform the joint analysis of multiple sets of data having the same structure with respect to all the occasions. This approach is similar to STATIS method (Lavit (1994)) however, in this case, the method is applied directly to the elementary matrices $X_k \in \mathcal{M}_{I\times J}(\mathbb{R})$.

The PTA relies on the properties of the Euclidean spaces and is carried out in three different steps, as below detailed:

1. **Interstructure**
   The Interstructure step is aimed at analyzing the data globally and, to this end, captures the similarity among the data structures related to the different occasions.
   Formally, the Interstructure matrix is based on the Frobenius matrix product, computed as:

$$\Gamma_{k,k'} = Tr\left(X_k X_{k'}\right) \qquad \forall k, k' \in K \tag{8}$$

   or, alternatively, as:

$$C_{k,k'} = \frac{Tr\left(X_k X_{k'}\right)}{\sqrt{Tr\left(X_k X_k\right) Tr\left(X_{k'} X_{k'}\right)}} \qquad \forall k, k' \in K. \tag{9}$$

   It is worth emphasizing that the generic element (k, k') represents either the vectorial covariance (Eqn. 8) or the Escoufier RV's correlation coefficient (Eqn. 9) between the two occasions k and $k^{'}$. Since, by using the latter, the assessment of the degree of similarity among the available occasions – controlling for the size effect – is allowed, we will choose them for the analysis.

   Once the matrix C is obtained, the analysis of the similarities among the occasions is based on the data projection onto the factorial space induced by its first eigenvectors.

2. **Compromise**
   The Compromise step is designed to define a common representation of the data through a factorial space with respect to the whole period considered. This representation is based on the matrix obtained by the following weighted sum:

$$X^* = \sum_{k=1}^{K} \alpha_k X_k, \tag{10}$$

   where the vector $\alpha = (\alpha_1, \ldots, \alpha_K)$ is the first eigenvector of the Inter-structure matrix C. It can be shown (Lavit (1994)) that the matrix X^* represents the best approximation of the matrices $\{X_k; k = 1, \ldots, K\}$ with respect to the norm induced by the Frobenius matrix product.

3. **Intrastructure**
   In this step, the specific information contained in the occasion matrices $\{X_k; k = 1, \ldots, K\}$ is analysed by decomposing the squared distances, computed using the matrix C, through the following equation:

$$D^2_{k,k'} = d^2(X_k, X_{k'}) = 2(1 - C_{k,k'}), \tag{11}$$

   D^2 being defined as the matrix of the square distances. By using D^2 with Eqn. 9, the contributions related to each unit are computed allowing a more insightful understanding of the changes occurring in the indicators and the detection of the more influencing units in the data set (Lavit (1994)).

5 The Resampling Method

Proposed in Vinod (2006) and subsequently improved in Vinod et al. (2009) and Vinod (2016), the Maximum Entropy Bootstrap (MEB) is the resampling scheme adopted here. At its core, MEB is a fully non-parametric method relying on a resampling mechanism which is different from other more traditinal solutions. In fact, what is generally done in standard procedures. In fact, while in the classic bootstrap an ensemble $\boldsymbol{\Omega}$ represents the population of reference the observed time series is drawn from, in *MEB* a large number of ensembles, say $\{\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_N\}$, becomes the elements belonging to $\boldsymbol{\Omega}$, each of them containing a number B of replicates $\{x_1, \ldots, x_B\}$. In addition, MEB is characterised by the following three important features:

1. it is distribution-free
2. it is able to handle any type of persistence
3. it can generate pseudo-replications satisfying both the ergodic and the central limit theorems

Let $\mathcal{H}_s$ be the Shannon entropy, defined as

$$\mathcal{H}_s = \mathbb{E}\left[-\log(\mathcal{G}(x))\right], \tag{12}$$

being $\mathcal{G}(\cdot)$ the probability density function, assigning a level of probabilities to each of the statuses a given system can be found in, and $\mathbb{E}(\cdot)$ the expectation function. The application of Eqn. 12 implies $\mathcal{G}(x)$ to be the argument of a standard optimization problem of the type

$$\mathcal{F}(x) = \underset{\mathcal{G}'}{\operatorname{argmax}} \mathbb{E}\left\{-\log \mathcal{G}'(x)\right\}, \tag{13}$$

which is solved within the maximum entropy framework by introducing a set of additional constraints, imposed on both the probability mass function and the probabilistic structure of each of the bootstrap replications, as explained below.

Let $\{X_t\}_1^T$ be the observed time series $x_1, x_2, \ldots, x_T$ and denote by $x_{(1)}, x_{(2)}, \ldots, x_{(T)}$ its order statistics: it holds that

$$x(t) \in \left[x_{(0)}, x_{(T+1)}\right]. \tag{14}$$

The set of midpoints, denoted by $z_1, z_2, \ldots, z_T$, is computed as

$$z_t = \frac{1}{2}\left(x_{(t)} + x_{(t+1)}\right) \qquad t = 1, \ldots, T-1, \tag{15}$$

being $z_0 = x_{(0)}$ and $z_T = x_{(T+1)}$. Using the midpoints, the T half open intervals $\mathbb{B}_t$, computed around each observation, are defined i.e. $\mathbb{B}_t = (z_{t-1} - z_t]$.

As for the above mentioned constraints, two of them are imposed on the density function $\mathcal{G}(x)$ to guarantee the whole information set to be correctly accounted for. Formally, they can be expressed as

$$\sum_{t=1}^{T} x_t = \sum_{t=x(t)}^{T} x_{(t)} = \sum_{t=1}^{T} m_t, \tag{16}$$

where $m_t = \mathbb{E}\mathcal{G}(\mathbb{B}_t)$. In particular, the last constraint is designed to preserve the probabilistic structure of the system and it is fulfilled by imposing the perfect rank correlation between the original series and its bootstrap replications.

6 MEB algorithm

The MEB algorithm generates a single bootstrap sample according to an algorithm which will be now broken down in a step-by-step fashion:

1. the order statistics for the observed data x_t are generated according to (14), being
$$x_{(0)} = x_{(1)} - \tilde{\delta}$$
and
$$x_{(T+1)} = x_{(T)} + \tilde{\delta},$$
with $\tilde{\delta} = \tilde{\mathbb{E}}\left[\left|x_{(t)} - x_{(t-1)}\right|\right]$.
2. A sorting matrix of dimension $T \times 2$, say W_1, accommodates in the first column the time series x_t and an index set – denoted by $\mathcal{I}_{ind} = \{1, 2, \ldots, T\}$ – in the other one;
3. W_1 is sorted according to the first column and the order statistics $\boldsymbol{x_{(t)}}$ is thus defined. We then compute the midpoints z_t and the half-open intervals $\mathbb{B}_t$;
4. T pseudo-random numbers are drawn from a uniform distribution, $p_s \sim U[0,1]; s \in \{1, \ldots, T\}$ and a range of values $\mathcal{R}_t = \left(\frac{t}{T}, \frac{t+1}{T}\right] \forall t \in \{0, 1, \ldots, T-1\}$ wherein each p_s falls, is defined;
5. both $\mathcal{R}_t$ and $\mathcal{I}_t$ are matched and a new set $\tilde{x}_t^*$ is drawn using the density function $\mathcal{G}(x)$;
6. a new sorting matrix W_2, exhibiting the same structure of W_1, is defined. It serves the purpose of sorting the T-dimension vector $\tilde{x}_t^*$ in increasing order and thus it defines the ordering statistic $\tilde{x}_{(t)}^*$;
7. the order statistics $x_{(t)}$ in W_1 is replaced by $\tilde{x}_{(t)}^*$ in W_2, generated in Step 6. Then, $\tilde{x}_{(t)}^*$ is sorted according to the first column of W_1 and x_t^* is recovered. The vector x_t^* stores a single bootstrap replication of the original series x_t.

7 Data bootstrapping

In this section, the resampling scheme introduced in Section 5 will be used to generate the bootstrapped data. Before that, the main issue related to the choice made of the number B of the bootstrapped series, is addressed. In practice, the final decision has been made using a trial and error approach, where three different bootstrap sample sizes B – i.e. $\{B = 75, 125, 175\}$ – have been tested. The number of pseudoseries has finally been set to $B = 75$, according to the best variance-bias trade–off criterion.

The bootstrap exercise has been carried out considering separately each and every time series, without attempting any explicit inclusion of the correlation structures present in the original data. The cross-variables connection accounted for is related to the bootstrap time rank, which has been kept constant for each run. Such a choice is justified by the limited sample size available, hardly suitable for multivariate approaches, and corroborated by the literature where several multivariate applications of the MEB algorithm, based on just one variable at time, can be found. For example, Srivastav et al. (2017) proposed a multivariate bootstrap–based scheme aimed at capturing the multicollinear dependency structure embedded in the investigated data set, related

to weather variables. Their approach envisions the employment of the MEB algorithm in conjunction with an orthogonal linear transformation. Lin Shang (2017) faces the problem of the online prediction of high frequency ecological data – as well as of the related forecasting intervals – under a high dimensional data context. In essence, the proposed forecasting method decomposes a functional time series into a number of functional principal components and their associated scores. In such a set up, which clearly poses challenges from a statistical point of view (summarised by the concept known as curse of dimensionality), the author successfully employed the MEB scheme to construct bona fide replications of the original time series. The performances recorded on real-life data can be defined remarkable. Finally, Vinod et al. (2009), discuss a Keynesian consumption function, using the Friedman's permanent income hypothesis, applying MEB scheme separately to each of the considered time series.

8 Data analysis

In this section we apply the PTA to the three-way array of the BES indicators, described in Section 2.

We first carry out the Interstructure step, by computing the matrix C of RV coefficients between pairs of occasions (see Eqn. 9). By inspecting Table 4, where the related results are reported, the high degrees of correlation between the occasions, always greater than 0.79, are noticeable.

Table 1 C matrix of RV coefficients

	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017
2008	1.000	0.956	0.931	0.915	0.861	0.841	0.822	0.807	0.789	0.799
2009	0.956	1.000	0.945	0.917	0.861	0.848	0.839	0.845	0.815	0.818
2010	0.931	0.945	1.000	0.947	0.887	0.871	0.883	0.871	0.863	0.849
2011	0.915	0.917	0.947	1.000	0.927	0.912	0.887	0.873	0.857	0.806
2012	0.861	0.861	0.887	0.927	1.000	0.971	0.942	0.920	0.902	0.854
2013	0.841	0.848	0.871	0.912	0.971	1.000	0.948	0.913	0.893	0.847
2014	0.822	0.839	0.883	0.887	0.942	0.948	1.000	0.963	0.958	0.918
2015	0.807	0.845	0.871	0.873	0.920	0.913	0.963	1.000	0.963	0.911
2016	0.789	0.815	0.863	0.857	0.902	0.893	0.958	0.963	1.000	0.936
2017	0.799	0.818	0.849	0.806	0.854	0.847	0.918	0.911	0.936	1.000

The factorial representation calculated from the matrix C is reported in Figure 1, where the first two axes account for 94.6% of the total inertia (the first axis explains 89.8%). It is worth mentioning that the data points relative to the different years are not too scatterd and, in fact, exhibit an approximatively regular pattern.

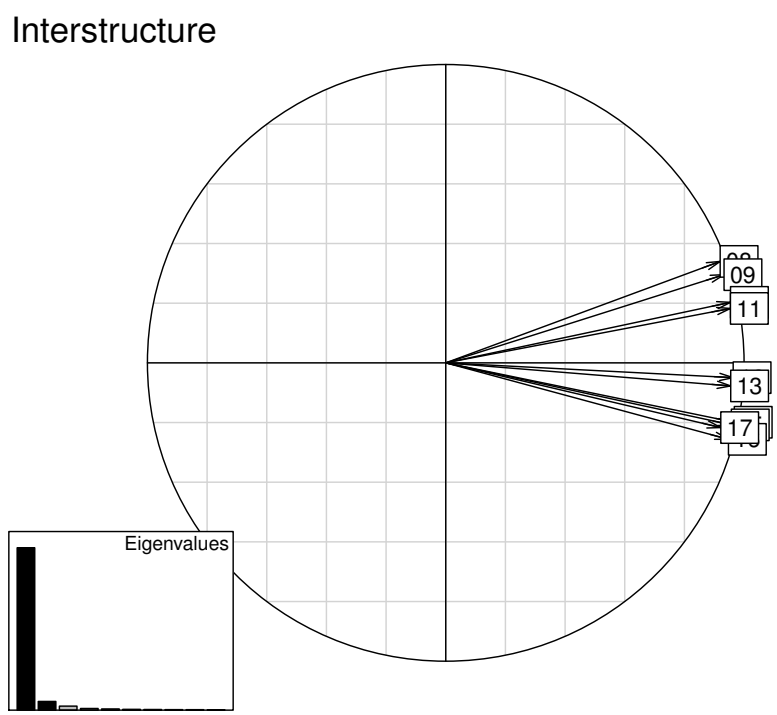

Fig. 1 Graphical representation of the C matrix

In order to assess the goodness of the eigendecomposition of the matrix C, B=75 MEB replicates of the whole data array have been generated. For each bootstrap replication, the matrix storing the RV coefficients, say C_b, as well as the related eigendecomposition, are computed. As a result of that, the B bootstrap distributions related to both the eigenvalues and the eigenvectors are obtained.

The outcomes relative to the bootstrap distributions of the first two eigenvalues are represented by the 0.05-0.95% quantile-based boxplots, as reported in Figure 2. From its inspection, it can be noticed that both the first and the second eigenvalue of the matrix C lie within their respective 5-95% bootstrap interval. The remaining eigenvalues are left out of the analysis since they account for a negligible amount of inertia.

By applying the same procedure to the first two eigenvectors, the 0.05-0.95% quantile-based boxplots of the first eigenvector components (one for each year) – reported in Figure 3 – are generated. Also in this case, all the first eigenvector components of the matrix C lie within their respective 5-95%

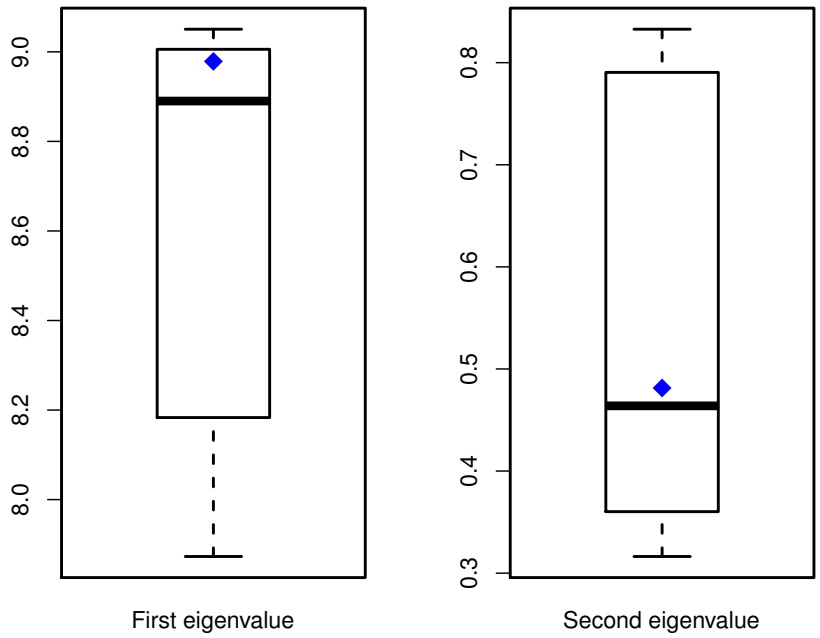

Fig. 2 Quantile-based boxplot (0.05-0.95) of the first two eigenvalues bootstrap distribution

bootstrap interval. The second eigenvector, even though not reported, has been observed to behave consistently with the first one.

Both these results seem to confirm that the representation obtained for the Interstructure step can be considered reliable. Recalling Eqn.10, we can therefore use these weights for the construction of the Compromise matrix H^*.

In order to assess the quality of the projection of each matrix $\{X_k; k = 1, \ldots, K\}$ on the Compromise space, we use Figure 4, where the typological values of the matrices X_k are depicted. Here, the weights α_k (Eqn.10) and the square of the cosinus of the projection on the Compromise space are reported, respectively, on the x and y axes. Both the high quality representation of these matrices (the squared cosinus is always greater than 0.91) and the fact that the estimated weights are approximately similar to each other, are necessary conditions to define a reliable Compromise space.

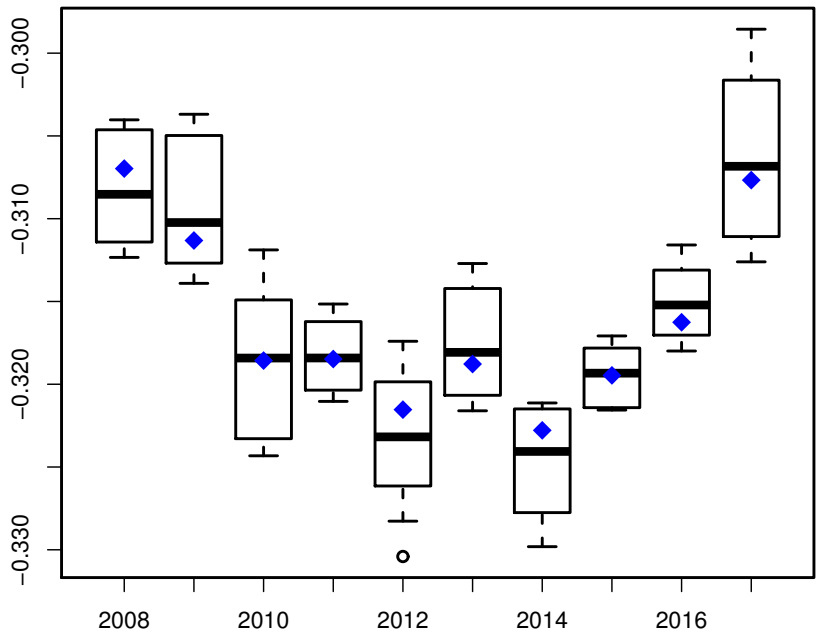

Fig. 3 Quantile-based boxplot (0.05-0.95) of the first eigenvector bootstrap distribution

Figures 5 and 6 show, respectively, the scree-plot and the biplot of the Compromise matrix X^*, which represents the "common structure" of the matrices $\{X_k; k = 1, \ldots, K\}$ (Bolasco (1999)). It is noted that the first two axes account for a percentage of the total inertia equal to 88.1%. Tables 2 and 3 report the contributions of the units to the principal components and their quality of representation, respectively.

By inspecting Figure 6, furthermore, we observe that the first axis is positively correlated with the variable *neet*, which represents the percentage of people not in education, employment or training, and negatively correlated to the variables representing secondary and tertiary education (*hi_sc* and *degree*) and, even though to a lesser extent, to the percentage of people aged 25-64 participating in formal or non-formal education (*cont_tr*). Based on these associations, it can be said that this axis account for the propensity in achieving higher level of education, even if not formally recognised.

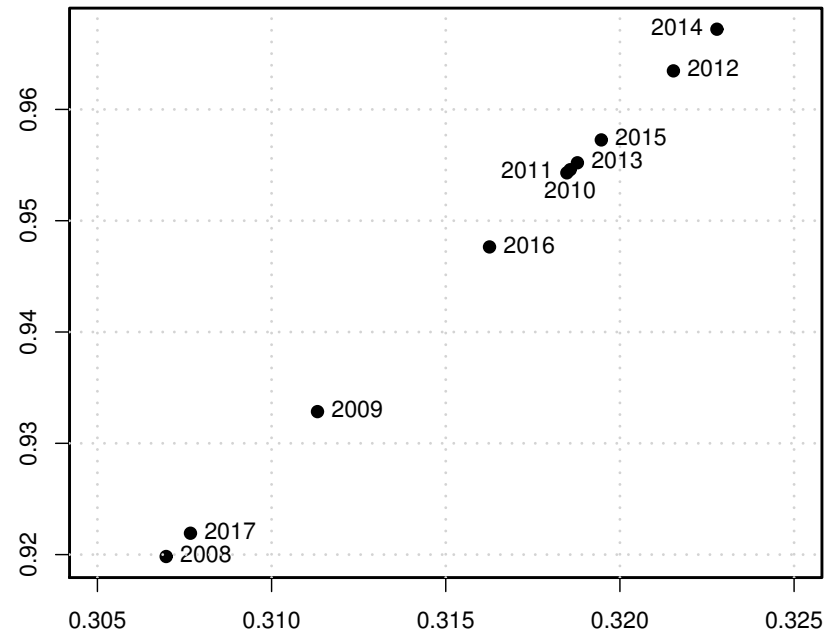

Fig. 4 Typological values of the observed data matrices

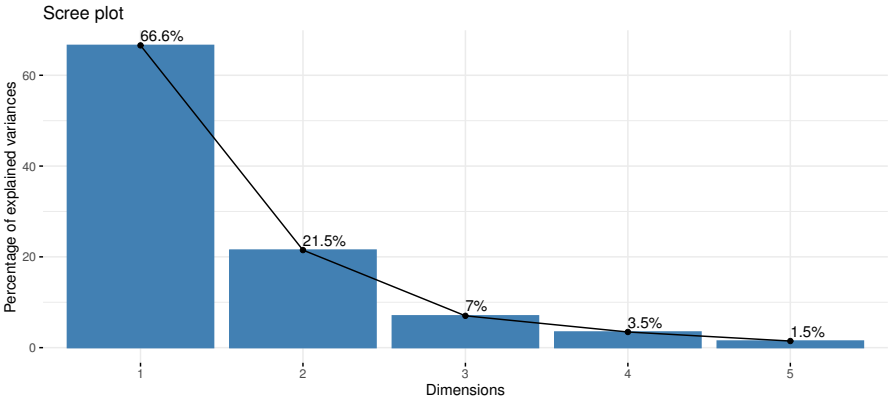

Fig. 5 Scree-plot of the Compromise matrix

Table 2 Contributions of the regions to principal components

	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5
PIEDMONT	1.047	0.980	3.155	14.945	0.145
AOSTA VALLEY	2.125	4.665	1.175	0.064	10.350
LIGURIA	6.322	0.072	0.009	21.466	2.840
LOMBARDY	14.394	0.002	0.205	1.957	1.503
TRENTINO - SOUTH TYROL	3.377	6.981	0.363	6.417	0.028
VENETO	3.787	1.664	4.400	0.083	1.241
FRIULI - VENEZIA GIULIA	4.898	23.242	1.703	3.773	1.696
EMILIA - ROMAGNA	3.585	0.960	2.323	9.143	0.121
TUSCANY	1.425	4.081	1.699	7.698	0.726
UMBRIA	1.041	1.160	10.774	1.207	0.414
MARCHE	0.011	16.347	0.073	1.050	34.483
LAZIO	0.066	2.006	5.685	0.203	2.292
ABRUZZO	12.808	0.299	0.307	1.666	2.924
MOLISE	5.043	9.374	29.351	1.414	8.669
CAMPANIA	22.258	2.533	0.012	0.090	13.752
APULIA	1.289	0.236	0.872	4.551	4.880
BASILICATA	9.618	12.105	15.359	0.178	3.120
CALABRIA	6.422	0.039	1.080	6.358	0.173
SICILY	0.119	13.215	18.808	5.256	0.839
SARDINIA	0.364	0.038	2.647	12.481	9.804

Table 3 Squared cosinus for the regions

	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5
PIEDMONT	0.423	0.128	0.134	0.313	0.001
AOSTA VALLEY	0.533	0.378	0.031	0.001	0.057
LIGURIA	0.841	0.003	0.000	0.148	0.008
LOMBARDY	0.989	0.000	0.001	0.007	0.002
TRENTINO - SOUTH TYROL	0.563	0.375	0.006	0.055	0.000
VENETO	0.786	0.111	0.096	0.001	0.006
FRIULI - VENEZIA GIULIA	0.382	0.586	0.014	0.015	0.003
EMILIA - ROMAGNA	0.777	0.067	0.053	0.103	0.001
TUSCANY	0.427	0.395	0.054	0.120	0.005
UMBRIA	0.397	0.143	0.433	0.024	0.003
MARCHE	0.002	0.864	0.001	0.009	0.124
LAZIO	0.048	0.472	0.436	0.008	0.037
ABRUZZO	0.979	0.007	0.002	0.007	0.005
MOLISE	0.441	0.265	0.271	0.006	0.017
CAMPANIA	0.952	0.035	0.000	0.000	0.013
APULIA	0.716	0.042	0.051	0.131	0.059
BASILICATA	0.632	0.257	0.106	0.001	0.004
CALABRIA	0.933	0.002	0.017	0.048	0.001
SICILY	0.018	0.641	0.298	0.041	0.003
SARDINIA	0.240	0.008	0.184	0.427	0.142

On the other hand, the second axis of the biplot is mainly associated to only one variable, i.e. the one accounting for the percentage of children aged 4-5 participating in pre-primary or primary education (*kinder*).

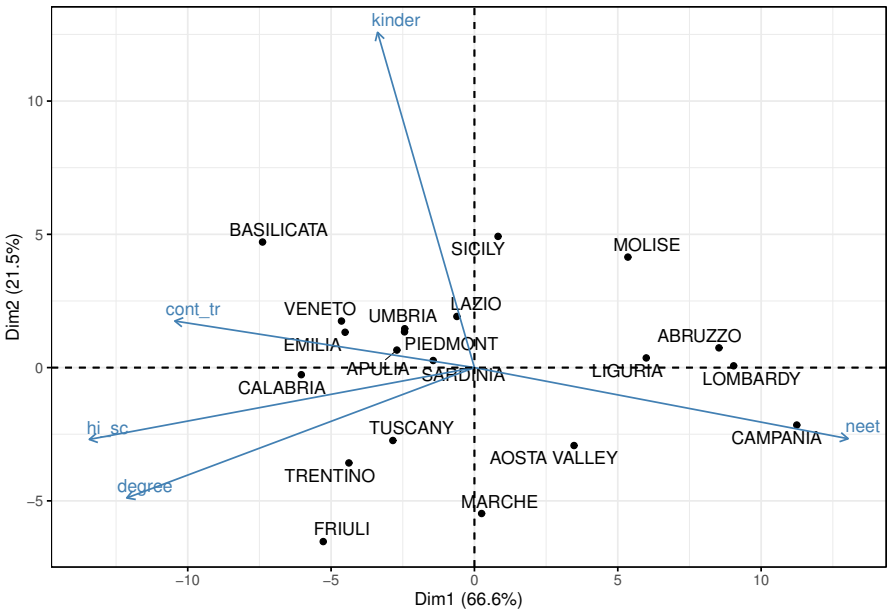

Fig. 6 Biplot of the Compromise matrix

Still, by inspecting Figure 6, it can be noticed a similar pattern among some regions. In particular, Veneto, Emilia-Romagna, Apulia and Calabria exhibit similar scores with respect to the first axis, which is a clear indication of their propensity in pursuing higher levels of education. On the contrary, Liguria, Abruzzo, Lombardy and Campania tend to show similar values located on the opposite side of the same axis.

Considering the second axis, it is noted that Marche, Aosta Valley, Trentino-South Tyrol and Tuscany exhibit low degrees of propensity towards sending their children to the pre-primary or primary schools. On the other hand, an opposite behaviour is showed by some regions, e.g. Sicily, Friuli-Venezia Giulia, Marche and Lazio.

As for the Intrastructure step, the matrix D^2 representing the squared distances among the occasions (years) and thus the differences expressed at the whole regional system level, is computed (see Eqn. 11) and reported in table 4. In order to define a clear and informative set of comparisons between the years, only the squared distances between the first year of the time span (with 2008 the reference year) and, one at a time, all the following years, are considered. The percentage contributions of each region to the squared distance – computed through Eqn. 11 and Eqn. 6 – are reported in Table 5.

Table 4 D^2 matrix of squared distances between years

	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017
2008	0.000	0.088	0.138	0.169	0.279	0.318	0.357	0.385	0.421	0.402
2009	0.088	0.000	0.111	0.166	0.278	0.303	0.322	0.309	0.371	0.365
2010	0.138	0.111	0.000	0.106	0.226	0.258	0.234	0.258	0.273	0.302
2011	0.169	0.166	0.106	0.000	0.146	0.177	0.226	0.254	0.286	0.389
2012	0.279	0.278	0.226	0.146	0.000	0.057	0.116	0.161	0.196	0.293
2013	0.318	0.303	0.258	0.177	0.057	0.000	0.104	0.173	0.213	0.305
2014	0.357	0.322	0.234	0.226	0.116	0.104	0.000	0.075	0.084	0.163
2015	0.385	0.309	0.258	0.254	0.161	0.173	0.075	0.000	0.074	0.178
2016	0.421	0.371	0.273	0.286	0.196	0.213	0.084	0.074	0.000	0.129
2017	0.402	0.365	0.302	0.389	0.293	0.305	0.163	0.178	0.129	0.000

Table 5 Percentage contributions of each region to the squared distance between 2008 and the other years

	2009	2010	2011	2012	2013	2014	2015	2016	2017
PIEDMONT	16.80	3.30	9.10	2.70	4.70	5.50	5.00	7.80	12.70
AOSTA VALLEY	1.90	10.20	5.20	2.80	3.10	3.00	5.40	3.00	3.70
LIGURIA	3.20	1.20	2.20	6.90	3.00	2.90	6.10	5.40	7.10
LOMBARDY	1.90	4.10	0.60	3.20	3.30	1.90	1.20	2.30	1.00
TRENTINO - SOUTH TYROL	6.20	3.30	2.40	1.80	1.30	1.40	1.30	1.90	4.30
VENETO	0.90	4.90	5.50	5.30	3.40	3.50	2.70	1.70	4.60
FRIULI - VENEZIA GIULIA	10.30	7.20	12.10	8.30	6.60	15.10	17.30	20.80	13.40
EMILIA - ROMAGNA	7.80	1.10	1.60	1.80	3.70	4.70	6.50	7.30	8.30
TUSCANY	1.40	1.90	1.30	1.60	3.20	3.80	2.70	3.50	3.00
UMBRIA	2.30	5.40	3.50	11.50	11.70	9.10	4.10	5.40	2.00
MARCHE	3.10	3.90	9.20	7.40	5.50	8.10	1.60	9.20	6.40
LAZIO	3.00	14.20	9.90	5.10	5.00	4.80	4.70	5.60	5.90
ABRUZZO	1.80	0.50	0.30	2.90	0.90	0.80	0.90	0.90	2.40
MOLISE	10.70	6.20	4.80	2.60	3.80	6.80	8.90	3.60	2.80
CAMPANIA	6.40	2.60	8.30	8.50	13.90	7.90	10.40	6.90	1.70
APULIA	8.40	9.50	9.20	8.00	6.10	4.80	5.20	5.00	2.80
BASILICATA	4.10	3.30	5.10	5.90	10.10	6.90	5.40	1.70	2.50
CALABRIA	0.80	3.60	2.20	1.40	2.50	1.80	0.90	0.60	1.30
SICILY	7.80	8.30	2.30	10.70	6.50	5.70	7.40	4.90	9.30
SARDINIA	1.40	5.50	5.10	1.80	1.80	1.50	2.30	2.50	4.90

According to Lavit (1994), the rows of the Table 5 can be interpreted as "trajectories". In fact, they account for the relative contribution given by each region to the actual change recorded at the whole regional system in the considered time span.

Some of the regions seem to give a relevant contribution to the squared distance. In particular, Piedmont, Friuli-Venezia Giulia, Umbria and Campania, which show an average contribution greater than 6% (and greater than 10% in at least two occasions). It is worth outlying how such regions belong to all the macro-regions Italy is usually broken into, i.e. North-West (Piedmont), North-East (Friuli-Venezia Giulia), Centre (Umbria) and South (Campania).

As already mentioned, the reliability of the observed trajectories has been assessed through their bootstrap distributions. In practice, these distributions

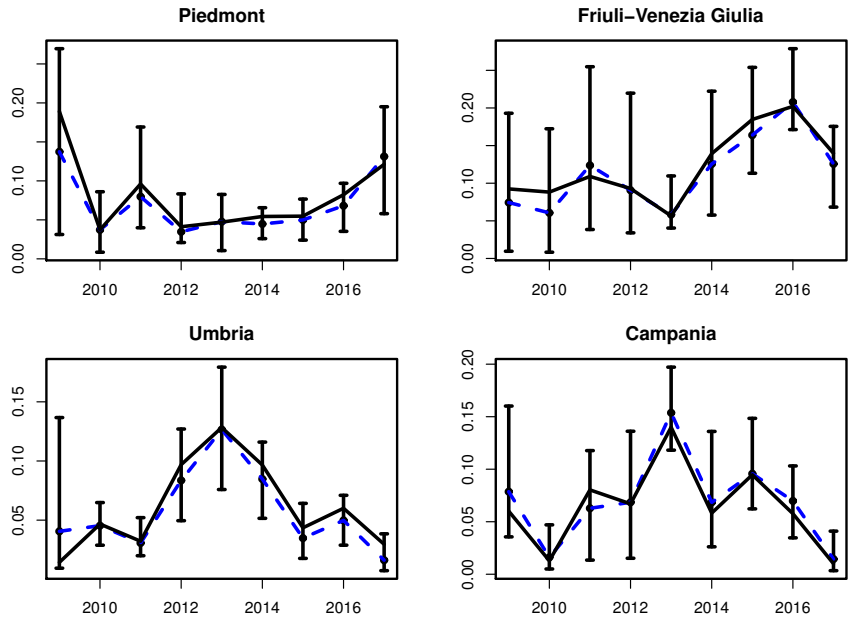

Fig. 7 Trajectories of the most influential regions

have been obtained by computing (see Eqn. 11 and 6) the relative contribution given by each region to the squared distance for each of the bootstrapped series, say $\left\{D_b^2; b = 1, \ldots, B\right\}$. Figure 7 reports the medians and the 5-95% bootstrap intervals related to the four above mentioned regions (the continuous line represents the original sample values, the dashed line the bootstrap median values). Also in this case, the data lie inside the bootstrap confidence limits, fact that confirms the reliability of the observed trajectories .

9 Final remarks

In this paper we proposed a bootstrap-assisted method for the analysis of the BES indicators through the PTA. Its application to a set of basic indicators time series – relative to the domain of education and training – proved its usefulness in capturing both the structural relationships as well as the dynamics generated at a regional level.

In more details, the three-stage PTA technique has made possible the global analysis of the whole set of indicators for each year, so that its dynamic has been investigated over the available timespan. Furthermore, the definition of a suitable compromise space allowed for a single factorial representation of the Italian regions, taking into account the information related to each of the considered years. Finally, the dynamics characterising the most inuential

regions, i.e. those more responsible for the changes recorded at the whole regional system, have been represented.

In addition, an assessment procedure for the PTA has been proposed on the basis of a suitable resampling scheme and applied to each step of the analysis. In particular, given the time-dependent nature of the data, we made use of the MEB algorithm which, as it is well known, is designed to perform satisfactorily under milder assumption than more traditional schemes. The use of the assessment procedure in each step of the the analysis is particularly important given the sequential nature of the PTA approach.

To do so, an adequate number of replicates of the observed data array has been artificially generated through the MEB algorithm, so that the corresponding matrices of RV coefficients have been computed. Using these distributions, we were finally able to carry out a bootstrap-based assesment of the key elements of the analysis (i.e. eigenvalues and eigenvectors of the RV matrix and the squared distances among the occasions).

Our findings point towards the MEB algorithm as a robust resampling tool able to deliver satisfactory performances in our particular context, where the number of occasions (i.e. the length of the time series) is of limited size. Furthermore, to the best of the authors' knowledge, this is the first time a bootstrap method for longitudinal data has been employed in PTA.

Future directions of the present research include the estimate of the minimum sample size of the data arrays for our procedure to be valid.

Disclaimer

The views and opinions expressed in this article are those of the authors and do not necessarily reflect the official policy or position of their institution.

Declarations

Funding

- The authors did not receive support from any organization for the submitted work.
- No funding was received to assist with the preparation of this manuscript.
- No funding was received for conducting this study.
- No funds, grants, or other support was received.

Conflicts of interest/Competing interests

- The authors have no relevant financial or non-financial interests to disclose.
- The authors have no conflicts of interest to declare that are relevant to the content of this article.
- All authors certify that they have no affiliations with or involvement in any organization or entity with any financial interest or non-financial interest in the subject matter or materials discussed in this manuscript.
- The authors have no financial or proprietary interests in any material discussed in this article.

Availability of data

All the data used in this paper are available at https://www.istat.it/en/well-being-and-sustainability/the-measurement-of-well-being/indicators.

Code availability

The software used for computations can be freely be downloaded at https://cran.r-project.org/. The code is available upon request.

References

Alaimo, L.S. and Maggino, F. (2020). Sustainable Development Goals Indicators at Territorial Level: Conceptual and Methodological Issues - The Italian Perspective. *Social Indicators Research*, 147:383-419.

Bolasco, S. (1986). L'evoluzione della struttura dei mercati del lavoro regionali negli anni 1977-1982 in Italia. *L'evoluzione delle strutture economiche regionali*. Franco Angeli, Milano, 123-–144.

Bolasco, S. (1999). *Analisi multidimensionale dei dati*. Carocci.

Bolasco, S. (1992). La complessità della condizione d'inoccupazione/occupazione: problemi di misurazione statistica. *La disoccupazione: interpretazione e punti di vista*. Liguori, Napoli, pp 65-–74.

Bolasco, S. and Coppi, R. (1991). Il ruolo dei metodi di analisi dei dati multiway nello studio della struttura e della dinamica di popolazione. *Annali di statistica*. 120:235-–249.

Caillez, F. and Pagès, J.P. (1976). *Introduction à l'analyse des données*. SMASH, Paris.

Cnel. and Istat. (2015). Report on equitable and sustainable well-being (BES 2014). Rome: Istat.

De L Cruz, O. and Holmes, S. (2011). The duality diagram in data analysis. *The Annals of Applied Statistics* 5:2266-–2277.

Dray, S. and Dufour, A. B. (2007). The ade4 package: Implementing the duality diagram for ecologists. *Journal of Statistical Software* 3:1-–20.

Dray, S., Dufour, A. B. and Chessel, D. (2007). The ade4 package — ii: Two-table and k-table methods. *R News* 7:47-–52.

Escoufier, Y. (2006). Operator related to a data matrix: A survey. *COMPSTAT 2006 — Proceedings in Computational Statistics* 285-–297.

Escoufier, Y. (2006). Operator related to a data matrix: A survey. *COMPSTAT 2006 — Proceedings in Computational Statistics* 285-–297.

Giovannini, E. and Rondinella, T. (2012). Measuring equitable and sustainable well-being in Italy. In F. Maggino and G. Nuvolati (Eds.), *Quality of Life in Italy Research and Reflections* (pp. 9–25). Cham: Springer.

Hall, J., Giovannini, E., Morrone A. and Ranuzzi, G. (2010). *A framework to measure the progress of societies*. France: OECD Publishing.

Istat. (2020). BES report 2019: Equitable and sustainable well-being in Italy. Roma: Istat. https://www.istat.it/en/archivio/237012. Accessed 19 March 2021

Jaffrenou, P. (1978). Sur l'Analyse des Familles Finies de Variables Vectorielles: Bases Algébriques et Applications à la Description Statistique. *Thèse de Troisième Cycle. Université de Lyon*

Kroonenberg, P.M. (1994). The analysis of multiple tables in factorial ecology. III. Three-mode principal component analysis: 'analyse triadique complète'. *Acta Oecologica, Oecologia Generalis* 10:245–256.

Lavit, C. (1994). The ACT (STATIS method). *Computational Statistics and Data Analysis* 18:97–115.

Lin Shang, H. (2017). Functional time series forecasting with dynamic updating: An application to intraday particulate matter concentration. *Econometrics and Statistics*. 1:184–200.

Mazziotta, M. and Pareto, A. (2019). Use and Misuse of PCA for Measuring Well-Being. *Social Indicators Research*, 142:451–476.

Monte, A. and Schoier, G. A Multivariate Statistical Analysis of Equitable and Sustainable Well-Being Over Time. *Social Indicators Research* (2020).

R Development Core Team (2009). R: A Language and Environment for Statistical Computing. *R Foundation for Statistical Computing, Vienna, Austria* http://www.R-project.org. Accessed 19 March 2021

R Package ade4 https://cran.r-project.org/web/packages/ade4/. Accessed 19 March 2021

R Package meboot https://cran.r-project.org/web/packages/meboot/. Accessed 19 March 2021

R Package Factominer http://factominer.free.fr. Accessed 19 March 2021

Srivastav, R.K. and Simonovic, S.P. (2015). Multi-site, multivariate weather generator using maximum entropy bootstrap. *Climate Dynamics* 44:3431–3448.

Stiglitz, J., Sen, A. and Fitoussi, J. P. (2009). Report by the commission on the measurement of economic performance and social progress. Paris: Commission on the Measurement of Economic Performance and Social Progress.

Thioulouse, J. and Chessel, D. (1987). Les analyses multitableaux en ecologie factorielle. i: De la typologie d'etat à la typologie de fonctionnement par l'analyse triadique. *Acta Oecologica, Oecologia Generalis* 8:463—480.

Thioulouse, J. and Dray, S. (2007). Interactive multivariate data analysis in R with the ade4 and ade4TkGUI packages. *Journal of Statistical Software* 22:1—14.

Thioulouse, J. (2011). Simultaneous analysis of a sequence of paired ecological tables: A comparison of several methods. *The Annals of Applied Statistics* 5:2300—2325.

Tomaselli, V., Fordellone, M. and Vichi, M. (2021). Building Well-Being Composite Indicator for Micro-Territorial Areas Through PLS-SEM and K-Means Approach. *Social Indicators Research* 153:407–429.

Vinod, H.D. (2006). Maximum entropy ensembles for time series inference in economics. *Journal of Asian Economics* 6:955–978.

Vinod, H.D. (2016). New bootstrap inference for spurious regression problems. *The Annals of Applied Statistics* 2:317–335.

Vinod, H.D., López-de Lacalle, J. et al. (2009). Maximum entropy bootstrap for time series: the meboot R package. *Journal of Statistical Software* 5:1–19.

Figures

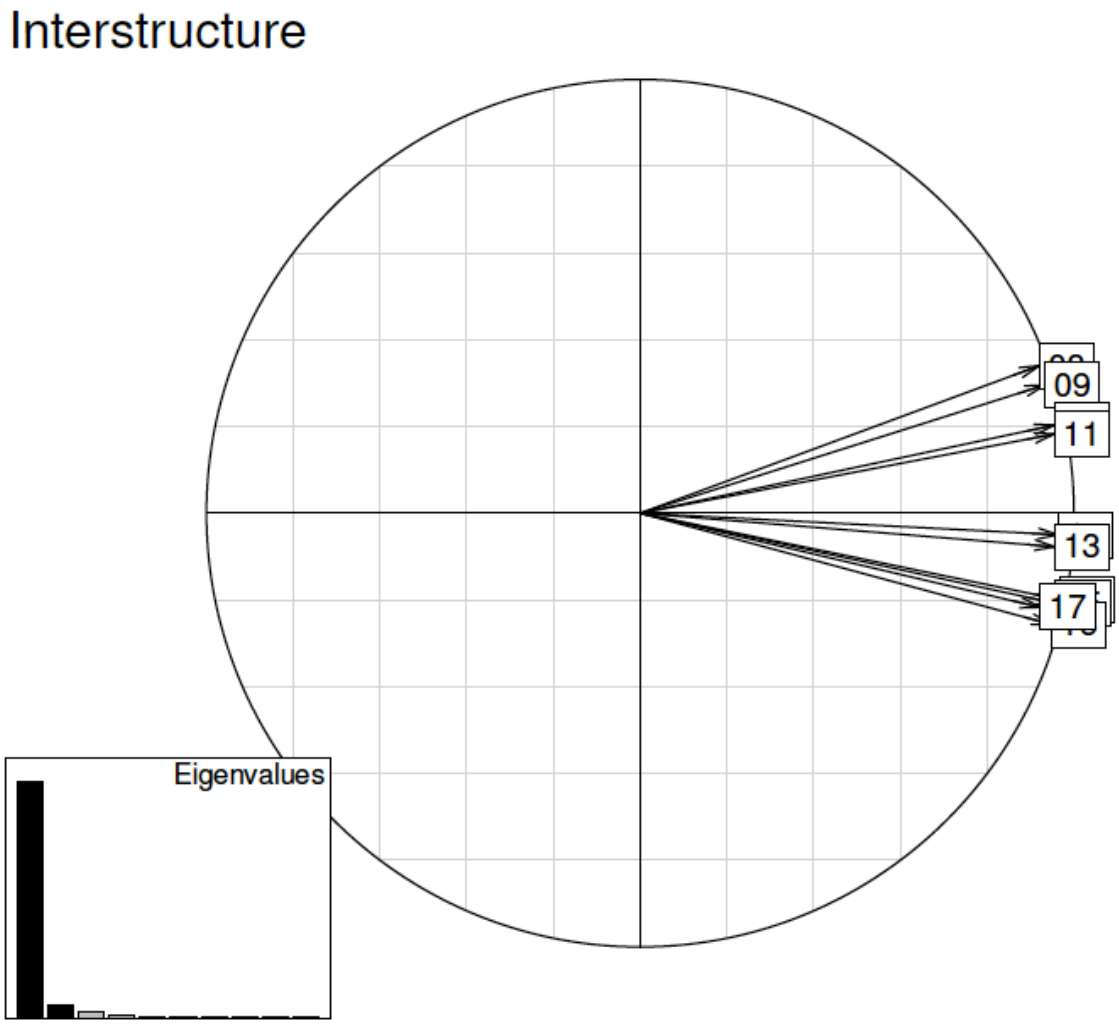

Figure 1

Graphical representation of the C matrix

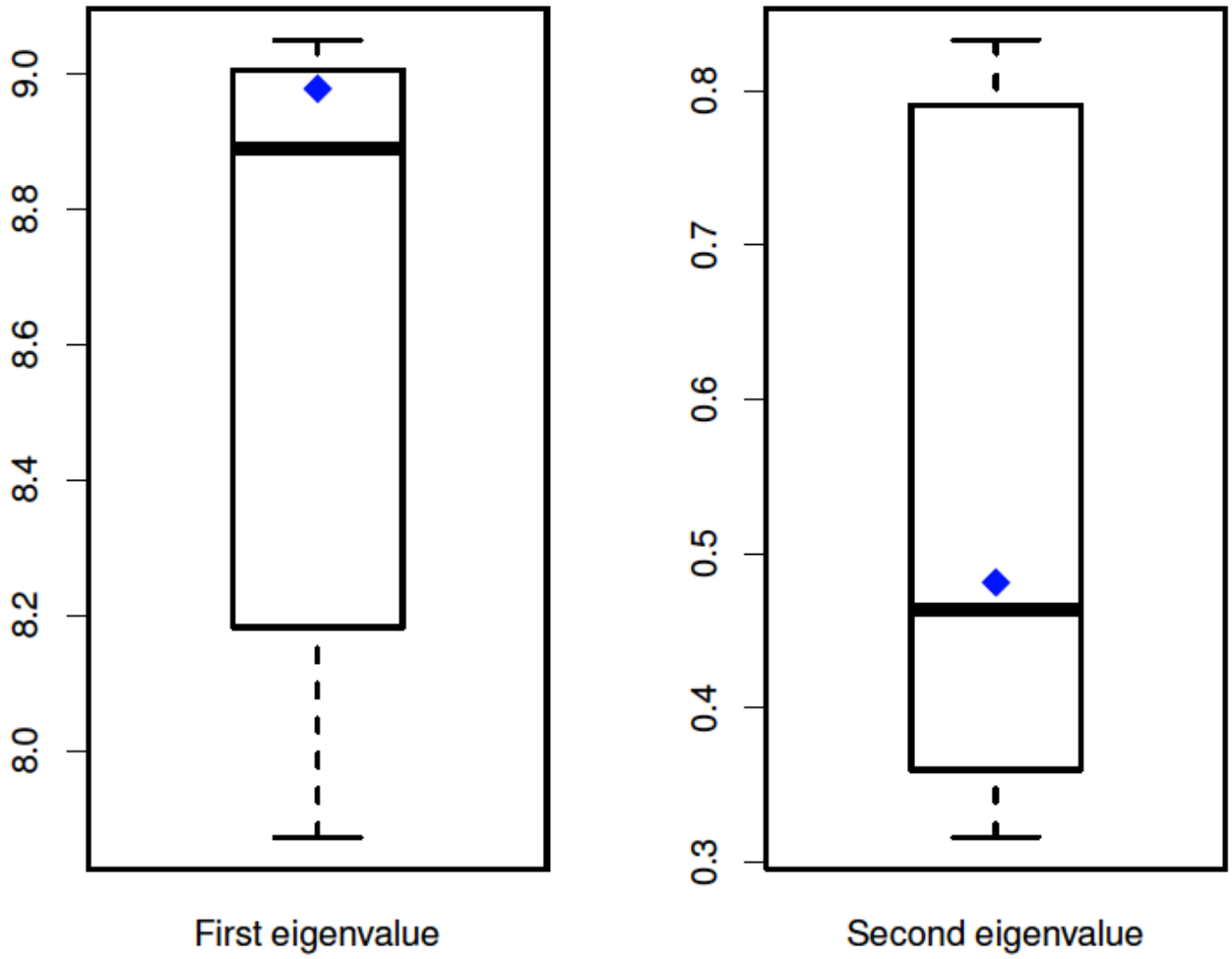

Figure 2

Quantile-based boxplot (0.05-0.95) of the first two eigenvalues bootstrap distribution

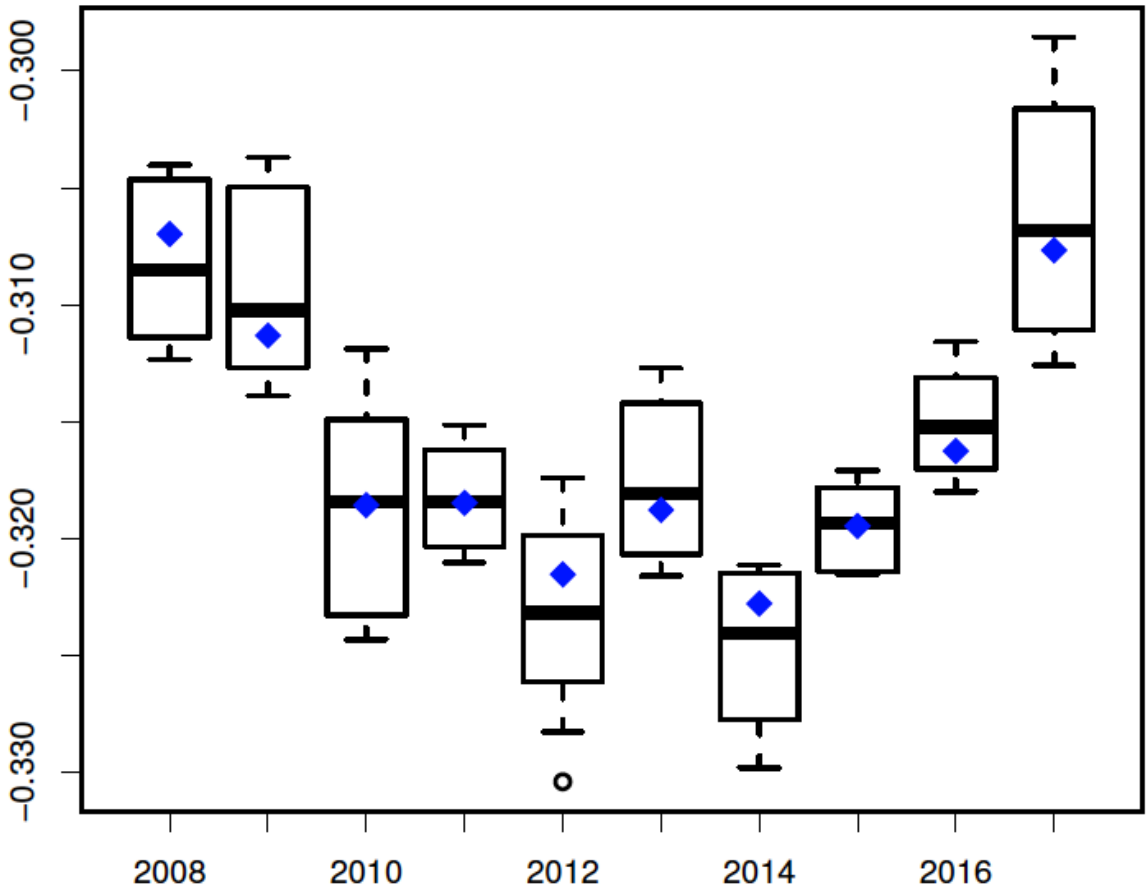

Figure 3

Quantile-based boxplot (0.05-0.95) of the first eigenvector bootstrap distribution

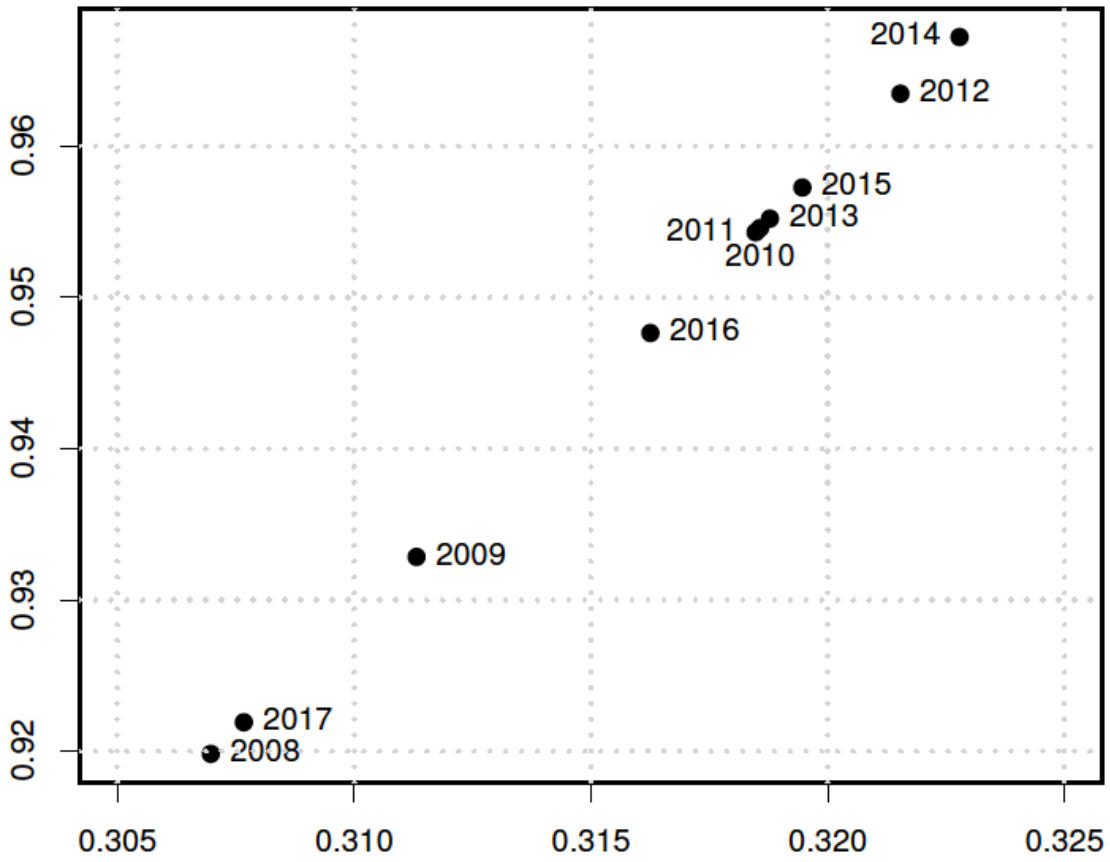

Figure 4

Typological values of the observed data matrices

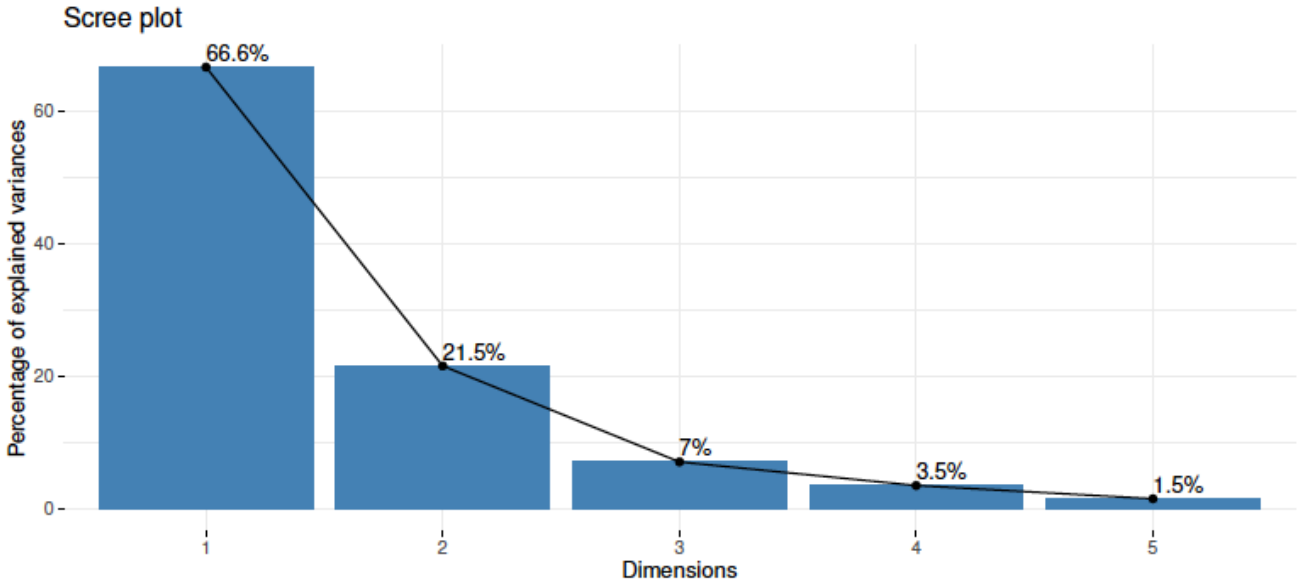

Figure 5

Scree-plot of the Compromise matrix

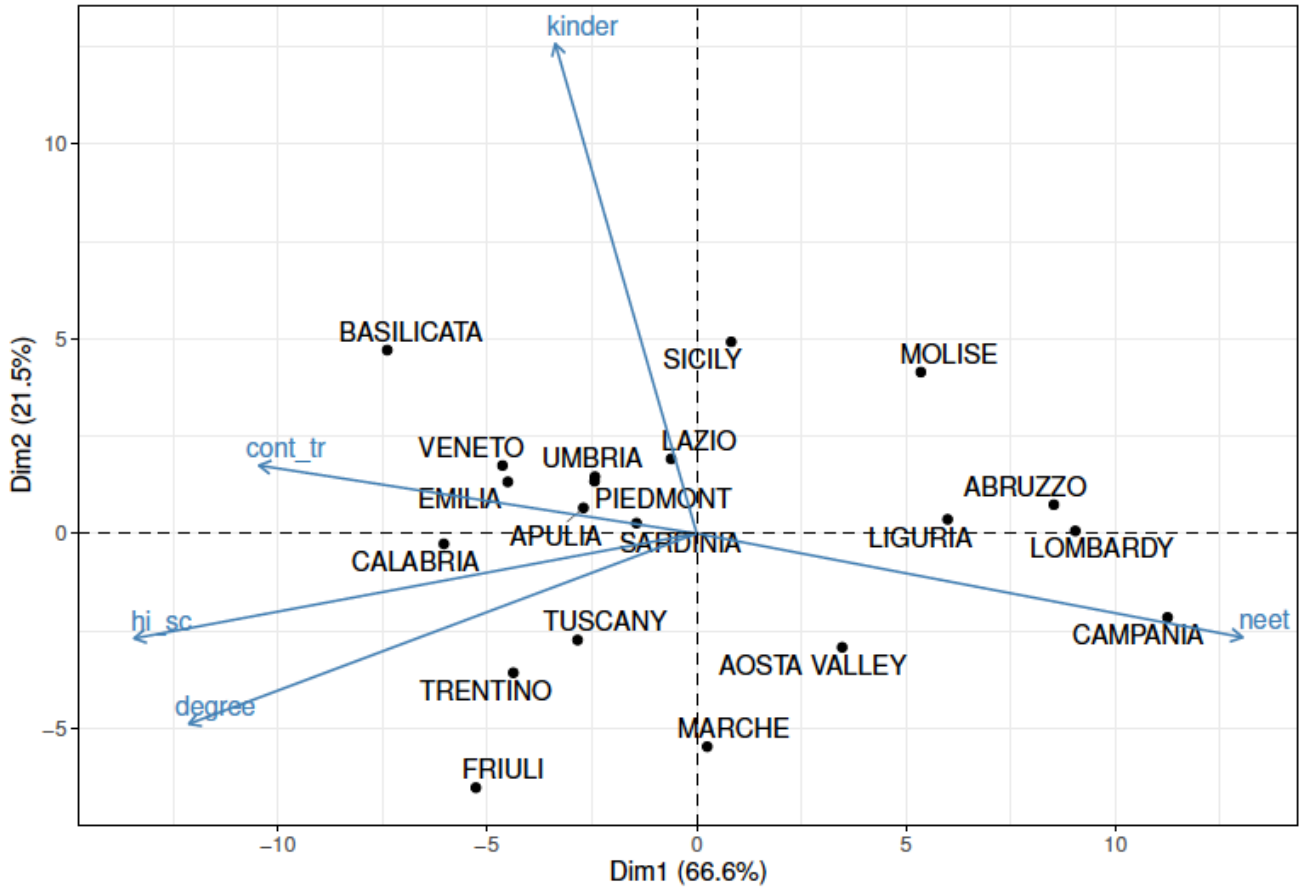

Figure 6

Biplot of the Compromise matrix

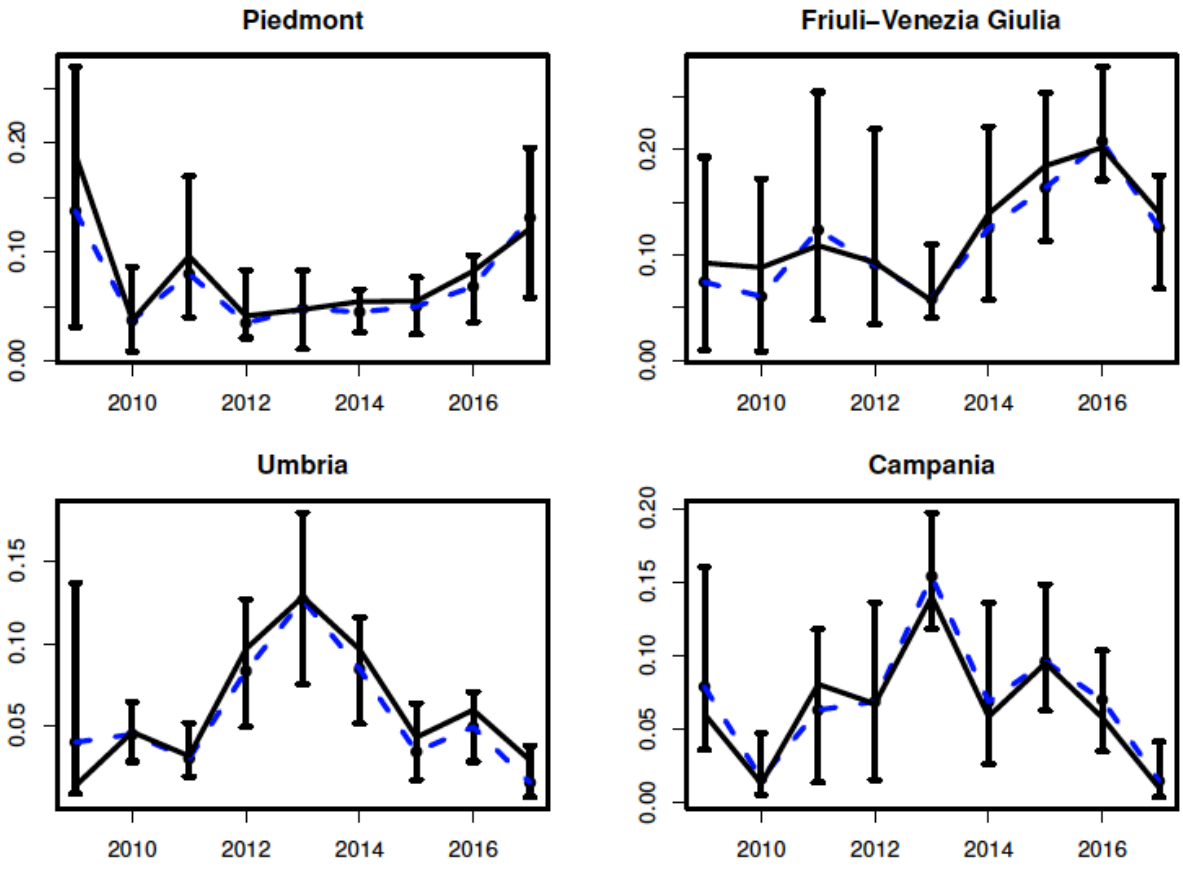

Figure 7

Trajectories of the most influential regions