AMB Àrea Metropolitana de Barcelona

Working Paper

New Models and Indices of Urban Network Sustainable Progress

Application at Regional and Megaregional Levels

Phase I - Report

Department of Ecology and Territory

Barcelona Institute of Regional and Metropolitan Studies

Department of Mathematics

Autonomous University of Barcelona

Project CP 6.1.2

December 2017

Coordinator

Joan Marull

Authors

Joan Marull

Mercè Ferrer

Alan Palacio

Aureli Alabert

Technicians

Francesc Coll

Núria Ruiz

Acknowledgements

This work was supported by the Metropolitan Area of Barcelona (AMB). The research was developed by the Department of Ecology and Territory of the Institute of Regional and Metropolitan Studies of Barcelona (IERMB), and the Department of Applied Mathematics of the Autonomous University of Barcelona (UAB).

Table of Contents

Abstract

The present study aims to understand the socioecological implications of a new economic unit of analysis consisting of networks of cities at regional and megaregional scales.

The proposed urban network sustainable progress indices, based on different conceptual scenarios, relies in three interrelated factors, the economic growth, the social cohesion and the urban ecology.

One of the most important concept of the paradigm of sustainable progress (beyond the GDP) is that economic growth is only a branch of this development, and therefore, the social and ecological factors are strategic elements of urban systems.

Even though there are formal ways of measure the sustainable progress (like the United Nations sustainable development index), the goal of this study is to propose a mathematical model, derived from Eurostat and satellite databases, to analyze different conceptual scenarios of urban growth in Europe.

To accomplish this objective, statistical methods will be applied to infer empirical models from data, that later are reconstructed according to conceptual information, detecting the hidden relationships between different variables, provide a way in which new unseen observations could be characterized according to this method, and to study different scenarios.

The statistics used are component analysis, factor analysis, cluster analysis, structural analysis, and a probabilistic method for the indices development. Besides the fact that the analysis are built upon a hypothesis, the model is built upon data.

The method used is designed in a way the results have a conceptual background. The indices – one for each scenario- are made in a way that their values provide an instinctively conceptual meaning. The index value is not an abstract number, but rather provides information about the position of an observation compared to the rest of them.

The index is a static measure, but allows tracking how the urban network evolves through time, influenced by the three before mentioned factors of the sustainable progress.

This study contributes to the debate on the essential properties of a regional and megaregional economy, optimizing the socioecological performance at the level of networks of cities.

Urban networks at regional and megaregional scales have emerged because of the densification and acceleration of economic processes. They concentrate a huge amount of world population, production, innovation and wealth, although they are also important consumers of resources.

The sustainable progress of urban networks is then a relevant issue. Our research question is whether existing regions and megaregions will evolve towards a sustainable path. This question is relevant and has direct implications for pro-active policy and planning.

Keywords

Beyond GDP, sustainable progress, urban network, socioecological system, megaregion, Europe

Introduction

Beyond GDP

The Gross Domestic Product (GDP) index was developed in the 1930s and 1940s amid the Great Depression and the First World War (Kuznets, 1934). Even before the United Nations (UN) began requiring countries to collect data to report national GDP, Simon Kuznets had warned against associating its growth with well-being.

The GDP index measures mainly market transactions. It ignores environmental impacts and social inequality. Yet since the end of the Second World War, promoting GDP growth has remained the primary national policy goal in almost every country (Van der Bergh, 2009).

In the meantime, researchers have become measuring what actually does make well-being. The environmental and social effects of GDP growth can be estimated (Kubiszewski, 2013). The psychology of human well-being can be analyzed comprehensively and quantitatively (Seligman, 2012). Therefore, many studies has produced alternative measures of sustainable development.

The chance to dethrone GDP is now in sight. In 2015, the UN announced the Sustainable Development Goals, a set of international objectives to improve global well-being. Developing integrated measures of urban sustainable progress attached to these goals offers the opportunity to define what well-being means, how to measure it, and how to achieve it.

Missing the opportunity of sustainable urban progress would condone growing inequality and the continued destruction of the natural capital on which all life on earth depends. Hence, GDP is dangerously inadequate as a measure of quality of live (Constanza et al., 2014).

Urban networks

In regional socioecological policy, there is the usual tradeoff between economic development and environmental quality (Batabyal and Nijkamp, 2009). In this study, we aim to show some evidence that it is possible to combine both objectives. Large urban agglomerations could rely on macroeconomic models more based on knowledge than on consumption of resources: this is the main challenge of sustainable progress.

Cities are not isolated systems, but are connected together to form networks. Traditionally, urban systems have been studied from a hierarchical point of view (Christaller, 1933). According to this view, the urban dimensions would reflect the existence of a hierarchy of goods and services, which would express the size of the market. However, later studies have shown that some urban structures are a mix of hierarchical (vertical) and non-hierarchical (horizontal) structures, in the form of “networks of cities” (Pred, 1977).

Networks of cities have been defined as a set of relationships between complementary or similar centres, relationships that allow the emergence of economies of specialization (division of labour) or the formation of economies of synergy (cooperation and innovation) (Camagni, 2005). In these networks, cities benefit from economic advantages stemming not only from their own dimension, but also from the size of the whole network. Therefore, the paradigm of urban networks implicitly suggests extending the scope of analysis beyond the metropolitan area.

The importance of this scaling is critical to achieve positive results in terms of economic efficiency, social equality and environmental sustainability. Night-time light (NTL) satellite data allows to analyze, on a global scale, the evolution of networks of cities towards structures that already exceed the metropolitan scale (Zhang and Seto, 2011), systems that are called urban

regions or even “megaregions”. Megaregions are emerging global economic units, the result of the concentration of production facilities, innovation and consumer markets (Florida et al., 2007). Their development is based on socioeconomic dynamics, processes that cause profound changes on their environment and accelerate global change (Grazi et al., 2008).

The demand for land to accommodate housing, economic activity, infrastructures and transport networks produces a significant pressure on the environment (Williams et al., 2000). Moreover, urban sprawl has been poorly managed (Breheny, 1992), which has led serious problems in quality of life and the ecological functioning. Applications developed on the basis NTL data framed by artificial satellites (Doll, 2008), allow us to define urban extensions, calculate energy consumption or estimate economic activity at the level of regions and megaregions (Figure 1).

Figure 1 Night-time light (NTL) satellite data (NASA, 2007)

The key factor in regional and megaregional development is that urban growth does not start from a central agglomeration towards an empty area, but can instead encompass many other smaller urban areas and also some of a similar size to the central one (Ross, 2009). Consequently, a metropolitan region and a megaregion trend to polycentric expanding urban networks.

Therefore, agglomeration economies can be achieved in urban polycentric networks (Marull et al., 2015), such as economies stemming from concentrated and diversified economic and social structure, and economies fed by the relationships that are developed in the network of cities. Thus, the study of economic growth, social cohesion and urban ecology, performed only through data from the city or country can be misleading.

This study departs from the question whether these concentrations of activity in expanding urban networks can contribute or not to the sustainable progress of European societies, that is, if metropolitan regions and megaregions should be considered as a problem or a potential solution for a more “inclusive growth”. Inclusive growth expands upon traditional economic

growth models to include focus on the equity of health, human capital, environmental quality, and social protection (Hasmath, 2015).

The main hypothesis of this study is that megaregions have emerged along with greater dissipation of energy (primary energy consumption –PEC) and the establishment of networks of cities (urban growth –URG, urban density –URD), that are more efficient in their economic activity (gross domestic product –GDP), knowledge creation (patents –PAT), and social cohesion (gross rate of employment –GRE). Suggesting the need for a new geographic scale to explore urban sustainability (Marull et al, 2013).

In short, the object of this study is to approach the evolution of towns, cities and metropolis towards more complex urban systems at regional and megaregional level, and analyze the consequences of these units of analysis in the context of inclusive growth, using new models and indices of urban network sustainable progress.

In order to develop the analysis we firstly study the statistical relationships between economic, social, ecological and urban variables of European regions (NUTS 3) in the period 1995-2010, considering that some of them pertain to a megaregion; and secondly, we develop ecological macroeconomic models and indices of urban network sustainable progress applied at regional and megaregional scales.

Sustainable progress

There is broad agreement that society should attempt for a high quality of life that is equitably shared and environmentally sustainable. Several reports have concluded that GDP is dangerously inadequate as a measure of quality of life, including those published by the Commission on the Measurement of Economic Performance and Social Progress (Stiglitz et al., 2009), the Center for the Study of the Longer-Range Future (Constanza et al., 2009) and the European Commission's ongoing Beyond GDP initiative.

However, GDP remains fixed, and some economic interest groups are partly responsible (Van den Bergh, 2009). However, much of the problem is that no alternative measure emerges as a clear successor. Creating that successor will require a sustained, transdisciplinary effort to integrate metrics and build consensus.

One opportunity for doing this consensus is the creation of the UN Sustainable Development Goals (SDG), a process that is now under way to replace the Millennium Development Goals (MDG). Established in 2000, the MDG comprise eight basic targets that include eradicating extreme poverty and establishing universal primary education, gender equality and environmental sustainability.

Currently, both the MDG and the SDG are only lists of goals with isolated indicators. The SDG process should be expanded to include comprehensive and integrated measures of sustainable well-being (Griggs et al., 2013). There are significant obstacles to doing this, including bureaucratic inertia, the tendency of academia to work in isolation, and political interests.

The successor to GDP should be a new set of metrics that integrates current knowledge of how social, economic and ecological factors collectively contribute to establishing and measuring sustainable progress. The new metrics must garner broad support from stakeholders. It is often said that what you measure is what you get. Building the future requires that we measure what we want, remembering that it is better to be approximately right than precisely wrong.

Our assumption is that urban networks, measured at regional and mega-regional levels, are complex adaptive systems, involving many variables and dimensions (territorial, social, economic, ecological), whose relations must be taken into account to understand the processes of change and to reverse certain trends through integrated sustainable urban planning.

For this work, a database composed by three socioeconomic traditional variables (GDP, GRE, PAT), obtained from the European Office for Statistics (Eurostat), and three more novel variables (PEC, URG, URB) which are constructed using satellite imaging, are used (Marull et al., 2013).

This technique based on satellite databases allows imputing values from a larger territory (i.e. countries) to smaller units of analysis (i.e. regions or megaregions) with enough precision to approximately measure urban network development.

The period of analysis is from 1995 to 2010, according to the satellite data availability. The new sustainable progress indices are constructed according to conceptual factors of interest, which are economic growth, social cohesion, and urban ecology.

Nevertheless, economic growth, social cohesion and urban ecology factors cannot be measured directly, but through the variables that are affected by them, which are the observable variables. Thus, factor analysis will be used to detect and measure these hidden factors of interest. They are called “scores”, and are used to: represent a parsimonious summary of original data; be more reliable than the observed variables; and get a measure of the latent factors.

These scores do not give any information a priori, as it would provide, for example, the Gini index of inequality. To do this, a statistical analysis is applied in order to transform the distribution of the scores (which probably does not fit any known distribution) into a Laplace distribution, and then estimate its density function. This way, the score is transformed into a number that reflects the position (percentile) of a given observation in the overall ranking.

In the present study, factors and variables of interest are analyzed in statistical terms. Observations are represented in graphics showing the variable average and the dispersion of values by the units of analysis –regions and megaregions. Several imputation methods and its consequences are studied and implemented. The factor analysis on the selected variables is applied, and the scores of this analysis are plotted. To verify the existence of specific homogenous groups a cluster analysis is made. The sustainable progress indicators are developed and finally, several urban network scenarios are studied.

Objective

The objective of the study is to develop indices of urban network sustainable progress (beyond the GDP) based on mathematical models. The models allow a better understanding of the optimal relationships of economic, social and ecological factors at different spatio-temporal scales. The study provides indices of the main scenarios that rule urban sustainable progress at European regional and megaregional levels, as well as tools for urban and regional planning.

Data Handling

Urban networks at megaregional scale

There are several methodologies that allow to define megaregions based on census data and a structured set of criteria (such as transport networks, population growth and land consumption) (Lang and Dhavale, 2005; Dewar and Epstein, 2007).

We use night-time light (NTL) satellite data to monitor the dynamics of urbanization at megaregional level (Marull et al., 2013). One of the benefits of NTL data, in front of national statistics (i.e. European NUTS 3) is that it allows delineating and estimate indicators for functional analysis units that do not necessary coincide with administrative boundaries.

We use the reference method proposed by Florida et al. (2007). According to this approach, a megaregion needs to fulfil two main criteria: it must be a contiguous lighted area with more than one major city or metropolitan region; and it must produce more than $ 100 billion in LRP (Light-based Regional Product). By that definition, there are 40 megaregions in the world, covering 18% of world population and producing 66% of its economic activity.

The main database for delineating megaregions is the series of images produced by the sensor of the DMSP-OLS and publicly distributed by the National Geophysical Data Center of NOAA (National Oceanic and Atmospheric Administration). The images used are in GeoTiff format with a spatial resolution of approximately 1 km^2 per pixel (30'). Each pixel sensor of the satellite assigns a specific value of light intensity. This value is explained as DN (Digital Number), has a radiometric resolution of 6 bits, and can vary between 0 and 63.

Given the definition of megaregion as an area characterized by a substantial physical contiguity of human settlements, a minimum threshold of light intensity (DN = 8) and a minimum distance of 2 km grouping have been introduced. Using this methodological procedure in accumulative way for annual series of historical data (from 1995 to 2010), we have measured the evolution of the 12 megaregions that exist in Europe (Figure 2).

It is important to stress that the urban network delineation approach of the megaregions used here is only a good approximation because it is not possible to define—owing to various technical problems (Small et al., 2005)—an exact relationship between bright areas detected by satellite and urbanized areas. However, the use of a single criterion and a common database to define different megaregions is a guarantee that the defined entities are effectively comparable.

The name of the 12 European megaregions is abbreviated for an improved understanding of the forthcoming graphics (Table 1).

Abbreviation	Name
NMR	No Megaregion
VIB	Vienna-Budapest
FRG	Frankfurt-Stuttgart
AMB	Amsterdam-Brussels-Antwerp
PRA	Prague
BER	Berlin
LIS	Lisbon
MAD	Madrid
BAL	Barcelona-Lyon
PAR	Paris
RMT	Roma-Milan-Turin
LON	London
GLB	Glasgow-Edinburgh

Table 1 Megaregions names

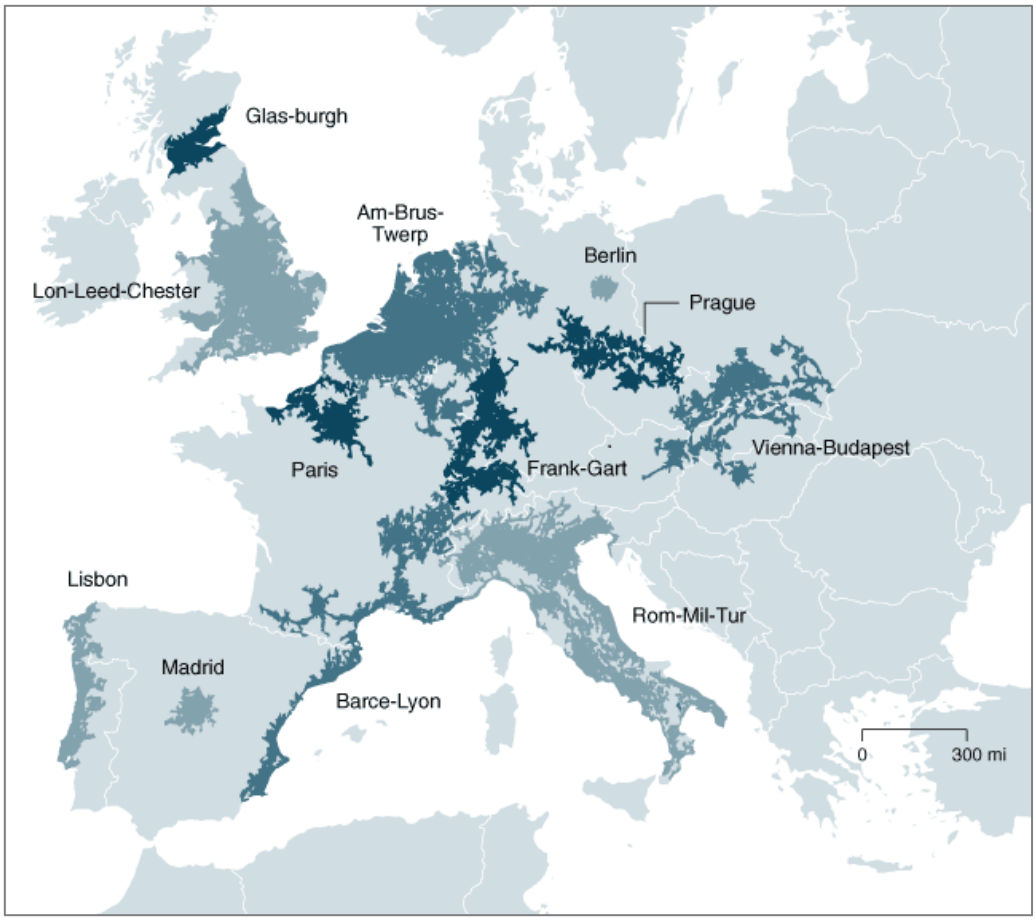

Figure 2 European Megaregions (Florida, 2008)

Territorial units at regional scale

The classification of territorial units for statistics (NUTS, for the French Nomenclature d'Units Territoriales Statistiques), is a geocode standard for referencing the administrative divisions of countries for statistical purposes. The standard was developed by the European Union.

There are three levels of NUTS defined, with two levels of local administrative units (LAUs) below. Not all countries (NUTS 0) have every level of division, depending on their size. For example, one of the extreme cases is Luxembourg, which has only LAUs; the three NUTS divisions each correspond to the entire country itself. The NUTS classification is a hierarchical system, dividing the territory of the EU in:

- NUTS 1: major socio-economic regions
- NUTS 2: basic regions for the application of regional policies
- NUTS 3: small regions for specific diagnoses

The goal is to study the relationships of urban network sustainable progress variables at the smallest scale possible; this is at NUTS 3 level (Figure 3).

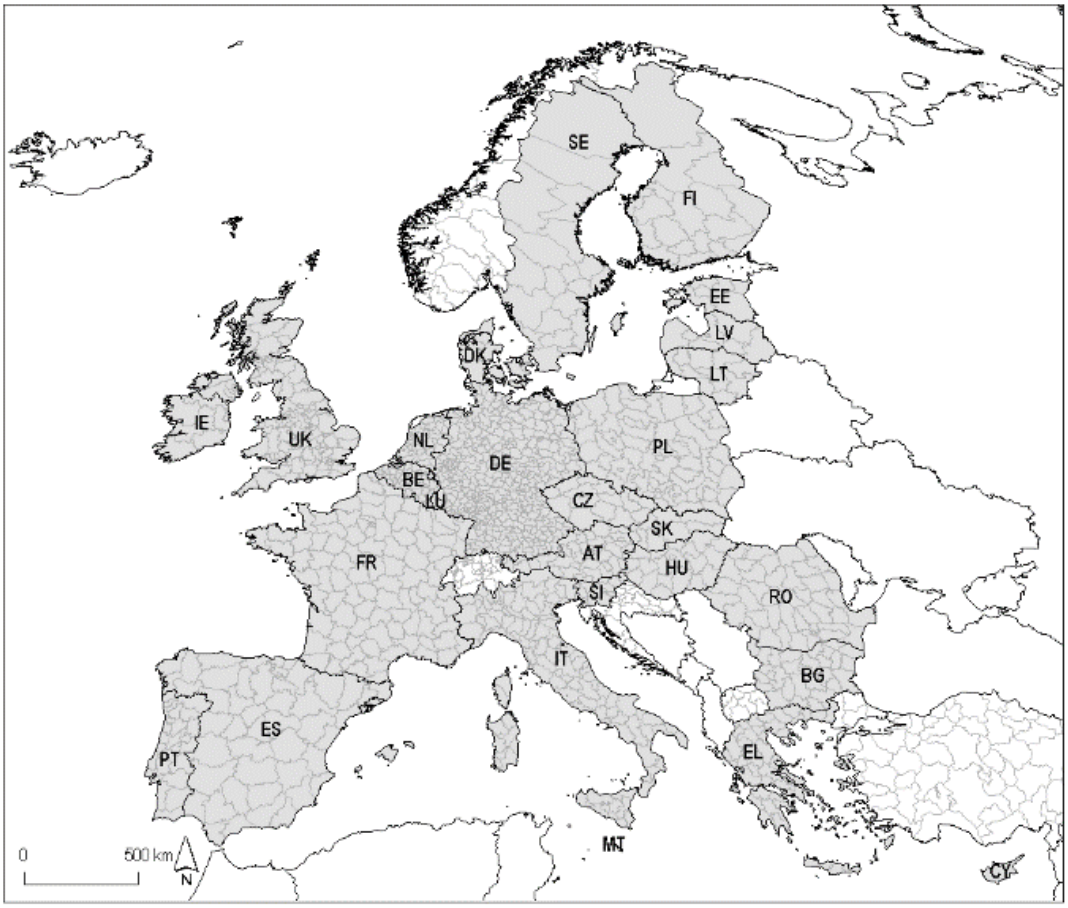

Figure 3 European territorial units for statistics at country (NUTS 0) and regional level (NUTS 3)

Period of analysis

Given the limitations on terms of data availability, the period of analysis is from 1995 to 2010 (Figure 4). This is mainly because the satellite imaging process requires to apply multiple layers of filters to identify the sources of information. The data is collected annually but, in order to simplify the analysis, we usually will represent four time points (1995, 2000, 2005 and 2010).

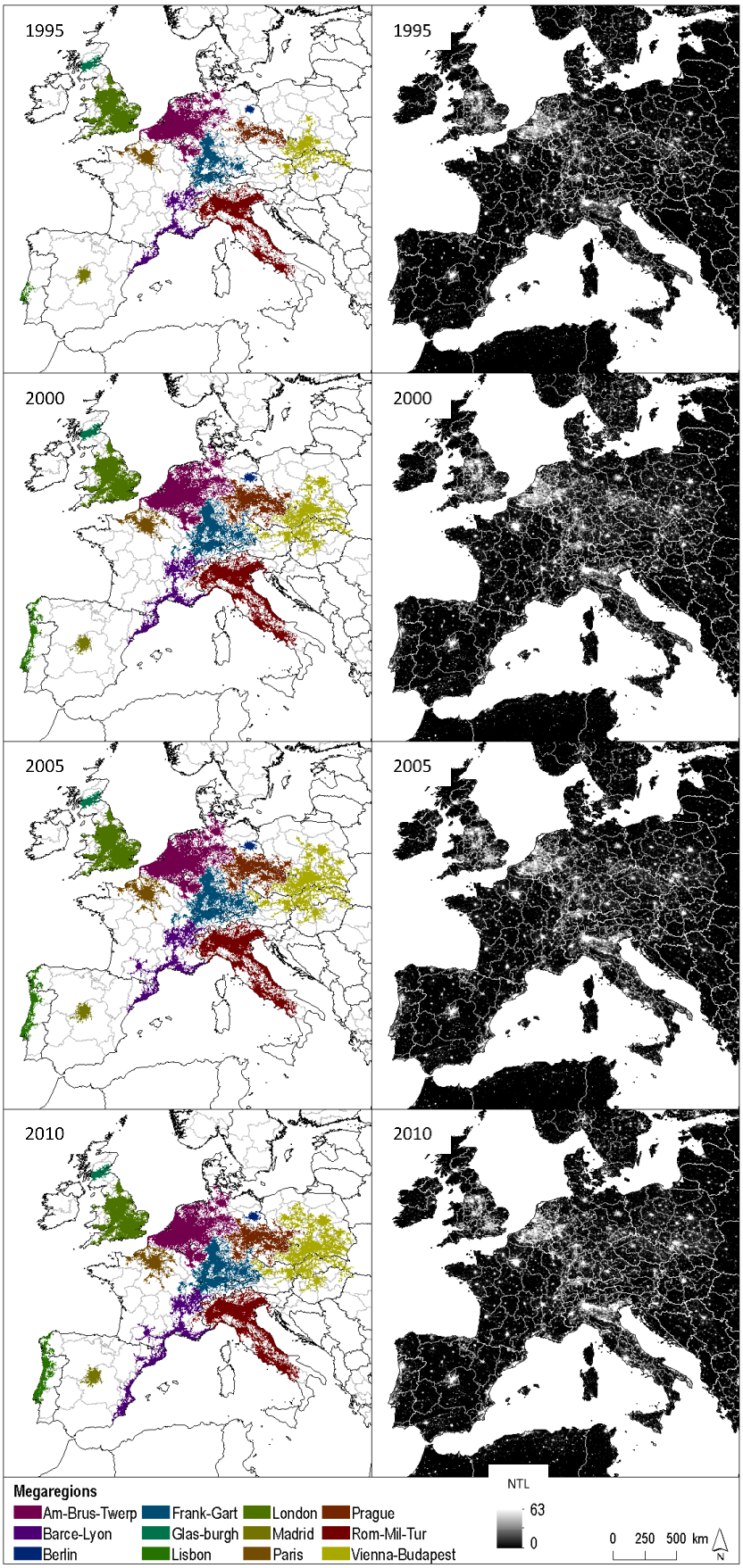

Figure 4 European megaregions' growth and changes in NTL satellite data (1995-2010)

Selected variables

A number of specific variables designed to evaluate the behavior of the units of analysis (countries, regions –NUTS 3, and megaregions) in different time frames are created (Table 2).

Abbreviation	Variable description
GDPpc	Gross domestic product, per capita
GREpc	Gross rate employment, per capita
PATth	Patents applications, per thousands of people
PECpc	Primary energy consumption, thousands of oil equivalent per capita
URDpsk	Urban density, people living in urban area per square kilometer
URGpor	Urban surface, percentage of urban area per NUTS3
COUn	Country name, at NUTS0 level
MGAn	Megaregion name, if NUTS3 belongs
NUTS3	Region code, at NUTS3 level
Year	Years, between 1995 to 2010

Table 2 Variables description

GPDpc - Gross domestic product

- Name: Gross domestic product –at current market prices
- Scale: NUTS 3 regions
- Source: Eurostat
- Code: nama r e3gdp
- Oldest data: 2000
- Most recent data: 2011
- Unit: Thousands of euros –purchasing power standards (PPS)- per inhabitant

GREpc – Gross rate employment

- Name: Gross rate employment –as proxy of social inequality
- Scale: NUTS 3 regions
- Source: Eurostat
- Code: nama 10r 3empers
- Oldest data: 2000
- Most recent data: 2014
- Unit of measure: Thousands of employed persons –per inhabitant

PATth – Patents applications

- Name: Patent applications to European Patent Office (EPO) –as proxy of knowledge
- Scale: NUTS 3 regions
- Source: Eurostat
- Code: pat_ep_rtot
- Oldest data: 1977
- Most recent data: 2012
- Unit of measure: Number of patents –per thousand inhabitants

URDpsk - Urban population density

- Name: Urban population density –as proxy of urban form
- Scale: NUTS 3 region, using satellite data
- Source: Eurostat, NASA
- Code: demo r pjanaggr3
- Oldest data: 1990

- Most recent data: 2014
- Unit of measure: Number of inhabitants per square kilometer of illuminated surface

URGpor - Urban surface percentage

- Name: Urban surface –as proxy of urban extension
- Scale: Imputed from 1x1 square kilometer data, measured by satellite data
- Source: NASA
- Oldest data: 1992
- Most recent data: 2012
- Unit of measure: Percentage of illuminated surface by NUTS 3

PECpc – Primary energy consumption

- Name: Primary energy consumption –as proxy of resource consumption
- Scale: Imputed from NUTS 2 to NUTS 3 using the satellite illumination intensity
- Source: Eurostat, NASA
- Code: tsdcc120
- Oldest data: 1990
- Most recent data: 2015
- Unit of measure: Million tons of oil equivalent (TOE) –per inhabitant

Variables density

The distribution of the variables is analyzed by a function that computes and draws a kernel density estimate, which is a smoothed version of the histogram. It can be seen that none of the variables follows a distinguishable distribution (Figure 5). When an index that characterize each region is constructed, these distributions will need to be normalized.

The Figure 5 shows the averaged values, but in order to understand the context of the study, plots of the variables for each one of the independent regions of analysis are going to be made. This will allow us to understand in a better way the dispersion of values, the presence of outliers and the number of observations per megaregion.

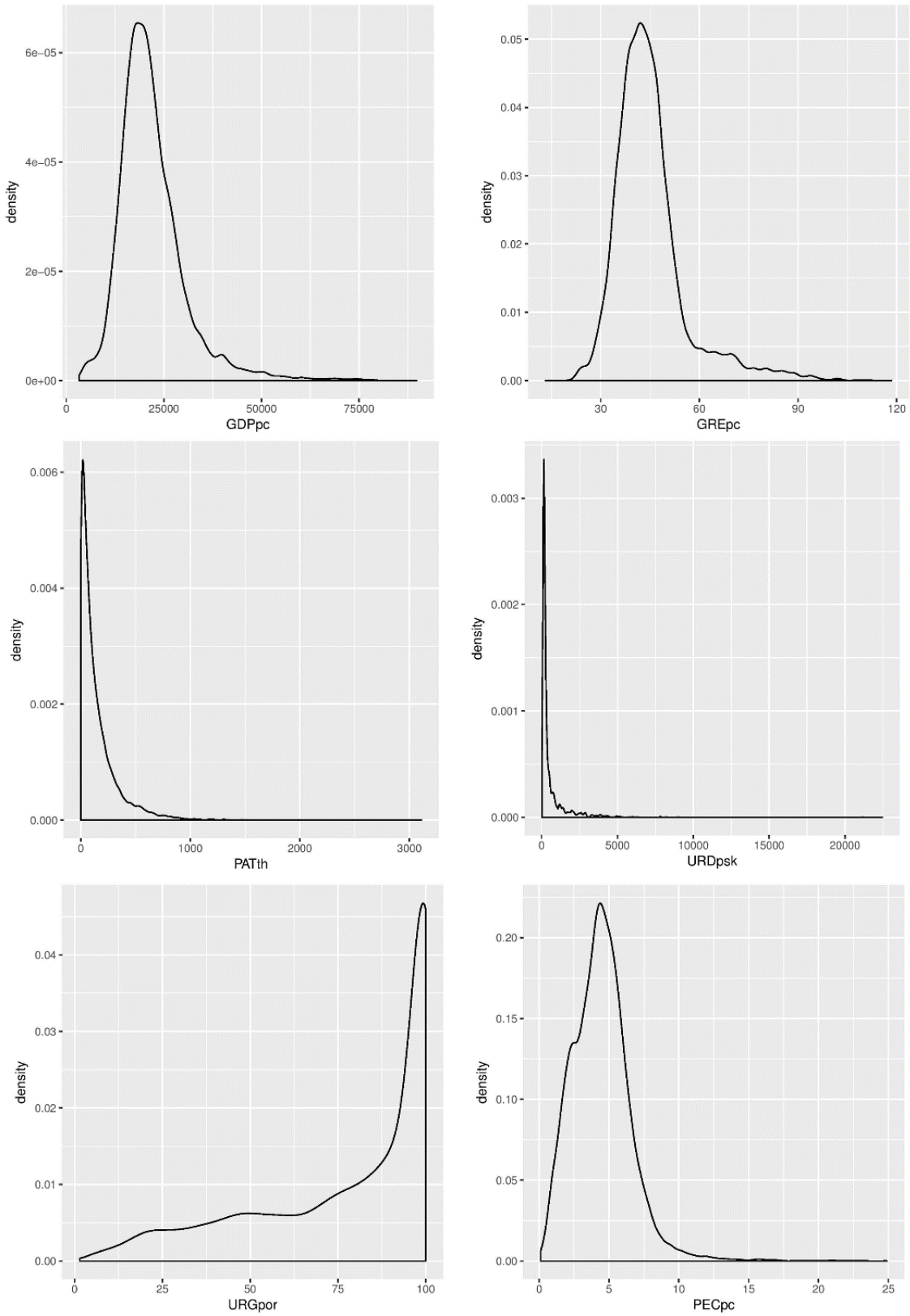

Figure 5 Variables density

Variables in time

The plot shows the behavior of the variables in time at NUTS 3 level, by country –to verify the results (from Figure 6 to Figure 11) and by megaregion –for the statistical proposal (from Figure 12 to Figure 17). Some interesting patterns can be seen in it, that later may be helpful to understand the results of the different statistical analysis.

Paris (PAR) is the megaregion with the highest average GDP per capita; the lowest is Lisbon (LIS) (Figure 12). Frankfurt-Stuttgart (FRG) has the highest number of employees per thousands of inhabitants; the lowest is Glasgow-Edinburgh (GLB) (Figure 13). The undisputed leader of patent generation is FRG; Madrid (MAD) and LIS are the last (Figure 14).

In terms of urban density, PAR has the highest values, followed in the distance by Berlin (BER); LIS is the mega-region with the lowest urban density (Figure 15). The megaregions with the highest percentage of urbanizated surface are Amsterdam-Brussels-Antwerp (AMB) and PAR; the lowest is GLB (Figure 16). The highest energy per capita consumption is matched between GLB, Prague (PRA), and FRG; the lowest are BER and MAD (Figure 17).

In general, GDPpc increase continuously, mostly without any change in its order, with a downturn in 2008 and a recovery by 2010. In 2009 the patents had a decrease probably because of the economic crisis. The differences in employment decreased during the years, probably because of the integration of the Eurozone.

Plots of variables by country

Gross domestic product –GDPpc

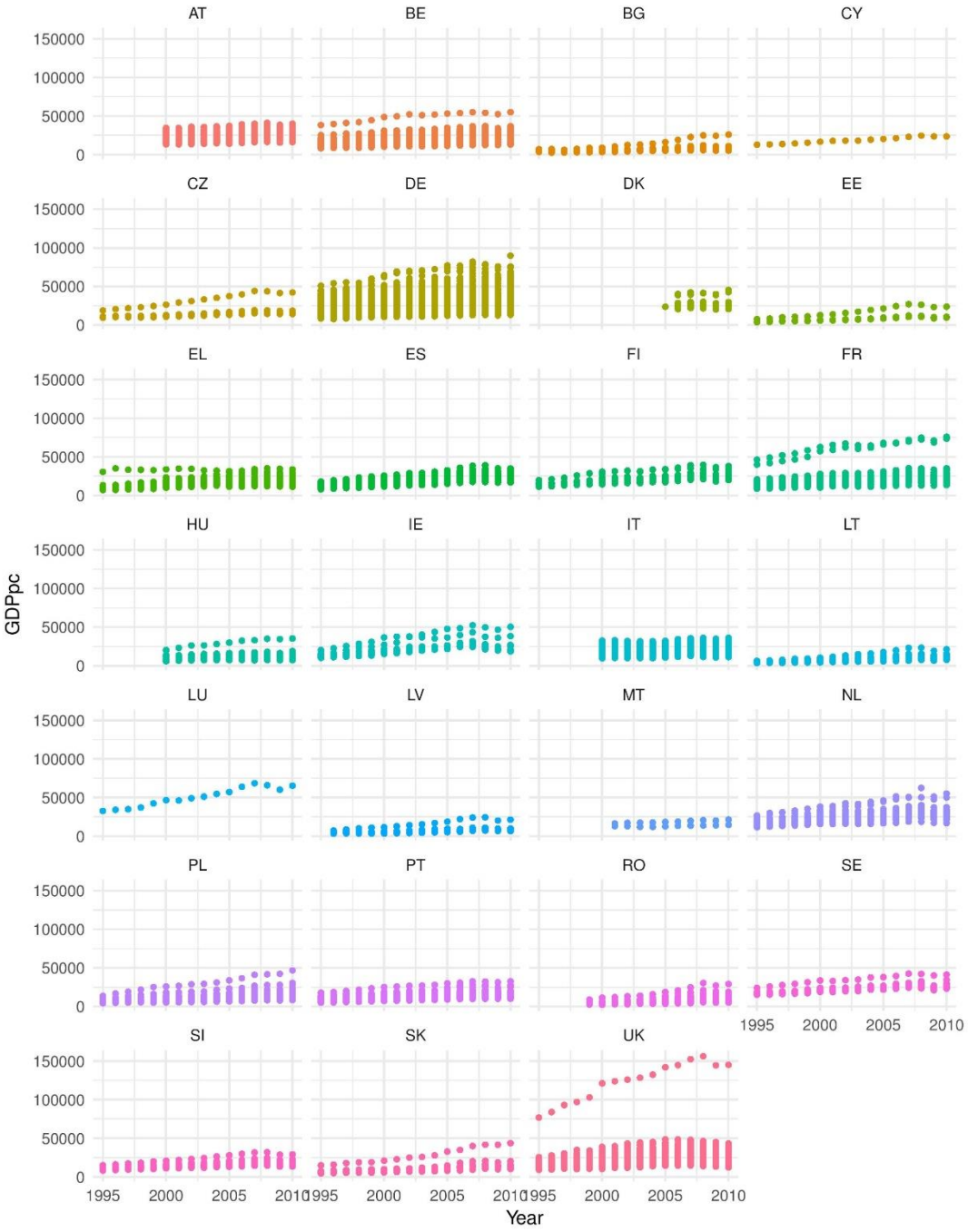

Figure 6 GDPpc at NUTS3 level by country

Gross rate employment –GREpc

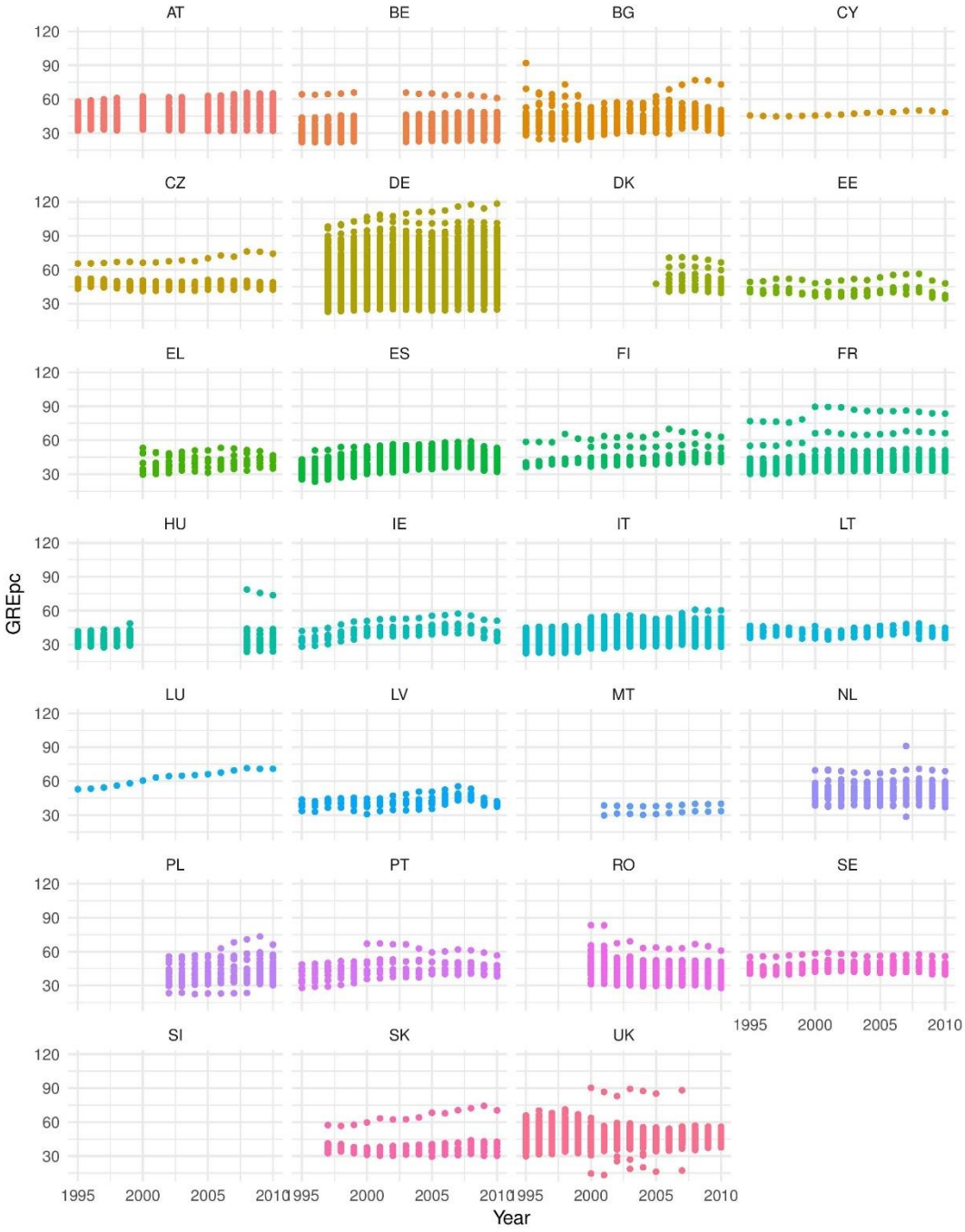

Figure 7 GREpc at NUTS3 level by country

Patents applications –PATth

Figure 8 PATth at NUTS3 level by country

Urban population density –URDpsk

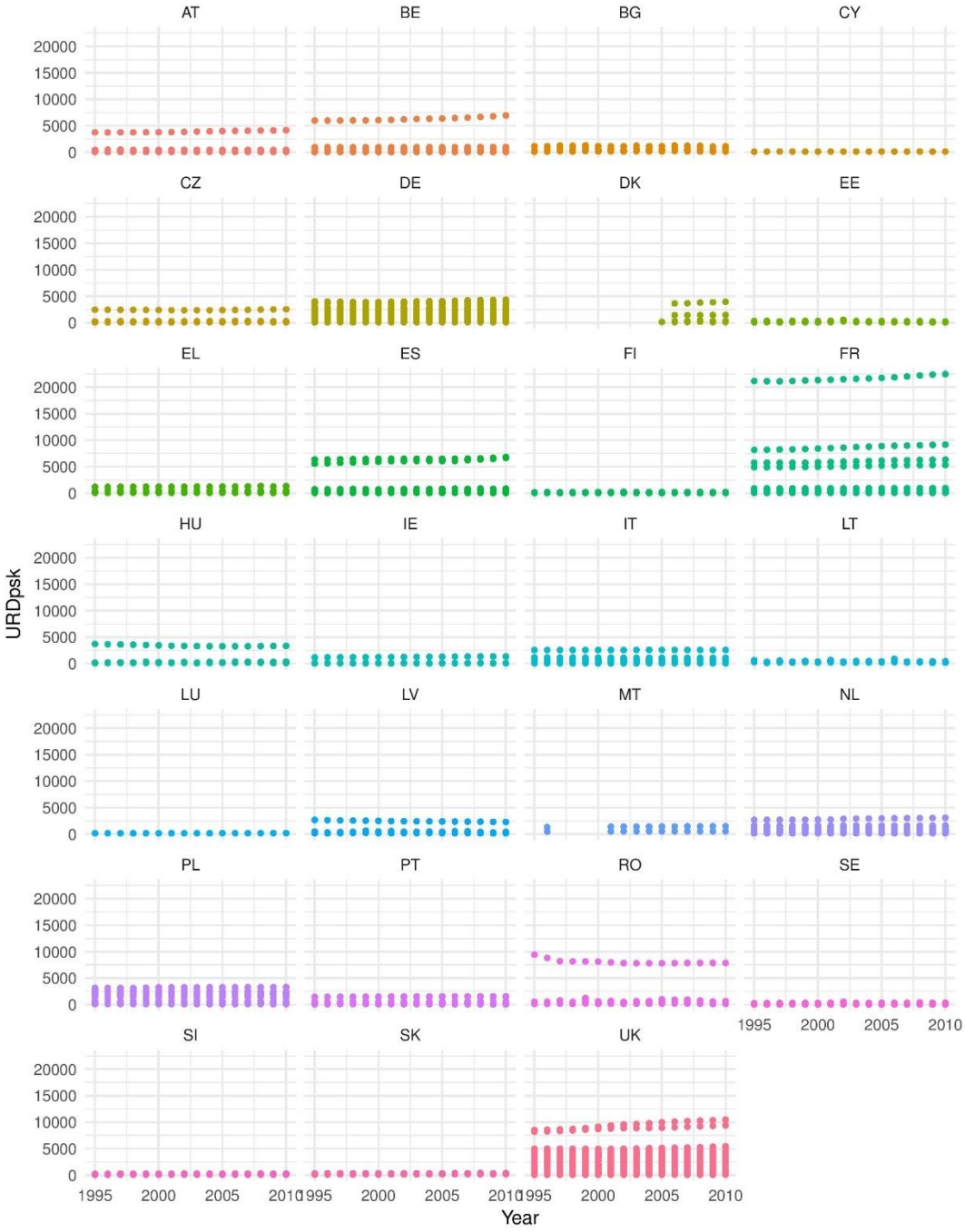

Figure 9 URDpsk at NUTS3 level by country

Urban surface percentage –URGpor

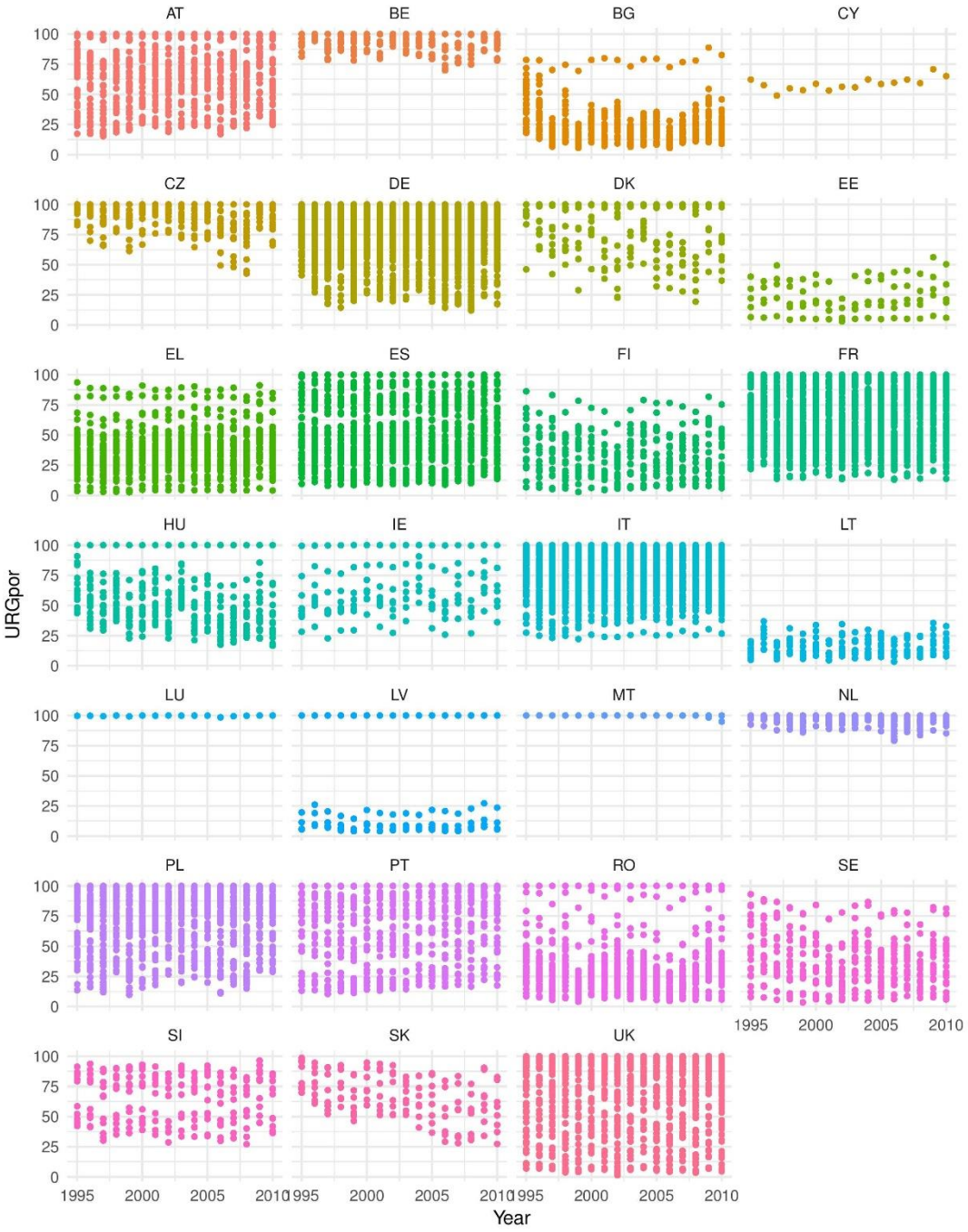

Figure 10 URGpor at NUTS3 level by country

Primary energy consumption –PECpc

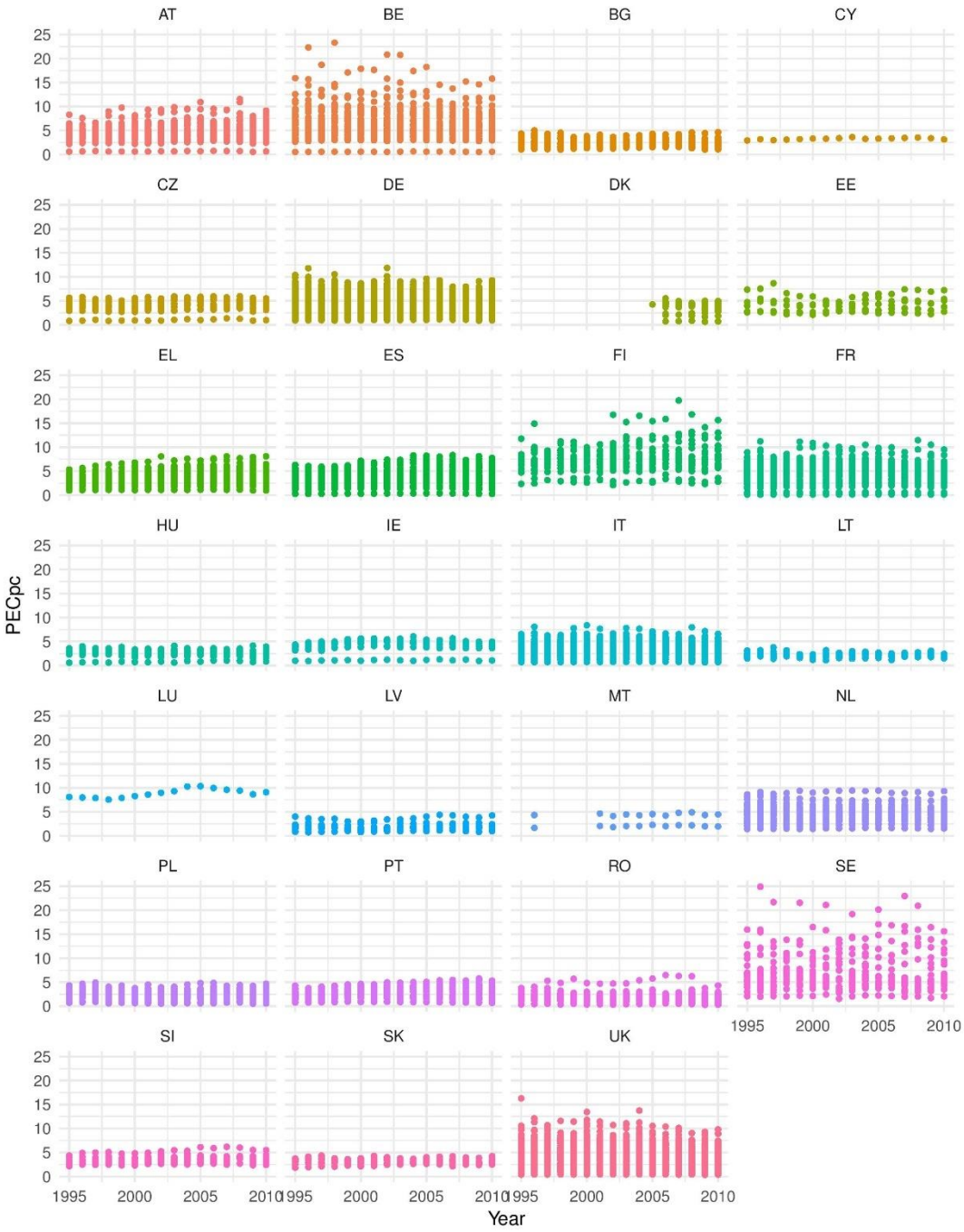

Figure 11 PECpc at NUTS3 level by country

Plots of variables by megaregion

Gross domestic product –GDPpc

Figure 12 GDPpc at NUTS3 level by megaregion

Gross rate employment –GREpc

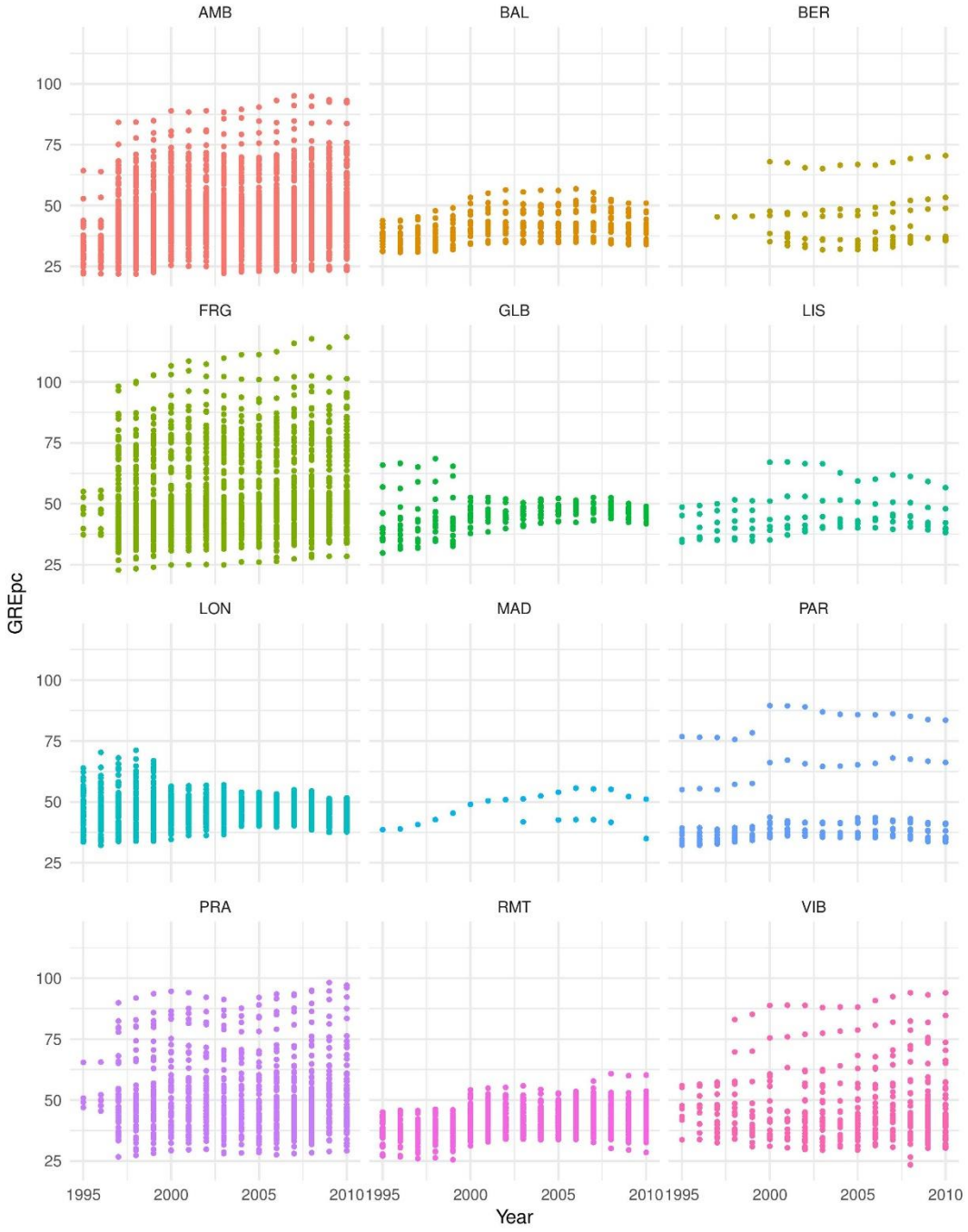

Figure 13 GREpc at NUTS3 level by megaregion

Figure 14 PATth at NUTS3 level by megaregion

Urban population density –URDpsk

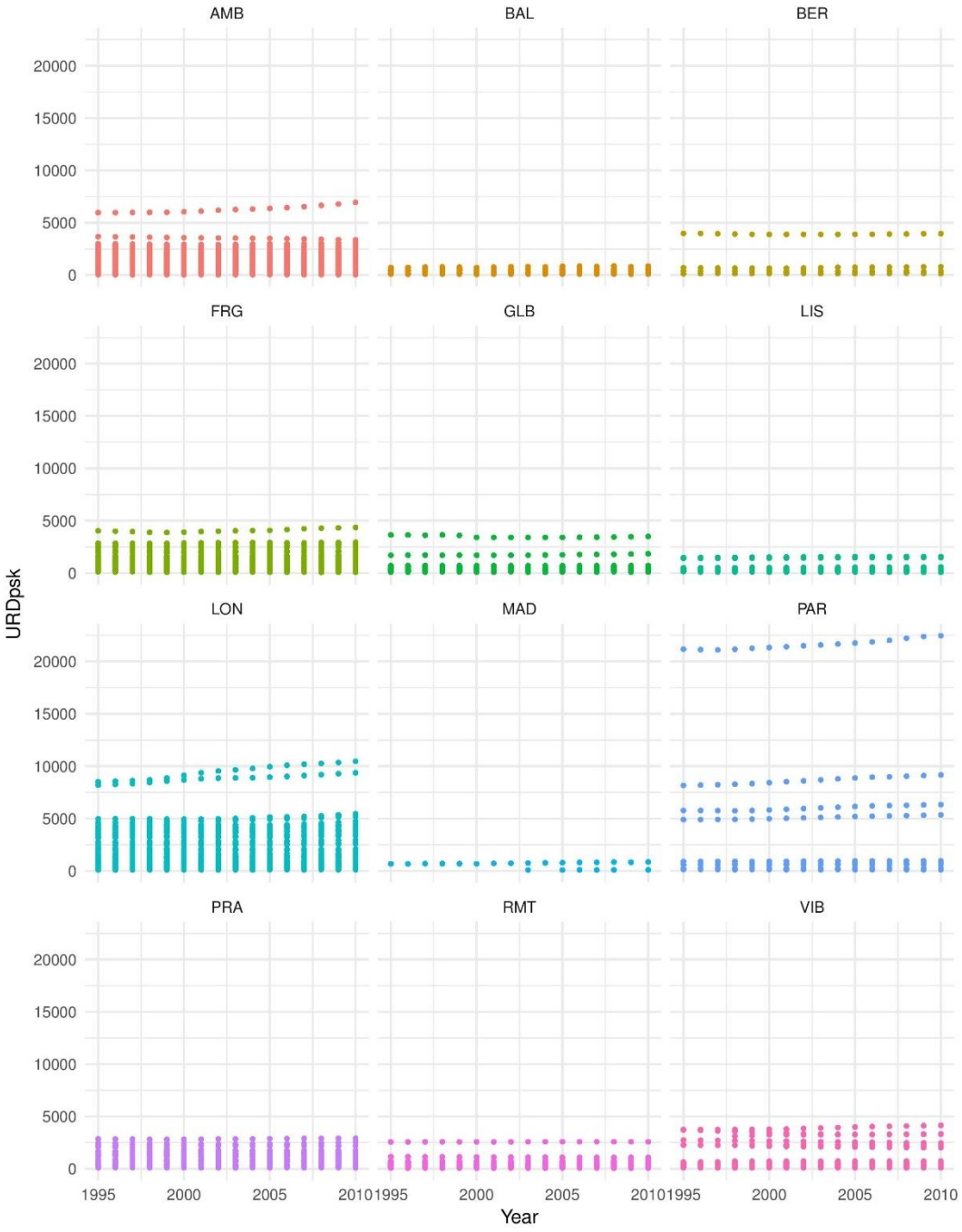

Figure 15 URDpsk at NUTS3 level by megaregion

Urban surface percentage –URGpor

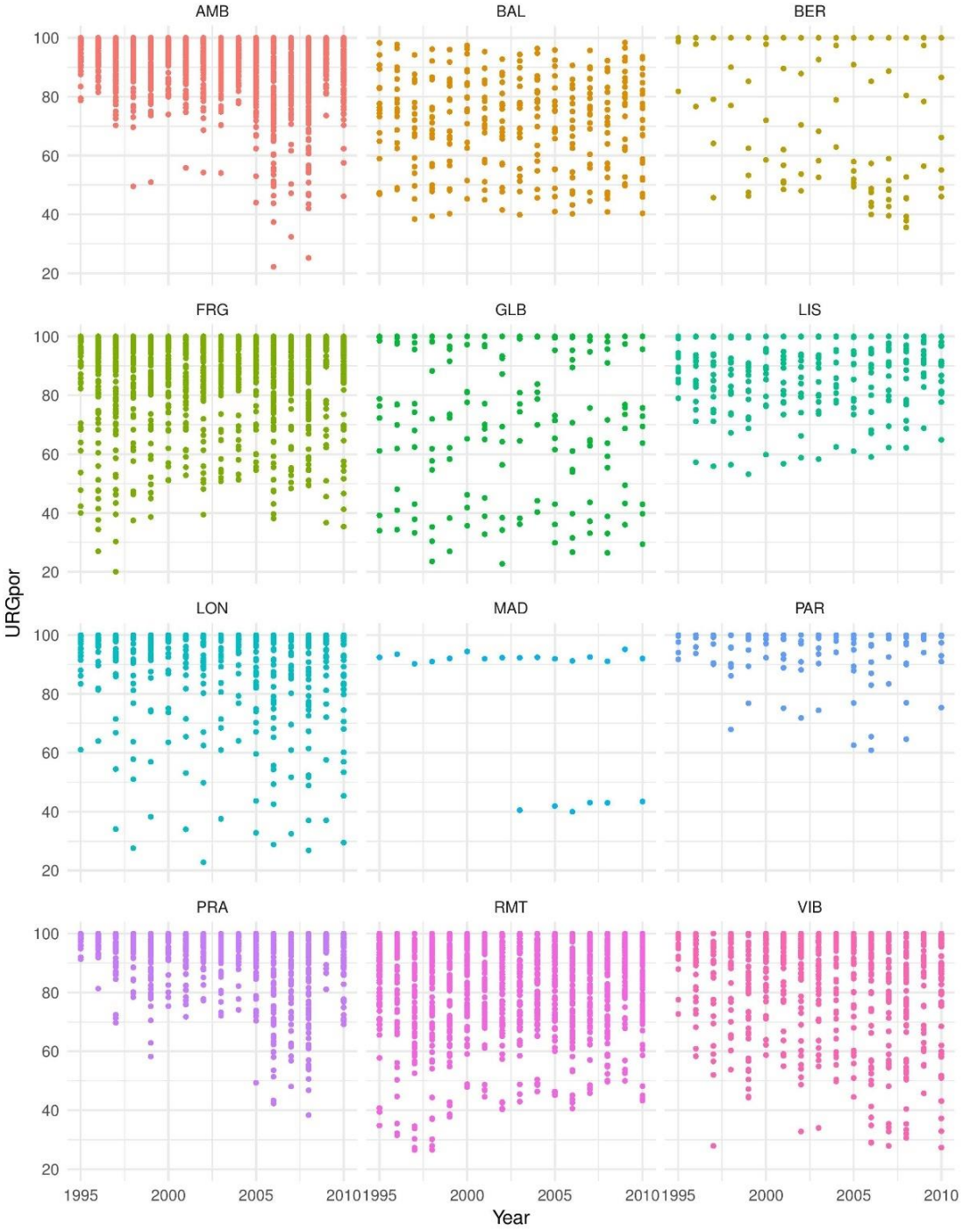

Figure 16 URGpor at NUTS3 level by megaregion

Primary energy consumption –PECpc

Figure 17 PECpc at NUTS3 level by megaregion

Missing data and imputation

Missing data arise in almost all serious statistical analyses and there are different methods to handle it, including some relatively simple approaches that can offer reasonable results. Multiple imputations (Rubin, 1987) is the method of choice for complex incomplete data.

Missing data that occur in more than one variable presents a special challenge. Two general approaches for imputing multivariate data have emerged: joint modeling and fully conditional specification, also known as multivariate imputation by chained equations (MICE).

Multiple imputation by chained equations

The Figure 18 illustrates the three main steps in MICE: imputation, analysis, and pooling. The software stores the results of each step in a specific class: *mids*, *mira* and *mipo*. We now explain each of these steps in more detail.

The analysis starts with an observed, incomplete data set $Yobs$. In general, the problem is that we cannot estimate Q from $Yobs$ without making unrealistic assumptions about the unobserved data. Multiple imputation is a general framework that several imputed versions of the data by replacing the missing values by plausible data values. These plausible values are drawn from a distribution specifically modeled for each missing entry.

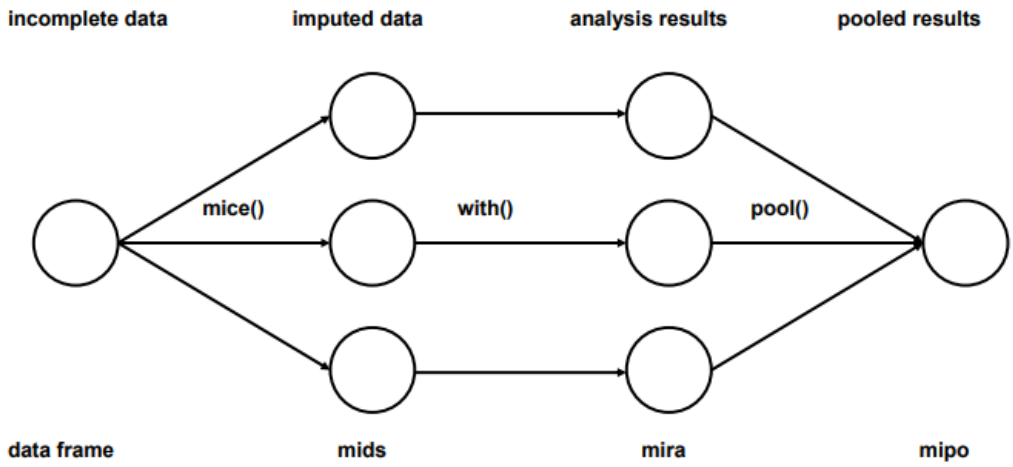

Figure 18 Multiple Chained Equations

In MICE, this task is being done by the function *mice()*. The first step portrays m = 3 imputed data sets $Y\ (1), \ldots, Y\ (3)$. The three imputed sets are identical for the non-missing data entries, but differ in the imputed values. The magnitude of these difference reflects our uncertainty about what value to impute. The package has a special class for storing the imputed data: a multiply imputed dataset of class *mids*.

The second step is to estimate Q on each imputed data set, typically by the method we would have used if the data had been complete. This is easy since all data are now complete. The model applied to $Y\ (1), \ldots, Y\ (m)$ is the generally identical. The estimates $\hat{Q}(1), \ldots, \hat{Q}(m)$ will differ from each other because their input data differ. These differences are caused because of our uncertainty about what value to impute. The analysis results are collectively stored as a multiply imputed repeated analysis within an R object of class *mira*.

The last step is to pool the m estimates $\hat{Q}(1),\ldots,\hat{Q}(m)$ into one estimate $\bar{Q}$ and estimate its variance. For quantities Q that are approximately normally distributed, we can calculate the mean over $\hat{Q}(1),\ldots,\hat{Q}(m)$ and sum the within- and between-imputation variance according to the method outlined in Rubin (1987).

Bayesian linear regression

Bayes' theorem is used to update the probability for a hypothesis as more evidence or information becomes available. When the regression model has errors that have a normal distribution, and if a particular form of prior distribution is assumed, explicit results are available for the posterior probability distributions of the model's parameters.

The Bayesian approach is based on the equal processing of the parameters of the model (θ) and the data analyzed (Y). Thus, the parameters θ are processed as usual random variables. In this way, the parameters are not unknown fixed values, but each parameter has its own probability distribution. The uncertainty is hereby introduced in the researched fixed value of the parameters. The Bayes rule is written in the special form as follows:

$$p(\theta|Y\ ) = \frac{p(Y|\theta)p(\theta)}{\int p(Y|\theta)p(\theta)d\theta} \propto\ p(Y|\theta)p(\theta)$$

The above-mentioned expression indicates that the posterior distribution of the parameters $p(\theta|Y\ )$ is proportional to the product of the sample distribution $p(Y|\theta)$ and the prior distribution of the parameters $p(\theta)$. This simple form of the rule encompasses the technical core of the Bayesian's inference. $Ymis$ denotes the missing data, and $Yobs$ that the data observed are noticed. The Bayesian approach for data imputation is based on a joint posterior distribution of the parameters θ and the missing data $Ymis$, which is conditional on the observed data and the model assumed: $p(Ymis,\theta)|Yobs,X)$.

The model represents both the explanatory variables X that are observed and the model of missing values, which is ignorable in our case. The model of missing values is ignorable, when the missing data is missing at random (MAR), and when the parameters of the missing data mechanism and the parameters of the probability model are distinct. The data Y may have $nmis$ missing values $Ymis$ with the accessory explanatory variables noticed $Xmis$. And the fully observed data are denoted by $Yobs$ in $Xobs$. The data may correspond to the model of linear regression:

$$Y\ =\ \beta_1 x_1\ +\ \beta_2 x_2\ +\ \ldots\ +\ \beta_k x_k\ =\ X\beta\ +\ \epsilon$$

$$\epsilon\ \sim\ N(0,\sigma^2)$$

By which the coefficient and variance estimation is calculated from the fully noticed data:

$$\hat{\beta}\ =\ \left(X_{obs}^T X_{obs}^{-1}\right)X_{obs}^T Y_{obs}$$

$$s2\ =\frac{\left(Y_{obs}-X_{obs}\hat{\beta}\right)^T\left(Y_{obs}-X_{obs}\hat{\beta}\right)}{n_{obs}\ -k}$$

Correlation between variables

The analysis methods that are going to be used to analyze the variables and to construct a model are funded in the correlation between the variables. Correlation is any of a broad class of statistical relationships involving dependence, though in common usage it most often refers to the extent to which two variables have a linear relationship with each other.

Formally, random variables are dependent if they do not satisfy a mathematical property of probabilistic independence. In an informal parlance, correlation is synonymous with dependence. Several correlation coefficients measure the degree of correlation. The most common of these is the Pearson correlation coefficient, which is sensitive only to a linear relationship between two variables (which may be present even when one variable is a nonlinear function of the other). The Pearson product-moment correlation coefficient is calculated as follows:

$$\rho_{x,y} = corr(X,Y) = \frac{cov(X,Y)}{\sigma_X \sigma_y} = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_y}$$

Where E is the expected value operator, cov means covariance, and corr is a widely used alternative notation for the correlation coefficient t. In our case, the data has missing values. This means that for some observations, some variables are present and some others not. The first option would be to impute these values, but this is a technique that creates new values to resemble the original data, but that there are not real.

To avoid the use of imputations, the correlation matrix is going to be calculated using a method called pairwise complete. This method search variable by variable and calculates the correlation between them, even though in that specific observation it may have missing values.

Implementation

Once the information has been mapped in order to assess the amount of missing information and given the aforementioned methods the imputation is made over the data. Given the distribution of missing values, it can be seen that the information it is to a certain degree missing at random but for the previous years the number of cases increases. The imputation is made using the package MICE, which stands for Multiple Imputation using Chain Equations. The method of imputation chosen is the Bayesian Linear Regression.

The following figures show the distribution of missing data over the variables (Figure 19), and the distribution over countries and years (Figure 20). If the amount of missing data is small relative to the size of the dataset, then leaving out the few samples with missing features is the best strategy in order not to bias the analysis, however leaving out available data points deprives the data of some amount of information.

Usually, a safe maximum threshold is 5% of the total for large datasets. If missing data for a certain feature or sample is more than 5% then you probably should leave that feature or sample out. Given the fact that the goal of the first analysis is to determine the relationships between the variables, a way to obtain the correlations of incomplete data sets is to determine this in a pairwise fashion. This way, a function is created that accomplishes the next objectives:

- Measure the percentage of missing values of a given variable.
- Imputes no more than 5% of these missing values.
- Deletes no logical values (i.e. negative values).
- Determines and saves only correlation matrix (because of memory allocation issues).
- Repeat the process a specified amount of times and determines the average of all the simulations.

The mean relative difference of imputed and not imputed correlation matrices is 0.09.

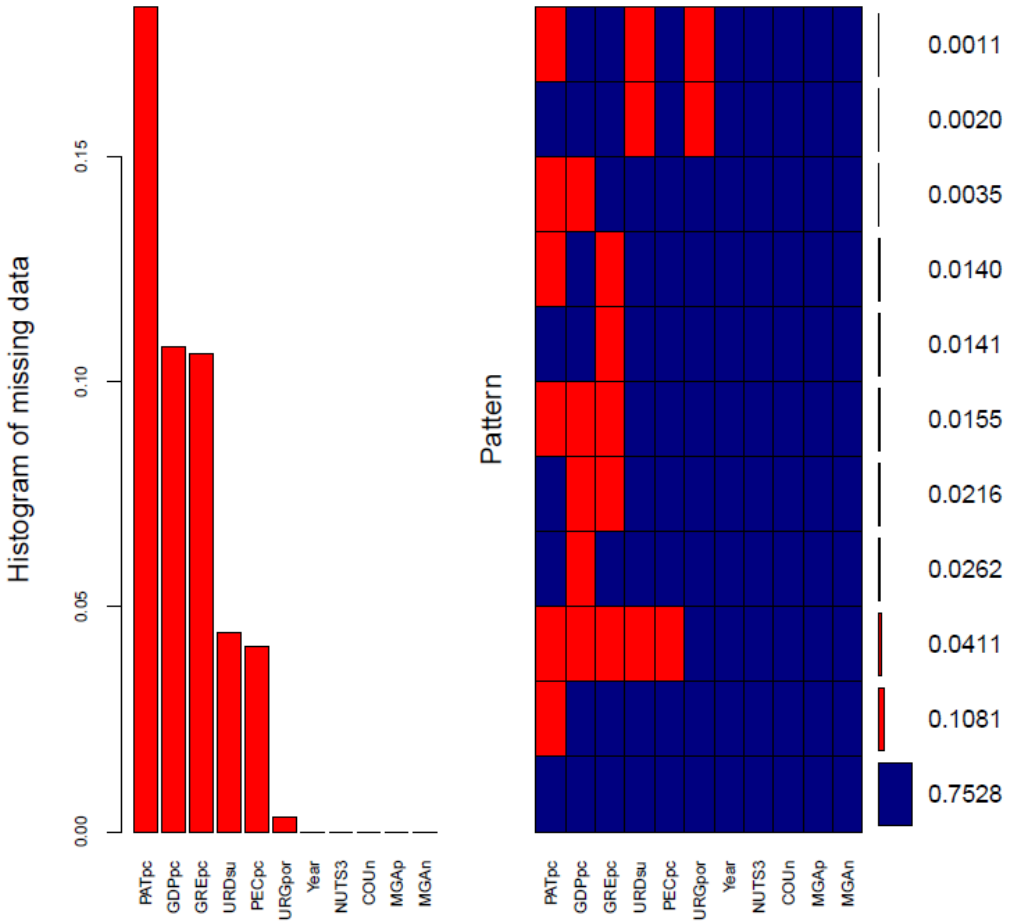

Figure 19 Distribution of missing values

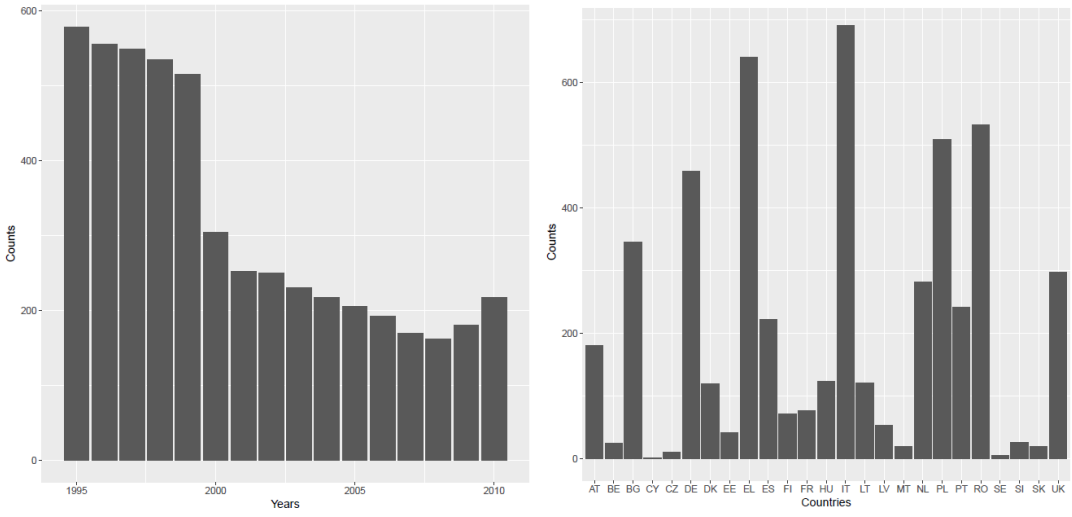

Figure 20 Missing values per country and year

Given the small difference between the imputed and original data in the correlation matrix, it is concluded that the imputations are going to be used only for visualization purposes, which means that they are not going to be used in the multivariate analysis. This is because the imputation itself means to introduce a bias in the correlations that could misrepresent the information.

The imputations are constructed using the observable relationships, but given the small difference between the imputed (Table 3) and non-imputed values (Table 4), and the fact that the goal of the further statistical analysis is to determine the hidden correlations between variables, the imputation may introduce correlations that are simply not there. Because the data is not missing at random (it is greater in the first years and for certain countries) and would induce instability in the statistical analysis.

Variable	PATth	GDPpc	PECpc	GREpc	URDpsk	URGpor
PATth	1.00	0.36	-0.00	0.27	0.07	0.28
GDPpc	0.36	1.00	-0.14	0.75	0.42	0.32
PECpc	-0.00	-0.14	1.00	-0.23	-0.46	-0.18
GREpc	0.27	0.75	-0.23	1.00	0.36	0.22
URDpsk	0.07	0.42	-0.46	0.36	1.00	0.29
URGpor	0.28	0.32	-0.18	0.22	0.29	1.00

Table 3 Correlation matrix: imputed data

Variable	PATth	GDPpc	PECpc	GREpc	URDpsk	URGpor
PATth	1.00	0.38	-0.05	0.28	0.08	0.30
GDPpc	0.38	1.00	-0.22	0.76	0.38	0.32
PECpc	-0.05	-0.22	1.00	-0.28	-0.47	-0.25
GREpc	0.28	0.76	-0.28	1.00	0.33	0.25
URDpsk	0.08	0.38	-0.47	0.33	1.00	0.29
URGpor	0.30	0.32	-0.25	0.25	0.29	1.00

Table 4 Correlation matrix: no imputed data

Statistical analysis

The economic, social and ecological factors of urban sustainable progress cannot be measured directly, but through the variables that are affected by them, which are the observable variables. Thus, several statistical analysis will be used to detect and measure these hidden factors: component analysis, factor analysis, structural equation model, and cluster analysis.

Component analysis

Briefly, Principal Component Analysis (PCA) is used to extract the important information from a multivariate data table and to express this information as a set of new variables called principal components. The information in a given data set corresponds to the total variation it contains. The goal of PCA is to identify directions along which the variation in the data is maximal. These directions (called also principal components) can be used to visualize graphically the data.

PCA is a statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components (or sometimes, principal modes of variation). The number of principal components is less than or equal to the smaller of the number of original variables or the number of observations.

This transformation is defined in such a way that the first principal component has the largest possible variance (that is, accounts for as much of the variability in the data as possible), and each succeeding component in turn has the highest variance possible under the constraint that it is orthogonal to the preceding components. The resulting vectors are an uncorrelated orthogonal basis set. PCA is sensitive to the relative scaling of the original variables.

Suppose a random vector X:

$$X = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_P \end{pmatrix}$$

With population variance-covariance matrix:

$$var(X) = \Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \dots & \sigma_{1p} \\ \vdots & \sigma_{22} & \dots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp} \end{pmatrix}$$

Consider the linear combinations:

$$Y_1 = e_{11}X_1 + e_{12}X_2 + \cdots + e_{1p}X_p$$

$$Y_2 = e_{21}X_1 + e_{22}X_2 + \cdots + e_{2p}X_p$$

$$\vdots$$

$$Y_p = e_{p1}X_1 + e_{p2}X_2 + \cdots + e_{pp}X_p$$

Each of these can be thought of as a linear regression, predicting $\mathrm{Y_i}$ from $\mathrm{X_1}, \mathrm{X_2}, \ldots, \mathrm{X_p}$. There is no intercept, but $\mathrm{e_{i1}}, \mathrm{e_{i2}}, \ldots, \mathrm{e_{ip}}$ can be viewed as regression coefficients.

Note that Y_i is a function of our random data, and so is also random. Therefore, it has a population variance:

$$var(Y_i) = \sum_{k=1}^{p}\sum_{l=1}^{q} e_{ik}e_{il}\sigma_{kl} = e_i'\Sigma e_i$$

Moreover, Y_i and Y_j will have a population covariance:

$$cov(Yi, Y_j) = \sum_{k=1}^{p}\sum_{l=1}^{q} \mathrm{e_{ik}e_{jl}\sigma_{kl}} = \mathrm{e_i'\Sigma e_j}$$

Here the coefficients e_{ij} are collected into the vector:

$$e_i = \begin{pmatrix} e_{i1} \\ e_{i2} \\ \vdots \\ e_{iP} \end{pmatrix}$$

ith principal component Y_i :

We select $\mathrm{e_{i1}, e_{i2}, \dots, e_{ip}}$ that maximizes:

$$var(Y_i) = \sum_{k=1}^{p}\sum_{l=1}^{q} \mathrm{e_{ik}e_{il}\sigma_{kl}} = e_i'\Sigma e_i$$

Subject to the constraint that the sums of squared coefficients add up to one along with additional constraint this new component will be uncorrelated with all the previously defined components.

$$e'_i e_i = \sum_{j=1}^{p} e_{ij}^2 = 1$$

$$cov(Y_1, Y_i) = \sum_{\mathrm{k=1}}^{\mathrm{p}}\sum_{\mathrm{l=1}}^{\mathrm{q}} \mathrm{e_{ik}e_{jl}\sigma_{kl}} = e_1'\Sigma e_i = 0,$$

$$cov(Y_2, Y_i) = \sum_{\mathrm{k=1}}^{\mathrm{p}}\sum_{\mathrm{l=1}}^{\mathrm{q}} \mathrm{e_{ik}e_{jl}\sigma_{kl}} = e_2'\Sigma e_i = 0,$$

$$\vdots$$

$$cov(Y_{i-1}, Y_i) = \sum_{k=1}^{p}\sum_{l=1}^{q} \mathrm{e_{ik}e_{jl}\sigma_{kl}} = e'_{i-1}\Sigma e_i = 0,$$

Therefore, all principal components are uncorrelated with one another.

PCA is mostly used as a tool in exploratory data analysis and for making predictive models. It is often used to visualize genetic distance and relatedness between populations. PCA can be done by eigenvalue decomposition of a data covariance (or correlation) matrix or singular value decomposition of a data matrix, usually after mean centering (and normalizing or using Z-scores) the data matrix for each attribute. The results of a PCA are usually discussed in terms of

component scores, sometimes called factor scores (the transformed variable values corresponding to a particular data point), and loadings (the weight by which each standardized original variable should be multiplied to get the component score).

PCA is the simplest of the true eigenvector-based multivariate analyses. Often, its operation can be thought of as revealing the internal structure of the data in a way that best explains the variance in the data. If a multivariate dataset is visualized as a set of coordinates in a high-dimensional data space (1 axis per variable), PCA can supply the user with a lower-dimensional picture, a projection of this object when viewed from its most informative viewpoint. This is done by using only the first few principal components so that the dimensionality of the transformed data is reduced.

PCA is closely related to factor analysis. Factor analysis typically incorporates more domain specific assumptions about the underlying structure and solves eigenvectors of a slightly different matrix.

Factor analysis

The factor analysis has the goal of determining the concepts of primary interest, called latent variables, that are not possible to be measured directly but can be observed in the manifest variables. The method of analysis most generally used to help uncover the relationships between the assumed latent variables and the manifest variables is factor analysis.

The model on which the method is based is essentially that of multiple regression, except now the manifest variables are regressed on the unobservable latent variables (often referred to in this context as common factors), so that direct estimation of the corresponding regression coefficients (factor loadings) is not possible.

Factor analysis comes in two distinct varieties. The first is exploratory factor analysis, which is used to investigate the relationship between manifest variables and factors without making any assumptions about which manifest variables are related to which factors. The second is confirmatory factor analysis, which is used to test whether a specific factor model postulated a priori provides an adequate fit for the covariances or correlations between the manifest variables.

Exploratory factor analysis

The basis of factor analysis is a regression model linking the manifest variables to a set of unobserved (and unobservable) latent variables. In essence, the model assumes that the observed relationships between the manifest variables (as measured by their covariances or correlations) are a result of the relationships of these variables to the latent variables. Since it is the covariances or correlations of the manifest variables that are central to factor analysis, we can assume that the manifest variables all have zero mean.

The model links the observed variables $xt = (x1, \ldots, xp)$ to a set of latent factors $ft = (f1, \ldots, xm)$, with m < p:

$$x_1 = q_{11}f_1 + q_{12}f_2 + \ldots + q_{1m}f_m + u_1$$

$$x_i = q_{i1}f_1 + q_{i2}f_2 + \ldots + q_{im}f_m + u_i$$

$$x_p = q_{p1}f_1 + q_{p2}f_2 + \ldots + q_{pm}f_m + u_p$$

Where the variables in random vector $ut = (u1, \ldots, up)$ are called specific or unique factors and the coefficients q_{ij} in the model are called factor loadings. In short:

$$x = Qf + u$$

The linear equation in factorial models a matrix form and applied to scaled data is:

$$Z = QF + U$$

Where Z are the scaled observed variables, F the latent factors, U the unique factors (or residuals) and Q are the factors loadings.

The correlation matrix ∑ can be calculated as:

$$\Sigma = QQ^t + \psi$$

Were psi being the specific factors.

The q_{ij}, which are the elements of Q, are essentially the regression coefficients of the observed variables on the common factors remembering that the common factors are hidden variables, so the meaning of regression is inappropriate. But in the context of factor analysis these regression coefficients are known as the factor loadings and show how each observed variable depends on the common factors. The factor loadings are used in the interpretation of the factors. In the described factorial model, with the given hypotheses, loadings are the correlation coefficients between observed variables and common factors.

We assume that the random disturbance (or specific factors) terms are uncorrelated with each other and with the factors. The elements are specific to each manifest variable and hence are generally better known in this context as specific variates. The two assumptions imply that, given the values of the common factors, the manifest variables are independent; that is, the correlations of the observed variables arise from their relationships with the common factors.

Factor analysis is essentially unaffected by the rescaling of the variables. In particular, if the rescaling factors are the ratio between its value and its standard deviation, then the rescaling is equivalent to applying the factor analysis model to the correlation matrix of the x variables and the factor loadings and specific variances that result can be found simply by scaling the corresponding loadings and variances obtained from the covariance matrix. Consequently, the factor analysis model can be applied to either the covariance matrix or the correlation matrix because the results are essentially equivalent.

Number of factors

Given the number of observable variables, the number of factors needs to be established. With too few factors there will be too many high loadings, and with too many factors, factors may be fragmented and difficult to interpret convincingly.

For our data set the Cumulative Variance criteria is applied. Choosing k might be done by examining solutions corresponding to different values of k and deciding subjectively which can be given the most convincing interpretation.

Factor rotation

There is no unique solution for the factor loading matrix.

We can see that this is so by introducing an orthogonal matrix *M* of order *k* ×*k* and rewriting the basic regression equation linking the observed and latent variables as:

$$Z = (QM)(M^T F) + U$$

This model implies that the covariance matrix of the observed variables is:

$$\Sigma = (MQ)(MQ)^t + \psi$$

This "new" model satisfies all the requirements of a k-factor model as previously outlined with new factors $F^* = MF$ and the new factor loadings QM. Rotations do not change the factorial characteristics: number of factors, communalities and specificities.

Structural equation model

An exploratory factor analysis is used in the early investigation of a set of multivariate data to determine whether the factor analysis model is useful in providing a parsimonious way of describing and accounting for the relationships between the observed variables. The analysis will determine which observed variables are most highly correlated with the common factors and how many common factors are needed to give an adequate description of the data. In an exploratory factor analysis, no constraints are placed on which manifest variables load on which factors.

The models will be constructed according to some theoretical criteria in order to obtain a model that has a valid interpretation. This may happen when the loadings of some variables on some factors are fixed at zero because they were "small" in the exploratory analysis and perhaps to allow some pairs of factors but not others to be correlated.

It is important to emphasize that whilst it is perfectly appropriate to arrive at a factor model to submit to a confirmatory analysis from an exploratory factor analysis, the model must be tested on a fresh set of data. Models must not be generated and tested on the same data. However, confirmatory analysis can be applied directly to data, without a previous exploring, if some common factors can be assumed to cause the observed variables by applying theoretical reasoning.

Confirmatory factor analysis models are a subset of a more general approach to modelling latent variables known as structural equation modelling or covariance structure modelling. Such models allow both response and explanatory latent variables linked by a series of linear equations. Although more complex than confirmatory factor analysis models, the aim of structural equation models is essentially the same, namely to explain the correlations or covariances of the observed variables in terms of the relationships of these variables to the assumed underlying latent variables and the relationships postulated between the latent variables themselves.

One way to look at SEM models is that they are simply an extension of linear regression. A first extension is that you can have several regression equations at the same time. A second extension is that a variable that is an independent (exogenous) variable in one equation can be a dependent (endogenous) variable in another equation. A third extension is that some of the variables are observable and others hidden.

Path analysis was introduced as a method for studying the direct and indirect effects of variables. The most important feature of path analysis is a diagram showing how a set of explanatory variables influence a dependent variable under consideration. How the paths are drawn determines whether the explanatory variables are correlated causes, mediated causes, or independent causes.

Model identification

If different sets of parameter values will lead to the same predicted covariance matrix, the model is said to be unidentifiable. Formally, a model is identified if and only if $\Sigma(\theta_1) = \Sigma(\theta_2)$ implies

that $\theta_1 = \theta_2$, which are the parameter values. Otherwise, it may happen that increased by some amount of a given parameter and other decreased by the same amount without altering the covariance matrix predicted by the model.

In confirmatory factor analysis models and more general covariance structure models, identifiability depends on the choice of model and on the specification of fixed, constrained (for example, two parameters constrained to equal one another), and free parameters. If a parameter is not identified, it is not possible to find a consistent estimate of it. Establishing model identification in confirmatory factor analysis models (and in structural equation models) can be difficult because there are no simple, practicable, and universally applicable rules for evaluating whether a model is identified, although there is a simple necessary but not sufficient condition for identification, namely that the number of free parameters in a model, t, be less than $q(q + 1)/2$.

In our case, the model converges within a certain number of iterations, and if it's not possible to compute the standard errors, it may be possible that the model has not been correctly identified.

Model fit

Structural equation models will contain a number of parameters that need to be estimated from the covariance or correlation matrix of the manifest variables. Estimation involves finding values for the model parameters that minimize a discrepancy function indicating the magnitude of the differences between the elements of S, the observed covariance matrix of the manifest variables and those of $\Sigma(\theta)$, the covariance matrix implied by the fitted model (i.e., a matrix the elements of which are functions of the parameters of the model), contained in the vector $\theta = (\theta_1, \ldots, \theta_t)$.

The function that is going to minimize the discrepancy between S and Σ in this group is the Unweighted Least Squares (ULS) discrepancy function:

$$ULS = \frac{1}{2}[tr(S - \Sigma)^2]$$

This discrepancy function is analogous to ordinary least squares estimation in regression. This function differs from the others in that it is not built on an assumption of multivariate normality in the data. As a result, this discrepancy function does not, in itself, lead to estimated standard errors or an overall chi-square fit statistic.

Confirmatory factor analysis models

In a confirmatory factor model the loadings for some observed variables on some of the postulated common factors will be set a priori to zero. Additionally, some correlations between factors might also be fixed at zero. Such a model is fitted to a set of data by estimating its free parameters; i.e., those not fixed at zero.

A series of models is going to be analyzed in the time frame. The most fitted models are going to be applied to time frames of 5 years. The models are constructed by a function that relates the latent variables with the observed variables by filtering the loadings of the factor analysis. This is made by establishing a threshold limit to these loadings. If loadings surpass this limit they are included, otherwise, specific observed variable is not included in that specific latent variable.

The strategy is to first create a measurement model, filtering the factors in observed variables in each latent variable according to a threshold that is going to be progressively increased until

the model converges. If there are any errors once the model has converged, this is going to be solved by introducing or modifying residual covariancies between the observed variables.

The parameters of analysis are going to be:

- Unweighted Least Squares (ULS) as parameter estimator.
- Fixed variances of all the latent variables in a CFA model to unity.
- Complete data only

The analysis is going to be performed using the package Lavaan (Table 5), a package for structural equation modeling implemented in the R system for statistical computing. Lavaan is an acronym for latent variable analysis, and its name reveals the long-term goal: to provide a collection of tools that can be used to explore, estimate, and understand a wide family of latent variable models, including factor analysis, structural equation, longitudinal, multilevel, latent class, item response, and missing data models.

Formula type	Operator	Mnemonic
Latent variable	=~	is manifested by
Regression	~	is regressed on
(Residual) (co)variance	~~	is correlated with
Intercept	~1	intercept

Table 5 Formula types that can be used to specify a model in the lavaan model syntax.

Cluster analysis

One of the most basic abilities of living creatures involves the grouping of similar objects to produce a classification. The idea of sorting similar things into categories is clearly a primitive one because early humans, for example, must have been able to realize that many individual objects shared certain properties such as being edible, or poisonous, or ferocious, and so on.

Classification of the phenomena being studied is an important component of virtually all scientific research. In the behavioral sciences, these “phenomena” may be individuals or societies, or even patterns of behavior or perception. The investigator is usually interested in finding a classification in which the items of interest are sorted into a small number of *homogeneous groups* or *clusters*, the terms being synonymous. Most commonly the required classification is one in which the groups are mutually exclusive (an item belongs to a single group) rather than overlapping (items can be members of more than one group).

But often a classification may seek to serve a more fundamental purpose. In psychiatry, for example, the classification of psychiatric patients with different symptom profiles into clusters might help in the search for the causes of mental illnesses and perhaps even lead to improved therapeutic methods. And these twin aims of prediction (separating diseases that require different treatments) and a etiology (searching for the causes of disease) for classifications will be the same in other branches of medicine.

Cluster analysis is a generic term for a wide range of numerical methods with the common goal of uncovering or discovering groups or clusters of observations that are homogeneous and separated from other groups. Clustering techniques essentially try to formalize what human observers do so well in two or three dimensions. Clusters are identified by the assessment of the relative distances between points.

The cluster analysis, in the context of the whole analysis itself, has the goal of discover the groups of the observed regions that share characteristics in terms of the hidden relationships

that govern the different axis of sustainable development, and it may give a helpful insight identifying trends or to understand its behaviors over time.

Hierarchical clustering

This class of clustering methods produces a hierarchical classification of data. In a hierarchical classification, the data are not partitioned into a particular number of classes or groups at a single step. Instead, the classification consists of a series of partitions that may run from a single "cluster" containing all individuals to n clusters, each containing a single individual. Agglomerative hierarchical clustering techniques produce partitions by a series of successive fusions of the n individuals into groups. With such methods, fusions, once made, are irreversible, so that when an agglomerative algorithm has placed two individuals in the same group they cannot subsequently appear in different groups. Since all agglomerative hierarchical techniques ultimately reduce the data to a single cluster containing all the individuals, the investigator seeking the solution with the best-fitting number of clusters will need to decide which division to choose. The problem of deciding on the "correct" number of clusters will be taken up later.

An agglomerative hierarchical clustering procedure produces a series of partitions of the data, $P_n, P_{n-1}, \ldots, P_1$. The first, *Pn*, consists of *n* single member clusters, and the last, *P1*, consists of a single group containing all *n* individuals. The basic operation of all methods is similar:

- Clusters *C1, C2,...,Cn* each containing a single individual
- Find the nearest pair of distinct clusters, say C_i and C_j, merge C_i and C_j, delete C_j, and decrease the number of clusters by one.
- If the number of clusters equals one, then stop; otherwise return to 1.

But before the process can begin, an inter-individual distance matrix or similarity matrix needs to be calculated. There are many ways to calculate distances or similarities between pairs of individuals, but here we only deal with a commonly used distance measure, Euclidean distance is calculated as:

$$d_{ij} = \sqrt{\sum_{k=1}^{q} \left(x_{ik} - x_{jk}\right)^2},$$

Where d_{ij} is the Euclidean distance between individual i with variable values $x_{i1}, x_{i2} \ldots x_{iq}$ and individual j with variable values $x_{j1}, x_{j2} \ldots x_{jq}$. The Euclidean distances between each pair of individuals can be arranged in a matrix that is symmetric because $d_{ij} = d_{ji}$ and has zeros on the main diagonal. Such a matrix is the starting point of many clustering examples, although the calculation of Euclidean distances from the raw data may not be sensible when the variables are on very different scales. In such cases, the variables can be standardized in the usual way before calculating the distance matrix, although this can be unsatisfactory in some cases (see Everitt et al. 2011).

Given an inter-individual distance matrix, the hierarchical clustering can begin, and at each stage in the process the methods fuse individuals or groups of individuals formed earlier that are closest (or most similar). So as groups are formed, the distance between an individual and a group containing several individuals and the distance between two groups of individuals will need to be calculated. How such distances are defined leads to a variety of different techniques. Two simple inter-group measures are

$$d_{AB} = \min_{i \in A, i \in B}(d_{ij})$$

$$d_{AB} = \max_{i \in A, i \in B}(d_{ij})$$

Where d_{AB} is the distance between two clusters A and B, and $d_{ij}$is the distance between individuals i and j found from the initial inter-individual distance matrix. The first inter-group distance measure above is the basis of single linkage clustering, the second that of complete linkage clustering. Both these techniques have the desirable property that they are invariant under monotone transformations of the original inter-individual distances; i.e., they only depend on the ranking on these distances, not their actual values. A further possibility for measuring inter-cluster distance or dissimilarity is

$$d_{AB} = \frac{1}{n_A n_B} \sum_{i \in A} \sum_{i \in B} d_{ij}$$

where n_A and n_B are the numbers of individuals in clusters A and B. This measure is the basis of a commonly used procedure known as group average clustering. Hierarchical classifications may be represented by a two-dimensional diagram known as a dendrogram, which illustrates the fusions made at each stage of the analysis.

In our case, the Ward's method criterion is applied in the hierarchical cluster analysis. Ward's minimum variance method is a special case of the objective function approach. Ward suggested a general agglomerative hierarchical clustering procedure, where the criterion for choosing the pair of clusters to merge at each step is based on the optimal value of an objective function.

K-means clustering

K-means clustering is a type of unsupervised learning, which is used when there is unlabeled data (i.e., data without defined categories or groups). The goal of this algorithm is to find groups in the data, with the number of groups represented by the variable K. The algorithm works iteratively to assign each data point to one of K groups based on the features that are provided. Data points are clustered based on feature similarity. The results of the K-means clustering algorithm are:

- The centroids of the K clusters, which can be used to label new data
- Labels for the training data (each data point is assigned to a single cluster)

Rather than defining groups before looking at the data, clustering allows to find and analyze the groups that have formed organically.

Each centroid of a cluster is a collection of feature values which define the resulting groups. Examining the centroid feature weights can be used to qualitatively interpret what kind of group each cluster represents.

Algorithm

The K-means clustering algorithm uses iterative refinement to produce a final result. The algorithm inputs are the number of clusters K and the data set. The data set is a collection of features for each data point. The algorithms start with initial estimates for the K centroids, which can either be randomly generated or randomly selected from the data set. The algorithm then iterates between two steps.

Data assignment step

Each centroid defines one of the clusters. In this step, each data point is assigned to its nearest centroid, based on the squared Euclidean distance. More formally, if ci is the collection of centroids in set C, then each data point x is assigned to a cluster based on

$$\arg min_{C_i \, \epsilon \, C} \, dist(C_i, x)^2$$

Where *dist* is the standard (L2) Euclidean distance. Let the set of data point assignments for each i^{th} cluster centroid be S_i.

Centroid update step

In this step, the centroids are recomputed. This is done by taking the mean of all data points assigned to that centroid's cluster.

$$C_i = \frac{1}{|S_i|} \sum_{x_i \epsilon S_i} x_i$$

The algorithm iterates between steps one and two until some stopping criteria is met (i.e., no data points change clusters, the sum of the distances is minimized, or some maximum number of iterations is reached).

This algorithm is guaranteed to converge to a result. The result may be a local optimum, which may be not necessarily the best possible outcome, meaning that assessing more than one run of the algorithm with randomized starting centroids may give a better outcome.

Choosing K

The algorithm described above finds the clusters and data set labels for a particular pre-chosen *K*. To find the number of clusters in the data, the user needs to run the K-means clustering algorithm for a range of *K* values and compare the results. In general, there is no method for determining exact value of *K*, but an accurate estimate can be obtained using some techniques.

One of the metrics that is commonly used to compare results across different values of *K* is the mean distance between data points and their cluster centroid. Since increasing the number of clusters will always reduce the distance to data points, increasing *K* will always decrease this metric, to the extreme of reaching zero when *K* is the same as the number of data points. Thus, this metric cannot be used as the sole target. Instead, mean distance to the centroid as a function of *K* is plotted and the "elbow point," where the rate of decrease sharply shifts, can be used to roughly determine *K*.

Statistical results

Factor analysis

For the factor analysis, three factors are used (Table 6), a Varimax rotation is applied to the loadings and the scores are calculated, once that the model has been identified, using the Thompson or regression method. To be able to use the entire set of data, instead of supplying the data, the correlation matrix is going to be loaded using a pairwise complete observation. In addition, an oblique rotation is going to be applied to test the stability of the model.

Variable	Factor1	Factor2	Factor3
PATth	0,414	-0,078	0,164
GDPpc	0,763	0,150	0,430
PECpc	0,102	-0,497	-0,160
GREpc	0,376	0,207	0,900
URDpsk	0,361	0,930	-0,019
URGpor	0,513	0,109	0,038

Table 6 Factor loadings

Statistics	Factor1	Factor2	Factor3
SS loadings	1.299	1.195	1.050
Proportion Var.	0.217	0.199	0.175
Cumulative Var.	0.217	0.416	0.591

Table 7 Explained variance

The three factors explain the 59,1% of the variance in our model (Table 7). This is enough to study the mainstream behavior of the relationships and complies with the necessity of having numerical results that can be interpreted in socio-economic-ecological terms.

The three-factor model resembles a statistical structure (Figure 21) that could be interpreted as *Economic Growth* (Factor 1), *Urban Ecology* (Factor 2), and *Social Cohesion* (Factor 3), judging by the loadings matrix (Table 6).

- In Factor 1, Gross Domestic Product (GDPpc), Patents (PATth), and the proportion of Urbanizated Area (URGpor) would explain the *Economic Growth*.
- In Factor 2, the negative correlation of Primary Energy Consuption (PECpc) and Urban Density (URDpsk) reveals that when urban networks are more concentrated and connected (i.e. policentric structure) could make more efficient use of energy, which would explain the *Urban Ecology*.
- And Factor 3 accounts for the correlation between Gross Rate Employment (GREpc) and GDPpc, which would explain the *Social Cohesion*.
- In the case of Factor 3 (used as proxy of social equality) is interesting to observe a low but negative correlation with PECpc (resources consumption), and positive with PATth (knowledge economy). This could be the future of an inclusive growth.

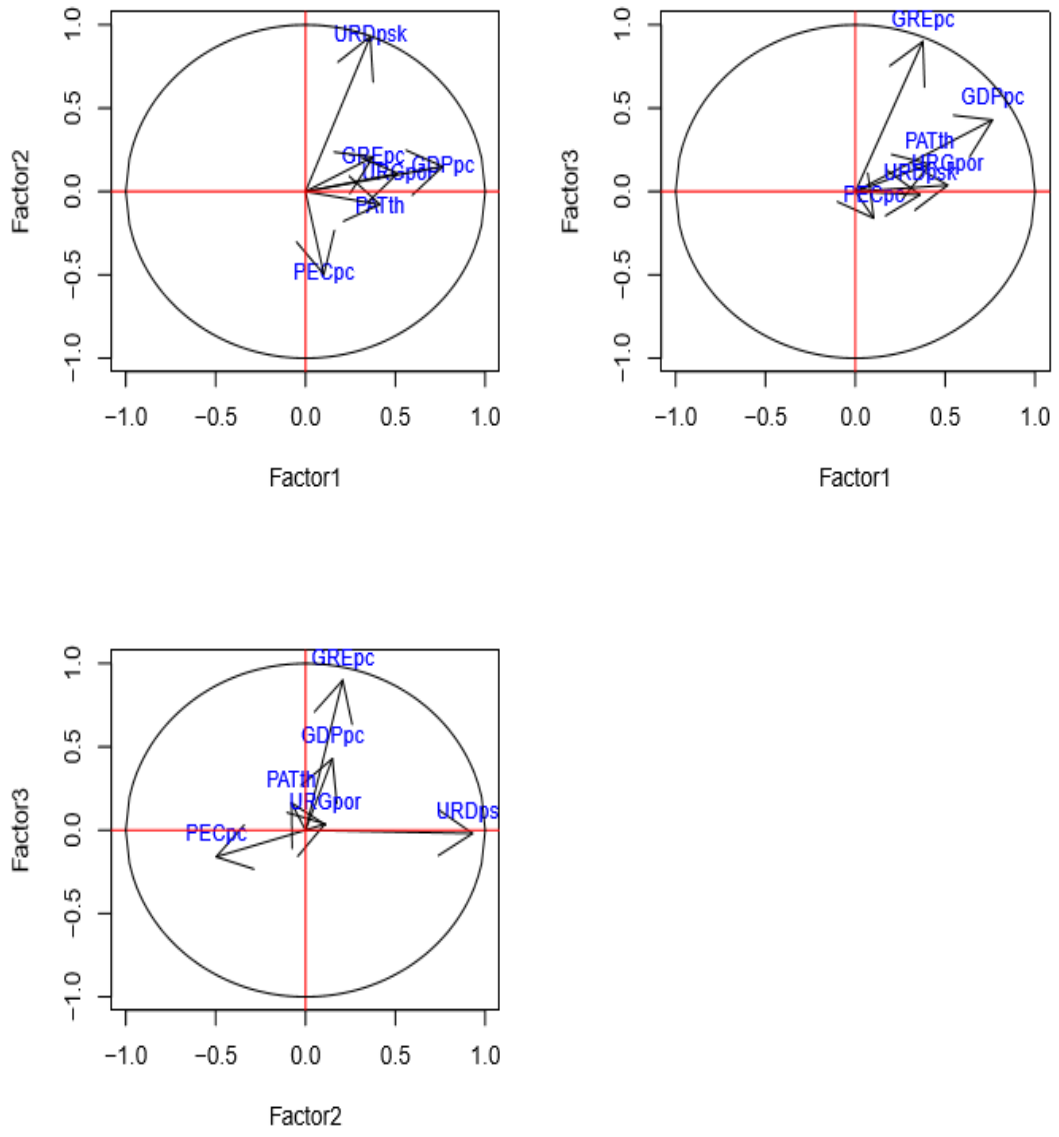

Figure 21 Exploratory Factor Analysis

Models evaluation

First, the model I is created using the function that analyze the loadings matrix (Table 8). Then, according to the results of the fit function a new model II is made (Table 9), taking into account the negative variances or other messages that are later included as residual covariances between the observed variables.

Each row in the tables of the models syntax corresponds to a single parameter. The 'rhs', 'op' and 'lhs' columns uniquely define the parameters of the model. All parameters with the '=~' operator are factor loadings, whereas all parameters with the '~~' operator are variances or covariances (Table 5). By default, the method includes the estimates, standard errors, *z* value, *p* value, and 95% confidence intervals for all the model parameters.

Model I complete observations

lhs	op	rhs	est	se	z	pvalue	ci.lower	ci.upper
Factor1	=~	PATth	0,505	NA	NA	NA	NA	NA
Factor1	=~	GDPpc	0,376	NA	NA	NA	NA	NA
Factor1	=~	URGpor	0,605	NA	NA	NA	NA	NA
Factor1	=~	URDpsk	0,202	NA	NA	NA	NA	NA
Factor2	=~	PECpc	0,905	NA	NA	NA	NA	NA
Factor2	=~	URDpsk	-0,453	NA	NA	NA	NA	NA
Factor3	=~	GREpc	0,829	NA	NA	NA	NA	NA
Factor3	=~	GDPpc	0,425	NA	NA	NA	NA	NA
GDPpc	~~	GREpc	0,228	NA	NA	NA	NA	NA
PECpc	~~	GREpc	-0,056	NA	NA	NA	NA	NA
URDpsk	~~	GREpc	0,126	NA	NA	NA	NA	NA
GDPpc	~~	URDpsk	0,147	NA	NA	NA	NA	NA
PATth	~~	PATth	0,745	NA	NA	NA	NA	NA
GDPpc	~~	GDPpc	0,499	NA	NA	NA	NA	NA
URGpor	~~	URGpor	0,634	NA	NA	NA	NA	NA
URDpsk	~~	URDpsk	0,696	NA	NA	NA	NA	NA
PECpc	~~	PECpc	0,180	NA	NA	NA	NA	NA
GREpc	~~	GREpc	0,313	NA	NA	NA	NA	NA
Factor1	~~	Factor1	1	0	NA	NA	1	1
Factor2	~~	Factor2	1	0	NA	NA	1	1
Factor3	~~	Factor3	1	0	NA	NA	1	1
Factor1	~~	Factor2	-0,316	NA	NA	NA	NA	NA
Factor1	~~	Factor3	0,564	NA	NA	NA	NA	NA
Factor2	~~	Factor3	-0,298	NA	NA	NA	NA	NA

Table 8 Model I complete observations

Model II complete observations

lhs	op	rhs	est	se	z	pvalue	ci.lower	ci.upper
Factor1	=~	PATth	0,505	0,010	52,667	0,000	0,487	0,524
Factor1	=~	GDPpc	0,273	0,037	7,359	0,000	0,200	0,345
Factor1	=~	URGpor	0,604	0,011	52,667	0,000	0,582	0,626
Factor2	=~	PECpc	-0,585	0,009	-63,059	0,000	-0,603	-0,567
Factor2	=~	URDpsk	0,800	0,013	63,059	0,000	0,775	0,824
Factor3	=~	GDPpc	0,679	0,036	18,643	0,000	0,608	0,750
Factor3	=~	GREpc	0,925	0,027	34,792	0,000	0,873	0,978
PATth	~~	PATth	0,745	0,013	57,155	0,000	0,719	0,770
GDPpc	~~	GDPpc	0,278	0,026	10,667	0,000	0,227	0,329
URGpor	~~	URGpor	0,635	0,016	38,840	0,000	0,603	0,667
PECpc	~~	PECpc	0,658	0,014	47,304	0,000	0,631	0,685
URDpsk	~~	URDpsk	0,361	0,022	16,345	0,000	0,317	0,404
GREpc	~~	GREpc	0,144	0,050	2,871	0,004	0,046	0,242
Factor1	~~	Factor1	1,000	0,000	NA	NA	1,000	1,000
Factor2	~~	Factor2	1,000	0,000	NA	NA	1,000	1,000
Factor3	~~	Factor3	1,000	0,000	NA	NA	1,000	1,000
Factor1	~~	Factor2	0,451	0,014	32,980	0,000	0,424	0,478
Factor1	~~	Factor3	0,505	0,021	23,535	0,000	0,463	0,547
Factor2	~~	Factor3	0,472	0,013	36,519	0,000	0,446	0,497

Table 9 Model II complete observations

The Lavaan package allows performing a comparison between the models (Table 10) in terms of an ANOVA test that includes several measure standards. Given that we do not use Maximum Likelihood Methods the Akaike (AIC) and Bayesian Likelihood is not computed, but we can compare the models using the Chisq parameter estimator, which in our case is the ULS.

ANOVA	Df	AIC	BIC	Chisq	Chisq diff	Df diff	Pr(>Chisq)
fit2b.p	0	NA	NA	2263,468	NA	NA	NA
fit2b.c	0	NA	NA	533,3426	-730,13	0	1
fit3X.p	5	NA	NA	4396,903	3863,56	5	0
fit3X.c	5	NA	NA	579,2454	-817,66	0	1

Table 10 Model comparison

The model I was able to compute the standard errors, which can be a symptom that the model is not completely identified. On the other side, the model II was the one that had a better performance, is the one that used the complete observations that outperformed the rest in terms of ULS, thus, is going to be the model that will be used to calculate the scores.

In Figure 22, Confirmatory Factor Analysis *Pattern Matrix* is represented. One direction arrows represent loadings (regression coefficients expressing variables in terms of latent factors), and double arrows represent implied correlations between factors, with red colored lines corresponding to negative coefficients and lines width being proportional to the absolute value.

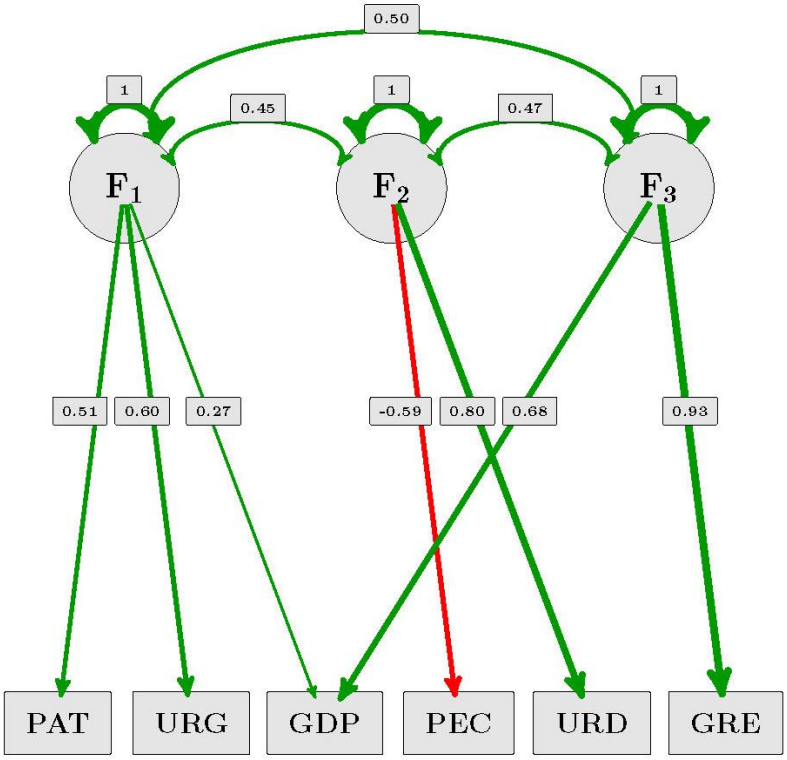

Figure 22 Confirmatory Factor Analysis *Pattern Matrix* of the complete observation model II

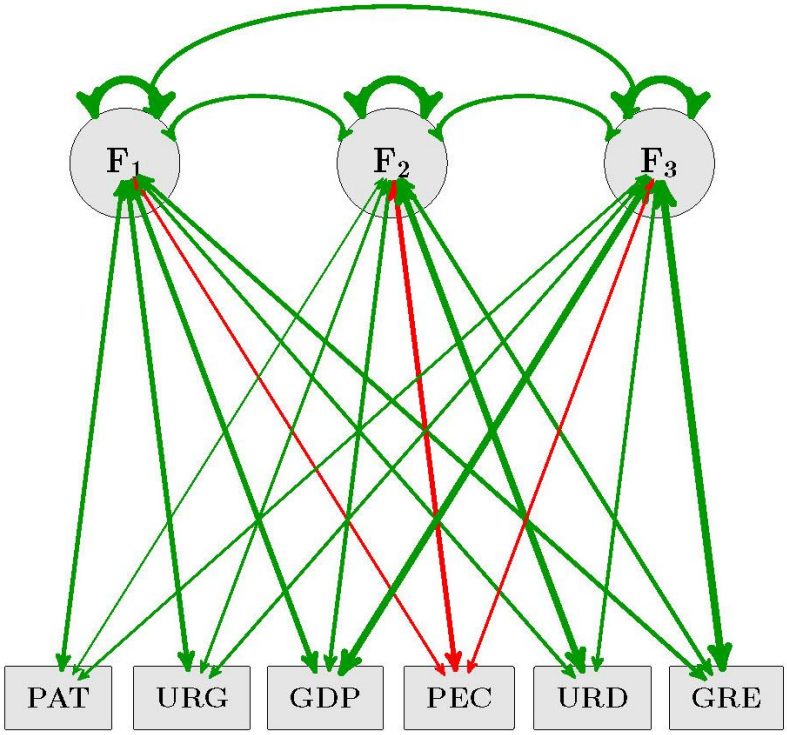

Figure 23 Confirmatory Factor Analysis *Structure Matrix* of the complete observation model II

In Figure 23, Confirmatory Factor Analysis *Structure Matrix* is represented. Double arrows represent correlations, with red coloured lines corresponding to negative coefficients and lines width being proportional to the absolute value. The Table 11 shows the numeric values in the structure matrix:

Variables	Factor1	Factor2	Factor3
PATth	0,50	0,23	0,25
GDPpc	0,62	0,44	0,82
PECpc	-0,26	-0,58	-0,28
GREpc	0,47	0,44	0,92
URDpsk	0,36	0,80	-0,38
URGpor	0,60	0,27	0,30

Table 11 Structure matrix of the selected model II

Confirmatory factor analysis

Given the obtained results, the model to be used is the II, which performs better. The Table 12 shows the correlation matrix of the factor model:

Variables	PATth	GDPpc	PECpc	GREpc	URDpsk	URGpor
PATth	1,00	0,31	-0,13	0,24	0,18	0,31
GDPpc	0,31	1,00	-0,26	0,76	0,35	0,37
PECpc	-0,13	-0,26	0,98	-0,26	-0,47	-0,16
GREpc	0,24	0,76	-0,26	1,51	0,35	0,28
URDpsk	0,18	0,35	-0,47	0,35	1,00	0,22
URGpor	0,31	0,37	-0,16	0,28	0,22	0,51

Table 12 Inferred correlation matrix of the factor model.

The model is defined by the following equations for the variables:

$$zPATpc = q_{11}F_1 + 0F_2 + 0F_3 + U_1$$

$$zGDPpc = q_{21}F_1 + 0F_2 + q_{23}F_3 + U_2$$

$$zPECpc = 0F_1 + q_{32}F_2 + 0F_3 + U_3$$

$$zGREpc = 0F_1 + 0F_2 + q_{43}F_3 + U_4$$

$$zURDpsk = q_{51}F_1 + q_{52}F_2 + 0F_3 + U_5$$

$$zURGpor = q_{61}F_1 + 0F_2 + 0F_3 + U_6$$

The notation zGDPpc, and so on, indicates that the observed variables are normalized in this analysis. The coefficients q_{ij} are the loadings, factors *F1*, *F2* and *F3* are the latent common factors and $U1, \ldots, U3$ are the unique or specific factors. We assume that there exist covariances between each pair of latent factors, being specific factors uncorrelated. Common and specific factors are uncorrelated, and we impose common factors to be standard. This model is simpler than the previous one, and the model assumptions imply the following decomposition:

$$R = Q\Sigma_F Q^t + \Psi$$

Where $R = (r_{ij})$ is the symmetric and positive definite correlation matrix between the observed variables (equivalent to the covariance of the normalized observations), Q is the matrix containing the loadings and Ψ the diagonal matrix containing the specific variances, also called residual variances:

$$Q = \begin{pmatrix} q_{11} & 0 & 0 \\ q_{21} & 0 & q_{23} \\ 0 & q_{32} & 0 \\ 0 & 0 & q_{43} \\ 0 & q_{52} & 0 \\ q_{61} & 0 & 0 \end{pmatrix}; \Sigma_F = \begin{pmatrix} 1 & \alpha & \beta \\ \alpha & 1 & \beta \\ \beta & \gamma & 1 \end{pmatrix}$$

$$\psi = \begin{pmatrix} \psi_{11} & 0 & 0 & 0 & 0 & 0 \\ 0 & \psi_{22} & 0 & 0 & 0 & 0 \\ 0 & 0 & \psi_{33} & 0 & 0 & 0 \\ 0 & 0 & 0 & \psi_{44} & 0 & 0 \\ 0 & 0 & 0 & 0 & \psi_{55} & 0 \\ 0 & 0 & 0 & 0 & 0 & \psi_{66} \end{pmatrix}$$

Scores calculation

Once the factor model has been identified, it is going to be used to calculate de scores of the observations. According to the factorial model, the Thompson type regression scores (Thomson, 1951) can be computed as:

$$F = ZR^{-1}Q\Sigma_F$$

Where Z is the data frame of normalized data, and the other matrices (observed variables correlation, loadings, and factors correlation matrices) are given above. The scores equation defining $B = R^{-1}Q\Sigma_F$, can be rewritten in the following form:

$$\begin{pmatrix} F_1 \\ F_2 \\ F_3 \end{pmatrix} = B^T \begin{pmatrix} Z_1 \\ Z_2 \\ Z_3 \\ Z_4 \\ Z_5 \\ Z_6 \end{pmatrix}$$

Where Z1 = zPATth, Z2 = zGDPpc, Z3 = zPECpc, Z4 = zGREpc, Z5 = zURDpsk and Z6 = zURGpor. This formula allows to compute the scores both for the sample observations and for new observations. Finally, taking into account that $Z = \frac{X-\bar{x}}{s}$ s and renaming $A^t = B^t S^{-1}$ where S^{-1} is the diagonal matrix whose diagonal elements are the reciprocal of the standard deviations 1/s of the variables, the scores can be written in terms of the original unscaled variables as follows:

$$\begin{pmatrix} F_1 \\ F_2 \\ F_3 \end{pmatrix} = A^T \begin{pmatrix} X_1 \\ X_2 \\ X_3 \\ X_4 \\ X_5 \\ X_6 \end{pmatrix} - A^T \begin{pmatrix} \overline{x_1} \\ \overline{x_2} \\ \overline{x_3} \\ \overline{x_4} \\ \overline{x_5} \\ \overline{x_6} \end{pmatrix}$$

The Table 13 shows the A^t matrix computed in this case:

Factors	X_1 PATth	X_2 GDPpc	X_3 PECpc	X_4 GREpc	X_5 URDpsk	X_6 URGpor
F_1 = Factor1	0,00129	4,00E-5	-0,01981	-0,00163	6,00E-05	0,01481
F_2 = Factor2	0,00075	0	-0,12097	0,00984	0,00056	-0,00206
F_3 = Factor3	-0,00034	3,00E-05	0,00402	0,06453	3,00E-05	0,00207

Table 13 A^t matrix

Finally, the factor scores in terms of the original unscaled variables can be calculated as:

$$F_1 = 0.00129X_1 + 4 \times 10^{-5}X_2 - 0.01981X_3 - 0.00163X_4 + 6 \times 10^{-5}X_5 + 0.01481X_6 - 2.17892$$

$$F_2 = 7.5 \times 10^{-4}X_1 + 0X_2 - 0.12097X_3 + 0.00984X_4 + 5.6 \times 10^{-4}X_5 - 0.00206X_6 - 0.2246$$

$$F_3 = -3.4 \times 10^{-4}X_1 + 3 \times 10^{-5}X_2 + 0.00402X_3 + 0.06453X_4 + 3 \times 10^{-5}X_5 + 0.00207X_6 - 3.73266$$

Table 14 and Table 15 show the averaged scores for countries and megaregions respectively.

Country	Factor1	Factor2	Factor3
AT	0,419	0,682	0,641
BE	0,939	0,36	-0,523
BG	-1,205	-0,086	-0,092
CY	-0,383	-0,724	0,441
CZ	-0,019	-0,417	0,559
DE	1,245	0,858	0,66
DK	0,816	0,5	1,041
EE	-1,183	-0,674	-0,077
EL	-0,772	-0,696	-1,028
ES	-0,169	-0,284	0,065
FI	-0,062	-1,066	-0,71
FR	0,603	1,295	0,134
HU	-0,572	0,051	-2,119
IE	0,3	-0,263	0,424
IT	0,087	-0,309	-0,656
LT	-1,309	-0,228	-0,347
LU	1,814	-1,149	2,745
LV	-0,933	0,804	-0,244
MT	0,456	1,161	-0,452
NL	1,109	0,157	-0,164
PL	0,021	0,71	0,088
PT	-0,423	-0,601	0,3
RO	-1,036	0,795	-1,33
SE	0,338	-0,65	0,547
SK	-0,544	-0,722	-0,376
UK	0,462	0,499	0,473

Table 14 Countries averaged scores

Megaregion	Factor1	Factor2	Factor3
NMR	-1,728	-0,697	-1,456
VIB	-0,589	-0,107	-0,715
FRG	1,326	0,143	1,356
AMB	0,594	-0,453	-0,26
PRA	0,006	-0,305	0,773
BER	0,413	1,23	0,004
LIS	-0,869	-0,616	-0,496
MAD	0,311	0,215	0,985
BAL	-0,353	-0,511	-0,292
PAR	1,57	2,364	1,094
RMT	-0,42	-0,546	-1,558
LON	0,045	-0,261	0,271
GLB	-0,306	-0,456	0,295

Table 15 Megaregions averaged scores

Factors in time

The plot shows us the behavior of the factors in time (1995, 2000, 2005, 2010) at megaregional level (Figure 24 to Figure 26). Some interesting patterns can be seen:

- In general, NUTS 3 not included in a megaregion (NMR) have worst patterns in *Economic Growth* (F1) – *Urban Ecology* (F2) relationship (Figure 24), comparing to NUTS 3 that belong to a megaregion.
- Three main NUTS 3 megaregional tendencies can be observed in the F1-F2 relationship (Figure 24): Frankfurt-Stuttgart (FRG) increases F1; London (LON) increases F2; and Paris (PAR) shows high scores of both factors.
- There is a positive lineal relationship between *Economic Growth* (F1) – *Social Cohesion* (F3) (Figure 25), with lowest values in no-megaregion NUTS 3 and highest values in advanced NUTS 3 megaregions (for example, FRG, AMB and PAR).
- More interesting is the relationship *Urban Ecology* (F2) – *Social Cohesion* (F3) (Figure 26). There is a tendency to a positive association between both factors, not observed in no-megaregion NUTS 3, but clearly observed in some advanced megaregions (like FRG, AMB, LON and PAR).
- The *Urban Ecology* (F2) – *Social Cohesion* (F3) relation is especially important because demonstrate that it is possible to create employment with lower energy consumption in European polycentric urban networks (Figure 26).

Plots of megaregions scores

Factor 1 vs Factor 2

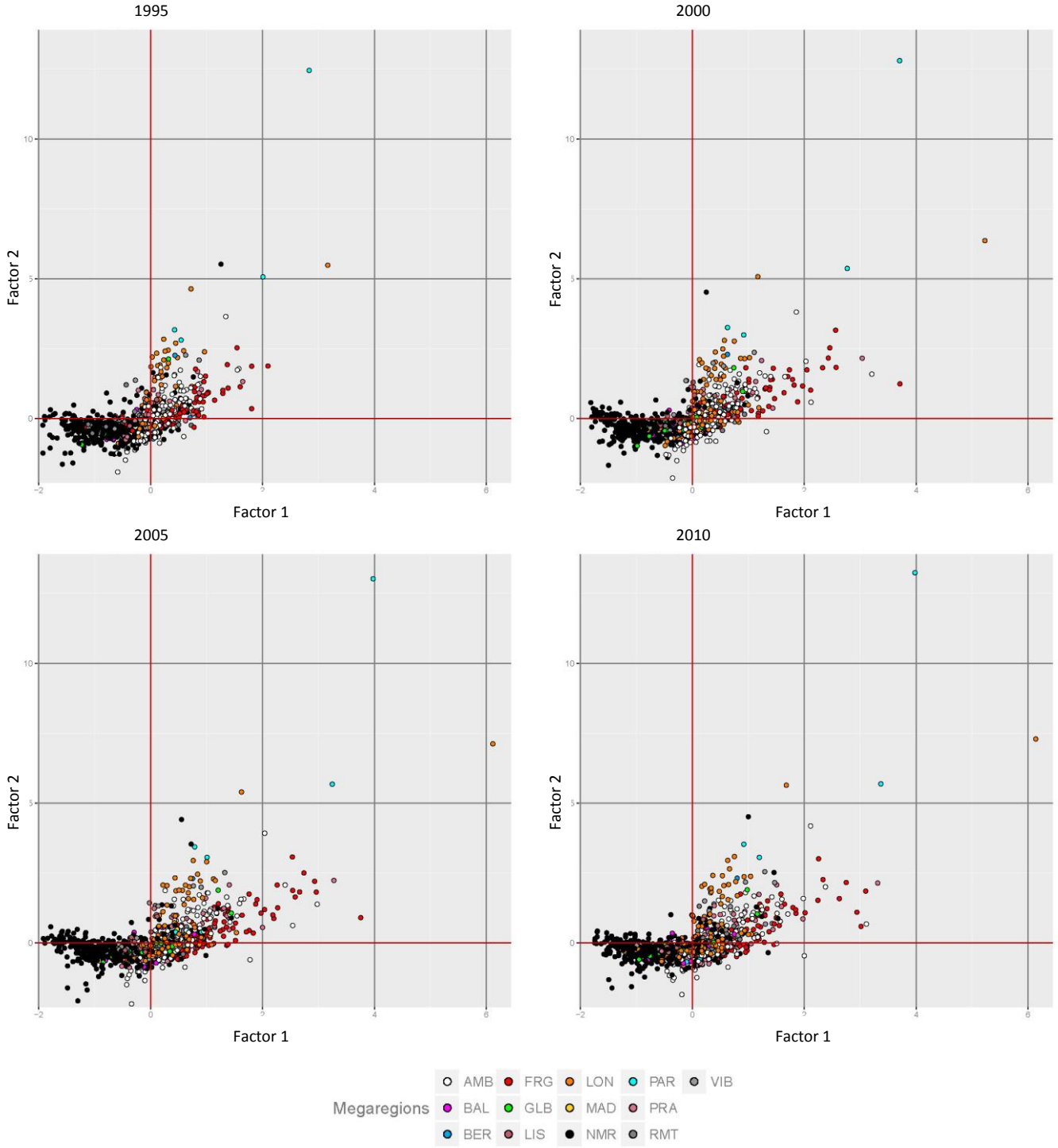

Figure 24 Factor 1 vs Factor 2, NUTS 3 belonging to a megaregion or not (NMR); 1995-2010

Factor 1 vs Factor 3

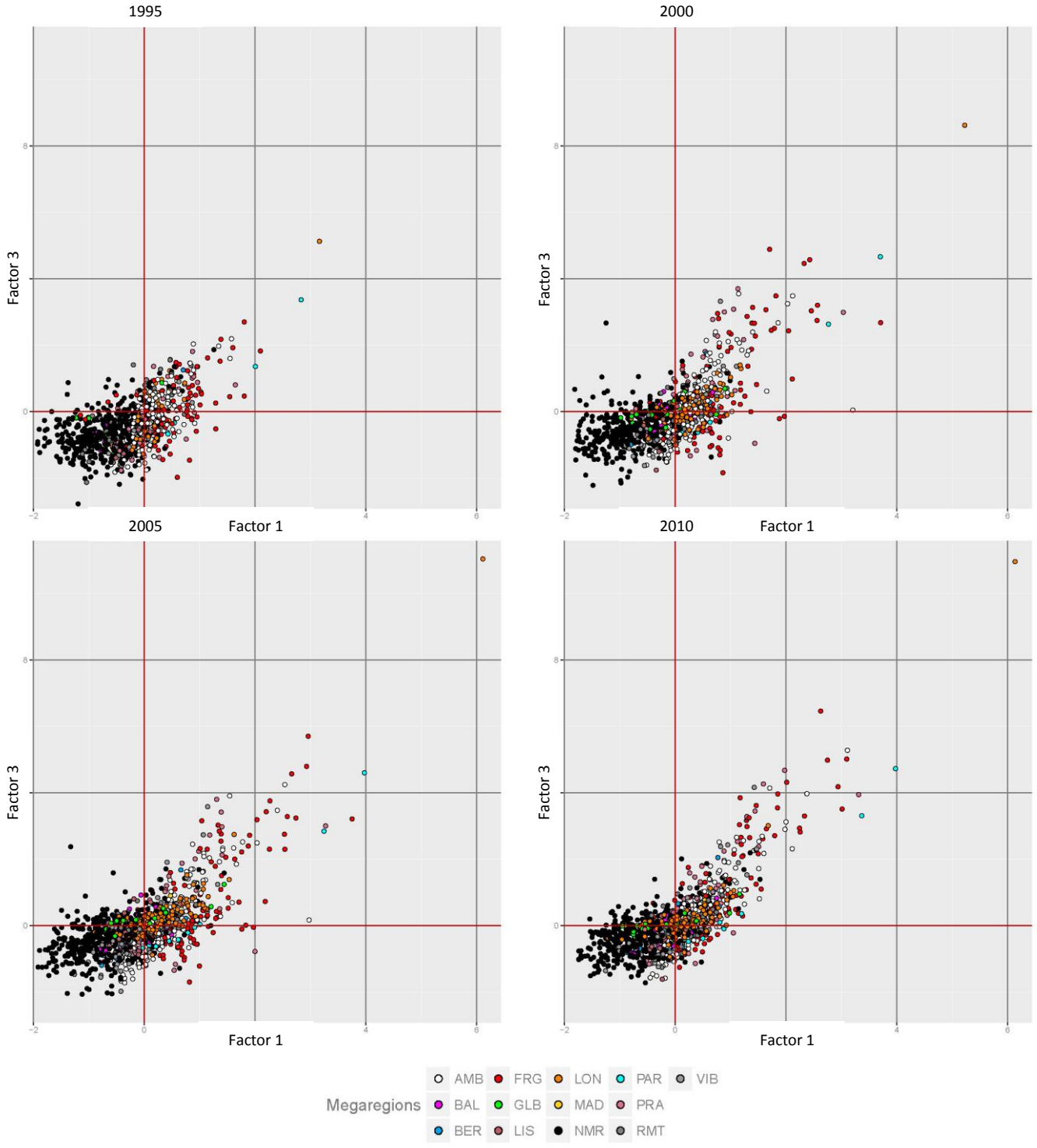

Figure 25 Factor 1 vs Factor 3, NUTS 3 belonging to a megaregion or not (NMR); 1995-2010

Factor 2 vs Factor 3

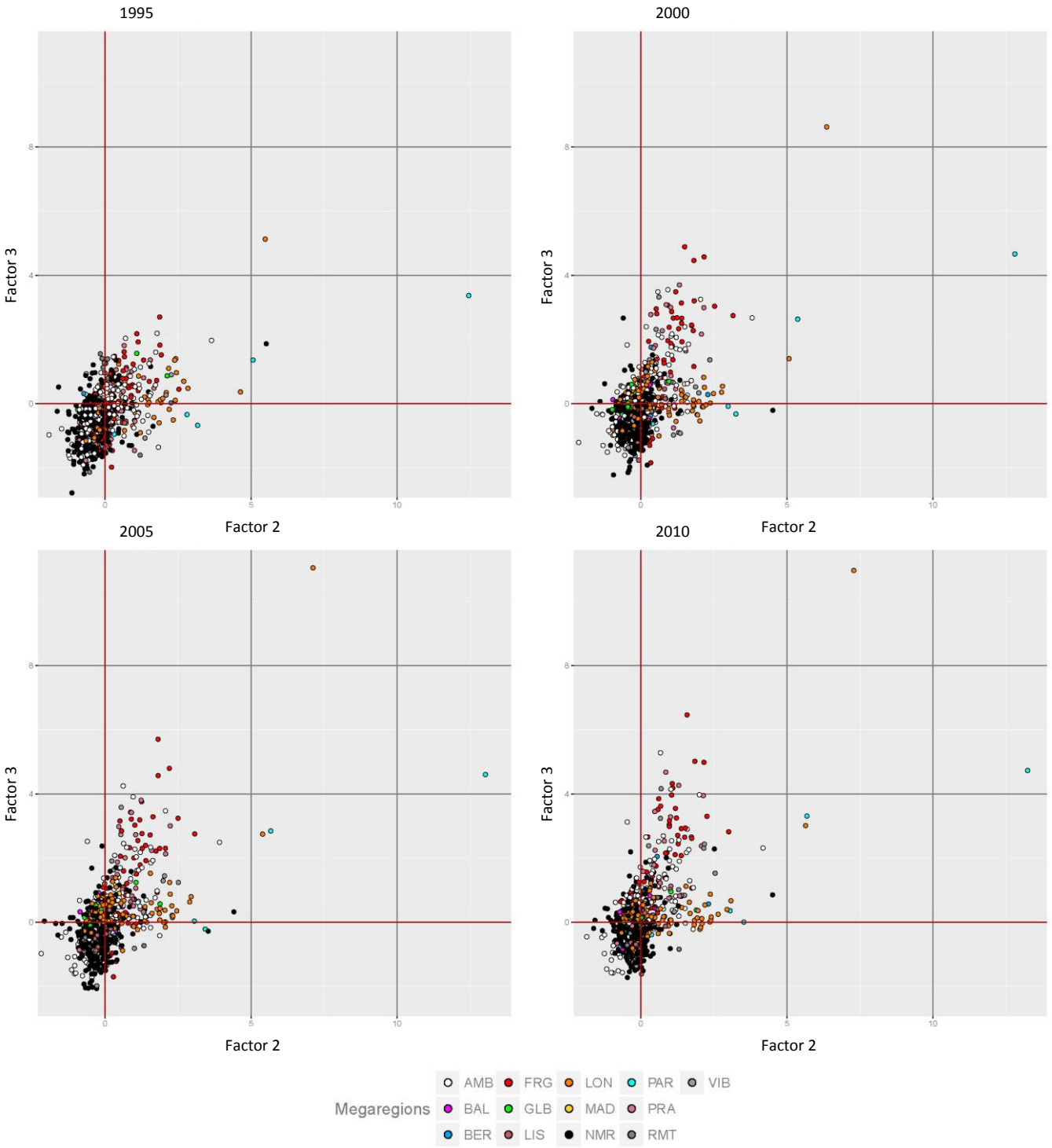

Figure 26 Factor 2 vs Factor 3, NUTS 3 belonging to a megaregion or not (NMR); 1995-2010

Cluster analysis

To determine and visualize the optimal number of clusters, using in this case the K-means partition method, the *fviz_nbclust* function is used. To determine the right number of clusters, the K-means algorithm minimize the Total within Sum of squares for an increasing of cluster. The number *k* is between 4 and 6, with no obvious point of break. This consideration also is based on previous theoretical interpretation and on the factor analysis results (Figure 27).

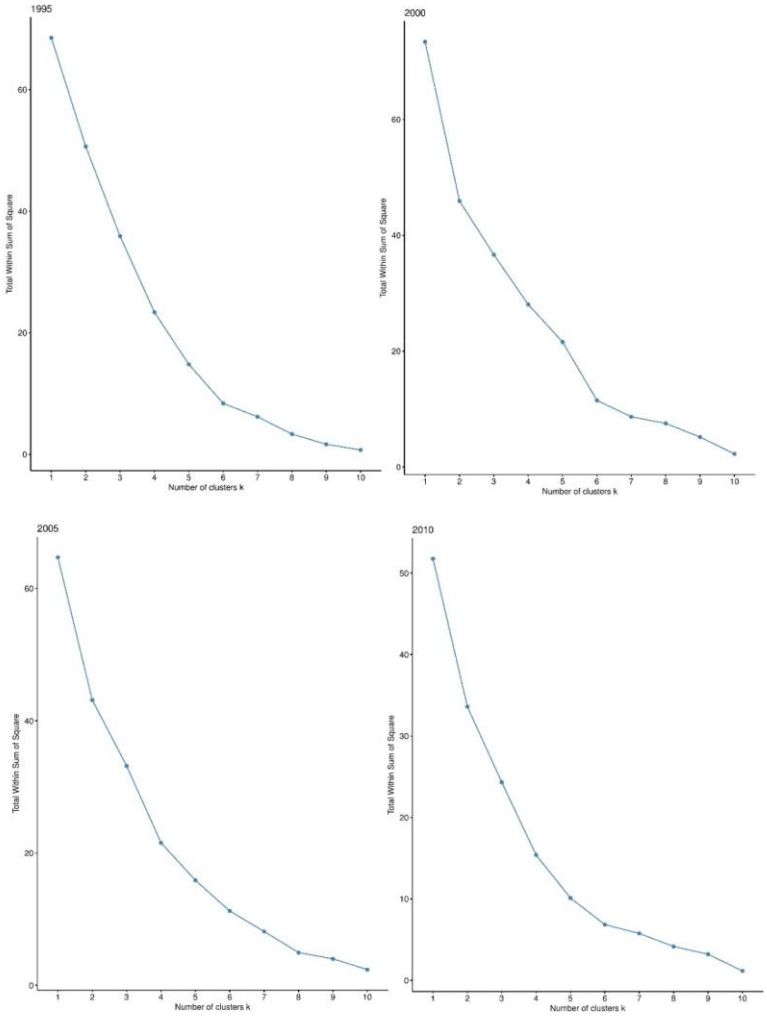

Figure 27 Total sum of squares for different values of k

Megaregions show a decrease in the similarities until the 2005 break, when they reached a maximum the agglomeration to start to disperse again in 2010 (Figure 28). The greatest difference in the megaregions is presented in 1995, were the difference between them makes difficult to stablish homogenous groups. That is because they have characteristics that varies from city to city more easily, like the degree of urbanization or the economic development.

For example, Madrid (MAD), Paris (PAR) and Berlin (BER) appears as a consistent group (from 2000 to 2010) because are very concentrated megaregions with high urban density. On the other side, Frankfurt-Stuttgart (FRG), Amsterdam-Brussels-Antwerp (AMB) and Prague (PRG) conform an instable group (2010) characterized by their socioeconomic development. Finally, some megaregions change their group over time (like Barcelona-Lyon –BAL), probably as consequence of the urban expansion (Figure 4) or the economic crisis.

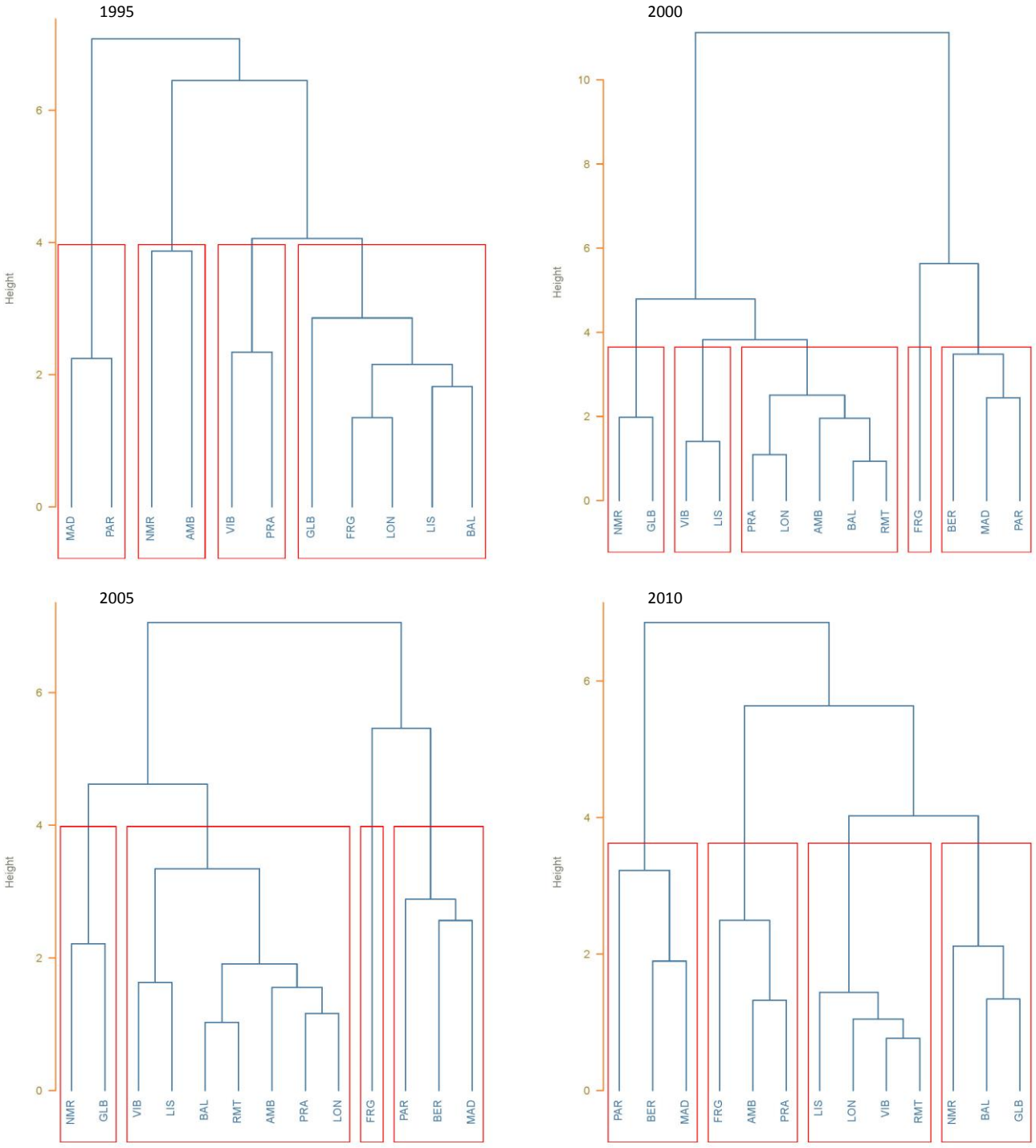

Figure 28 Dendrogram of megaregions clustering; 1995-2010

Model construction

Conceptual approach

When GDP was introduced seven decades ago, it was a relevant notice of progress: increased economic activity was recognized with providing employment, income and amenities to reduce social conflict and prevent another world war. However, the world today is very different from the one faced by the leaders who met to plan the post-war economy in 1944 (see introduction).

The emphasis on GDP in developed countries now fuels social and environmental instability. It also blinds developing countries to possibilities for more sustainable models of urban progress. Rising economic activity has depleted natural resources. Much of the generated wealth has been unequally distributed, leading to a host of social problems (Wilkinson and Pickett, 2009).

The United Nations General Assembly –World Summit Outcome (2005 resolution) has identified economic growth, social cohesion and urban ecology as interdependent and mutually reinforcing pillars of sustainability. While such as the Club of Rome (1970 report) highlighted the unsustainability of current rates of resource depletion.

Elkington (1999) created the term 'triple bottom line' to represent this emerging focus on the three factors of social, environmental and economic added value. This concept has been modified over time and is now often summarized as "people, planet and profit".

More recently, the World Confederation of Productivity Science (WCPS; http://www.wcps.info) has addressed these three factors, suggesting that had to be viewed holistically as a business issue, not as an 'add-on' issue of corporate social responsibility. The current recognition of the impact of climate change clearly demonstrates the complexity of the interaction between social, environmental and economic factors in global urban systems.

The main goal of this study is to create integrated indices of urban network sustainable progress according to different conceptual scenarios, able to measure the performance and the dynamics of urban systems at regional and megaregional scales, based on official (Eurostat) and satellite (NASA) data, and given standardized socio-economic-ecological factors.

The absolute or relative measures of these factors would be helpful in terms of: understanding and diagnosing current urban systems' dynamics; comparing current urban networks performance and long-term behaviors; driving efforts to improve performance. It therefore sought to identify ways in which urban sustainable progress can be measured or assessed.

There are numerous measures for the individual components of sustainable progress but there seems to be no consensus on a measurement or assessment model scenarios representing integrated the combined factors. Any measurement or assessment regime for urban sustainable progress should: provide information in a timely fashion; tailored view so that subsequent improvement actions are not sub-optimal; present information clearly and concisely.

The methodological development is based on the discussion that arose from that presentation and is an attempt to take forward the concepts and propose simple and up to date, but robust and rigorous, urban network sustainable progress indices at regional and megaregional scales.

Methodological development

The OECD (2008) describes how a composite indicator is formed when individual indicators are compiled into a single index on the basis of an underlying model. The idea is that the composite

index captures multi-dimensional concepts that cannot be handled by a single indicator and is therefore relevant combining separate aspects of country performance (i.e. inclusive growth).

The OECD suggests a ten-step guide to building a composite index: theoretical framework, data selection, imputation of missing data, multivariate analysis, normalization, weighting and aggregation, robustness and sensitivity analysis, back to the detail, links to other indicators, and visualization of the results. In general, these ten steps will be used to carry forward our discussion of the construction of the urban network sustainable progress indices.

According to the previously performed multivariate analysis, that have under covered the hidden relationships that govern the observable variables, the goal is now to represent these relationships in a way that they can be used to represent the data in a more integrated way. This means that the indices must be able to be used in new observations and convey a more meaning than the one that can be inferred from a factor analysis and from the scores.

After all, the indices must have a meaning that serves the interest in more socio-economic-ecological way, maintaining the mathematical tools as a support for these statements.

The idea of the urban network sustainable progress indices can be summarized in the next statement: How likely is to find an urban region with the same underlying relationships that are reflected on the observable variables? This statement enables us to use a known probability distribution function to express the indices according to different conceptual scenarios.

Scores distribution

In statistics, a power transform is a family of functions that are applied to create a monotonic transformation of data using power functions. This is a technique used to stabilize variance, make the data more normal distribution-like, and improve the validity of measures of association –such as the Pearson correlation and other data stabilization procedures.

A translation to get positive scores is applied to each one of the factors (F1, F2 and F3), followed by a Box-Cox transformation (Box and Cox, 1964) in order to approximate better to the Gaussian, gives us the transformed scores tFj, for j = 1,2,3:

$$tFj = \frac{F_j^{\lambda_j} - 1}{\lambda j},$$

$$\lambda_1 \approx -2.019;\ \lambda_2 \approx -4.999;\ \lambda_3 \approx -4.346$$

The value of the λj is obtained by function *powertransform* in the library car.

The marginal distributions of the transformed scores seem to adjust better to the Laplace or double exponential distribution than to the Gaussian law, as it can be seen in Figure 29. Recall that the Laplace density and distribution functions are, respectively:

$$f(x) = \frac{1}{2\beta}.exp(-\frac{|x-m|}{\beta})$$

$$\Phi(x) = \frac{1}{2} + \frac{1}{2}sgn(x-m)(1-exp(-\frac{|x-m|}{\beta}))$$

The Laplace density is represented over the histograms of tFj in the right-hand side of the pictures over the histogram of the transformed scores in Figure 30. Given a sample $y1,\ldots,yn$,

the MLE of the parameters m and β are the sample median (50th-percentile) and the mean absolute deviations from the sample median respectively:

$$\hat{m} = C_{50}$$

$$\hat{\beta} = \frac{1}{n}\sum_{i=1}^{n} |yi - \hat{m}|$$

For each transformed factor tFj, the correspondant Laplace distribution function Φj is estimated and compared to the empirical distribution function of the same transformed scores (Figure 30). This can be seen in the Figure 29, pictures on the right-hand side. On the left-hand side and the center, the pictures show that Gaussian does not give a good fit to Fj, fits better to tFj, and Laplace give rise to the best fit in all cases.

From top to bottom, for j = 1,2,3, in each row of the graph we show the histogram of factor *Fj* (left) and tFj (center) jointly with Gaussian density (Figure 29). On the right, the best fit correspond to the Laplace density adjusted to the histogram of tFj.

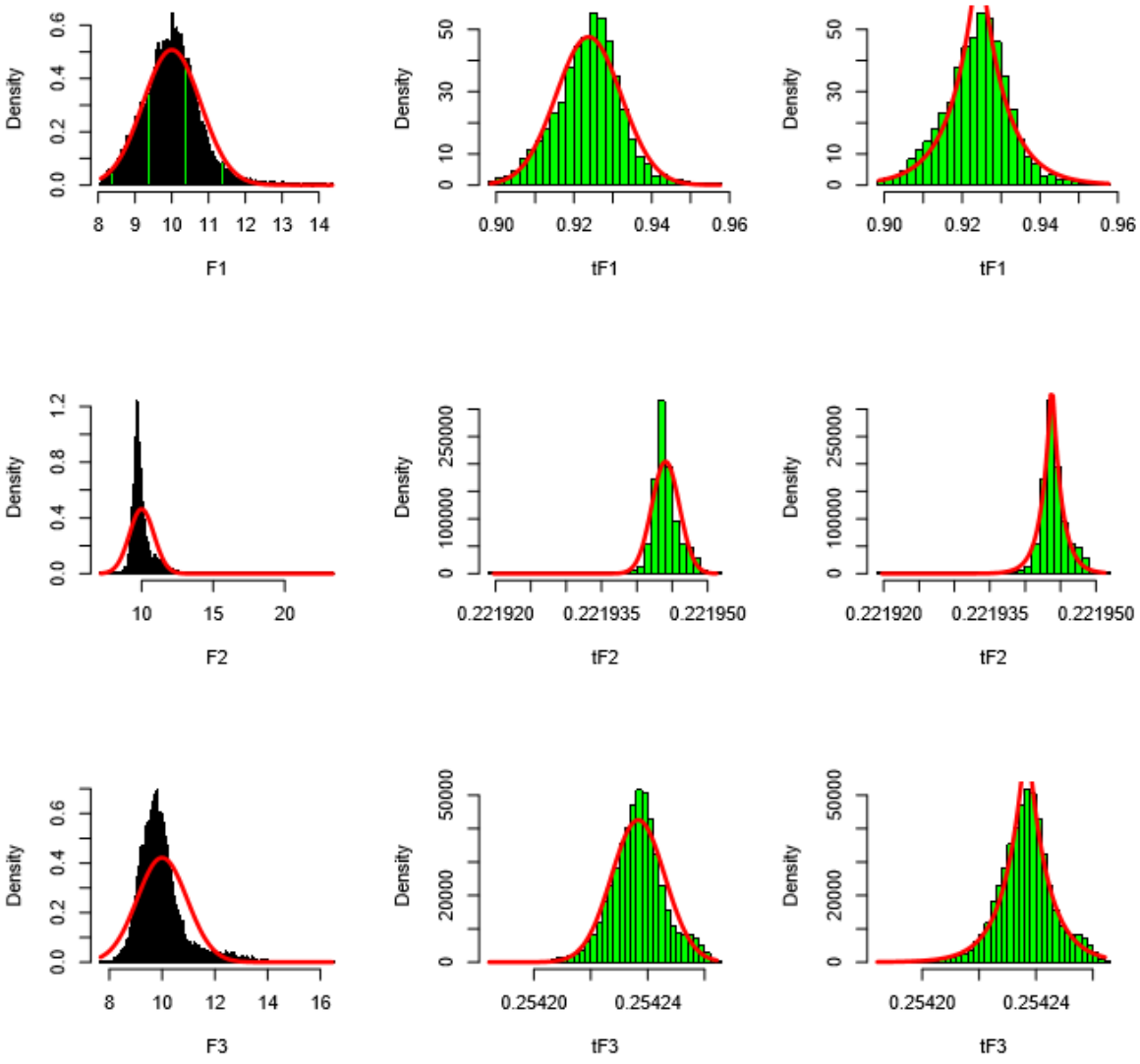

Figure 29 Cumulative Laplace distribution function (black line) and empirical distribution function (gray line) for the transformed factors

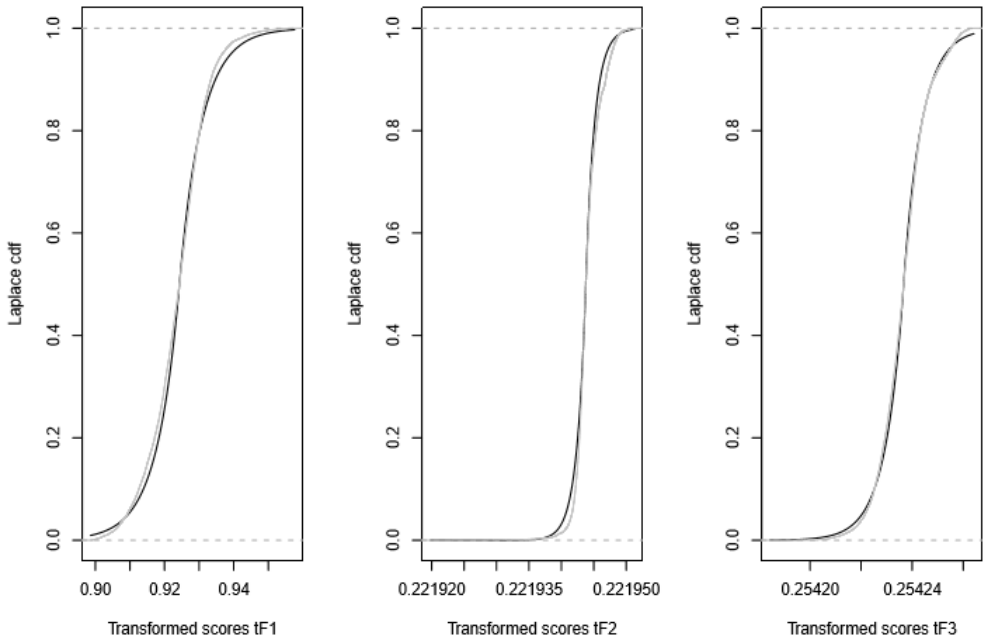

Figure 30 Transformed scores

Indicators calculation

We obtain the indicators for every NUTS 3 region. The head and the tail of the sorted data in decreasing order with respect to F1 are shown in Table 16 and Table 17. Recall that the same order is equivalently defined by either F1, tF1 or the indicator I_1 (*Economic Growth*).

NUTS3	Year	COUn	MGAn	PATth	GDPpc	PECpc	GREpc	URDpsk	URGpor	I_1
UKI11	2008	UK	LON	155,5425	156173	0,42256	161,0082	10265,74	100	0,998973
UKI11	2007	UK	LON	166,4727	152333,6	0,390922	168,6853	10185,19	100	0,99888
UKI11	2006	UK	LON	186,696	144721,5	0,434152	152,3477	10091,67	100	0,998702
UKI11	2009	UK	LON	169,8498	144366,1	0,355486	151,6699	10355,56	100	0,998688
UKI11	2005	UK	LON	175,6068	141928,8	0,43169	147,0692	9956,481	100	0,998601
UKI11	2010	UK	LON	54,06051	145101,7	0,385424	154,7227	10466,67	100	0,998588
UKI11	2004	UK	LON	186,5008	132256,2	0,414322	149,0575	9787,963	100	0,998228
UKI11	2003	UK	LON	167,8657	128314,6	0,401109	143,1429	9652,778	100	0,99801
UKI11	2002	UK	LON	140,5025	125890,6	0,436448	143,6821	9544,444	100	0,997816
UKI11	2001	UK	LON	187,8776	123819,3	0,416944	136,5443	9379,63	100	0,99779

Table 16 Best performing NUTS3 regions

NUTS3	Year	COUn	MGAn	PATth	GDPpc	PECpc	GREpc	URDpsk	URGpor	I_1
BG311	2001	BG	NMR	0,832702	4360,333	1,543374	29,10123	494,7566	8,803165	0,012716
BG311	2006	BG	NMR	8,598452	5193,465	1,969058	33,53478	642,5414	5,967689	0,012649
LT007	2002	LT	NMR	7,473842	4977,578	1,56212	41,32506	390,0875	7,883245	0,01255
LV005	2006	LV	NMR	2,799552	5688,69	1,571649	39,31342	453,2995	5,488612	0,012385
EE008	2002	EE	NMR	0,284657	6726,445	3,799642	36,28638	371,3531	5,701543	0,01232
LV005	2005	LV	NMR	2,761668	5144,988	1,605885	35,46318	386,4461	6,526433	0,01219
EE008	1995	EE	NMR	5,409792	3657,019	3,968758	40,03404	150,7749	14,77821	0,011935
LV005	2000	LV	NMR	16,37972	3433,186	1,60531	30,69662	291,4958	9,173226	0,011809
FI1D7	1999	FI	NMR	42,96675	7422,67	6,346453	31,63193	68,11847	2,814139	0,011778
FI1D4	1996	FI	NMR	10,54852	7000,642	10,03367	30,11119	41,12798	9,331984	0,010557

Table 17 Worst performing NUTS3 regions

The same proceeding is applied to the factor's scores F2 and F3 in order to obtain the indicators I_2 (*Urban Ecology*) and I_3 (*Social Cohesion*).

Plots of the regions and its indicators values

The next plots show each region (NUTS 3) in a color scale according to the indicators in consideration: I_1 (*Economic Growth*, Figure 31), I_2 (*Urban Ecology*, Figure 32), and I_3 (*Social cohesion*, Figure 33). The regions with no assigned value are not included in the analysis, because the information is missing (white color in the maps).

It must be taken into account that despite the fact that only real information was used to analyze the relationships between the variables and to create the indicators, in the plots a small amount of imputation is made, according to the imputation method before mentioned.

It is important to note that the three indicators measures the urban network performance (for instance, there could be a region with a great percentage of open spaces and environmental quality, but with inefficient urban ecological functionality –like disperse urbanizations).

In general, the maps represent consistently the economic (Figure 31), ecological (Figure 32) and social (Figure 33) urban system dynamics of European NUTS 3 regions from 1995 to 2010. These indicators will be used to calculate the integrated index of urban network sustainability progress according to different scenarios (using indicator adjustment –with penalization).

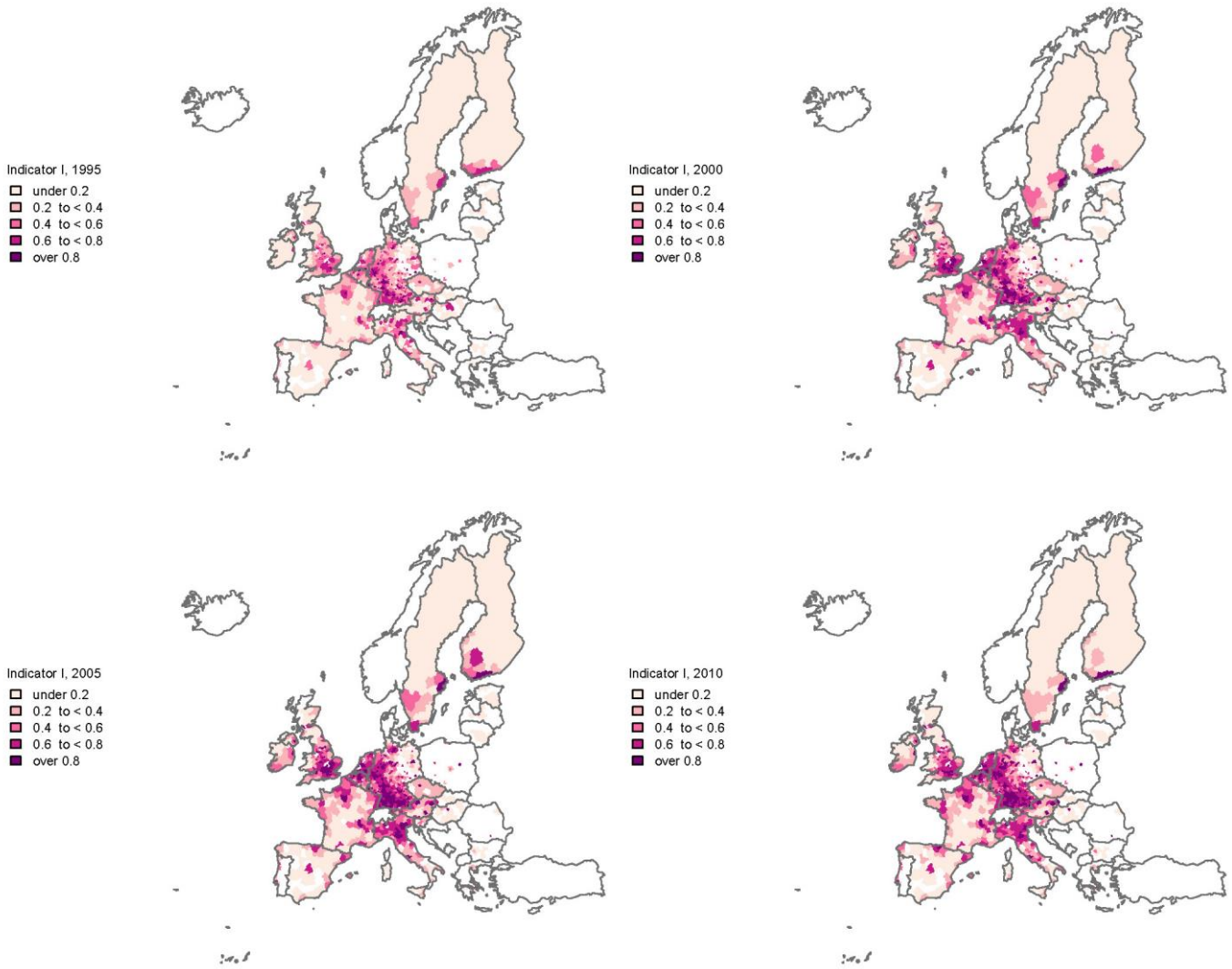

Figure 31 Indicator I –*Economic Growth* at NUTS3 level (1995-2010)

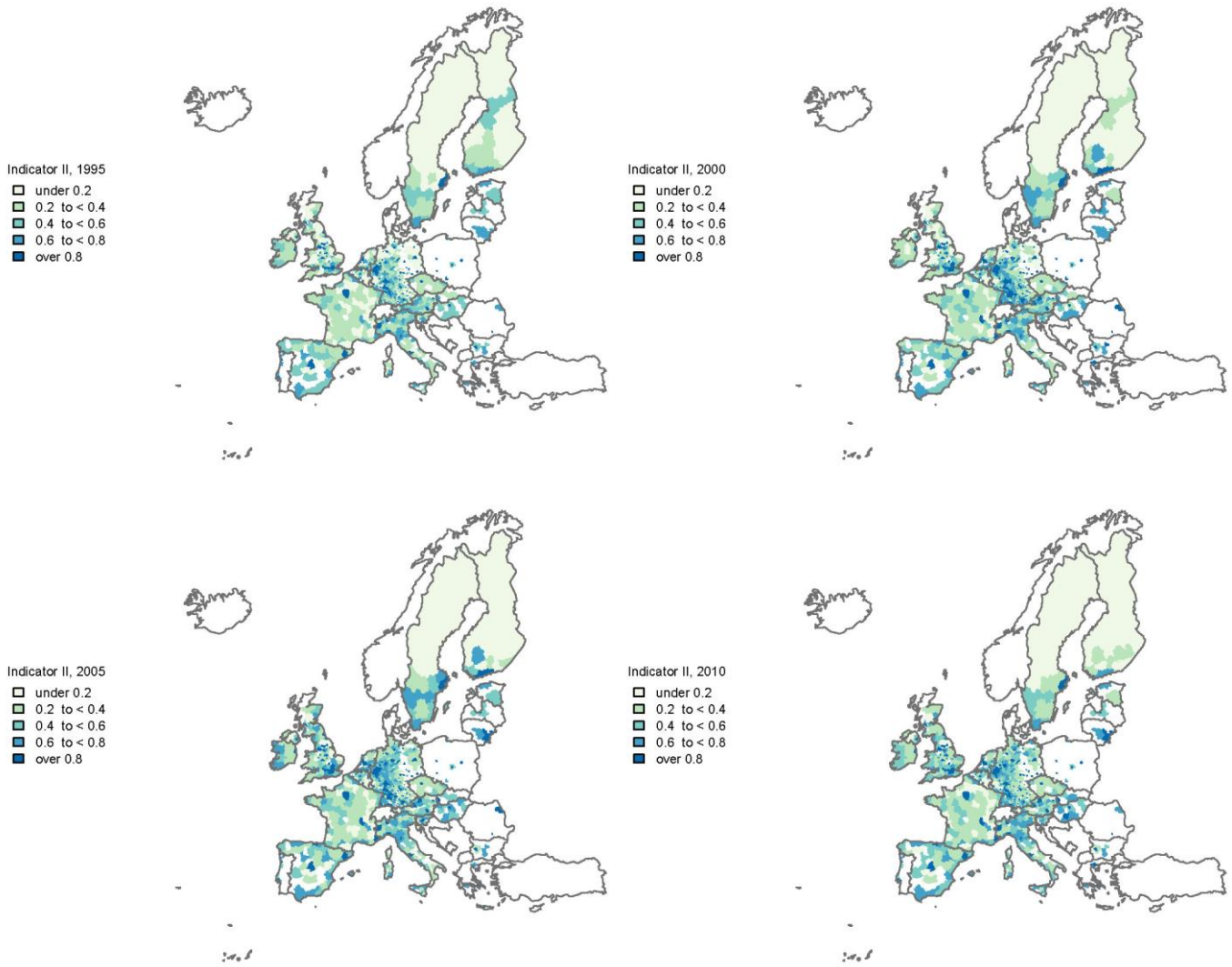

Figure 32 Indicator II –*Urban Ecology* at NUTS3 level (1995-2010)

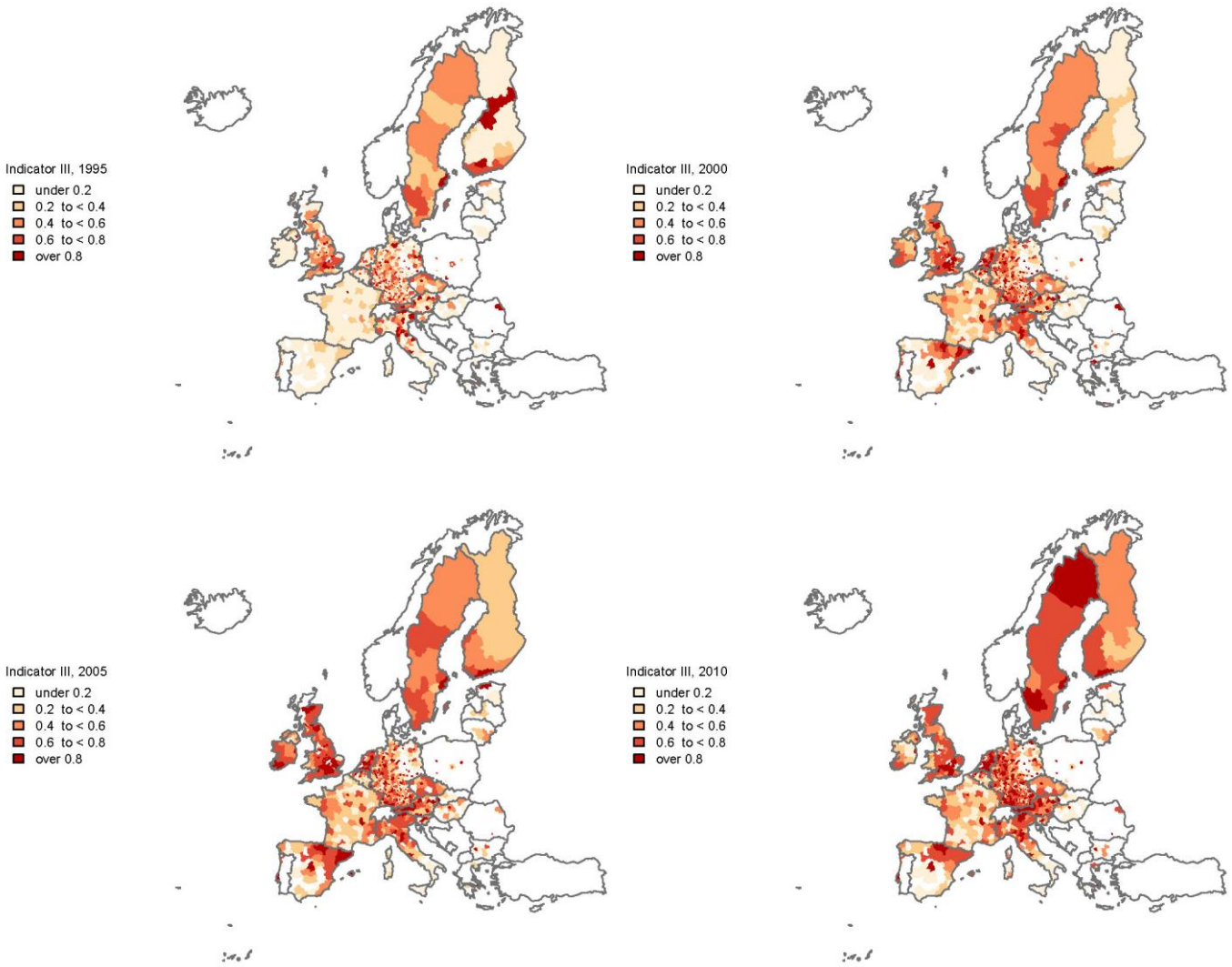

Figure 33 Indicator III –*Social Cohesion* at NUTS3 level (1995-2010)

Adjustments: penalization

Consider a latent factor, denoted F_1 and provided by some factorial method (factor analysis, structural equations, among others). F_1 is a random variable satisfying the assumptions of the undelaying model. The corresponding scores $\{f_{11}, \ldots, f_{1n}\}$ can be seen as observed values of this random variable. The natural order in R, defines an obvious index on the scores, giving rise to a total ordering between any pair of cases, in this way:

$$i \prec j \; if \; and \; only \; if \; f_{1i} \leq f_{1j} \; (1)$$

Assuming that scores indicate best performance of the case as greater their value is, the downside is that are unbounded, and their isolated value cannot be interpreted. To circumvent this problem, some monotonic function Φ: $R \rightarrow [0,1]$ (or to another bounded interval instead of the unitary one) should be applied to F_1. We propose using the transformation given by the distribution function Φ_1 of the random variable F_1. As the distribution function is increasing, we have:

$$i \prec j \; if \; and \; only \; if \; f_{1i} \leq f_{1j} \; if \; and \; only \; if \; \Phi_1(f_{1i}) \leq \Phi_1(f_{1j}). \; (2)$$

This function is intrinsic to the observations and the nature of the variable. The value $\Phi_1(f_{1j})$ is the cumulative percentage of cases whose scores are lower or equal to f1i. In this way, it is the percentile position of the case, according to the scores distribution.

Definition 1. Given any random variable F_1 taking real values, consider the index I_1 defined by the distribution function $\Phi 1$ of F_1 : $I_1 := \Phi_1(F_1)$ (3). This index takes values between 0 and 1, and preserves the order given by F_1. Moreover, as close to 1 as better the case position is and, conversely, values near 0 indicate bad positioning. Index I_1 is a random variable following a uniform distribution in the unit interval. If Φ is not known but it belongs to some parametric family, parameters can be estimated as Φ_1 from the scores, and the index defined by means of this estimator. If any parametric family can not be assumed, the empirical distribution function (denoted by Φ_1^n) could be used instead of the actual distribution. Habitually, factorial models involve more than one factor. Consider two factors F_1, F_2 jointly evaluated on the same cases. The goal could be indexing the cases by some kind of ordering taking into account both I_1 and I_2, in a way that prioritizes the first index I_1, but penalizes negatively cases showing low values in I_3, conversely, and penalizes positively taking higher values on the second index. In this way, we suggest a new kind of index depending on a weight parameter $w \in [0,1/2]$ expressing the penalty degree.

Definition 2. Given any pair of random variables F_1 and F_2 taking real values, consider the index Iw 12 defined by the distribution functions Φ_1, Φ_2 and scalar $w \in [0, 1\,2]$:

$$I_{12}^w = \Phi_1(F_1) - w(\Phi_1(F_1) - \Phi_2(F_2)) = (1-w)\Phi_1(F_1) + w\Phi_2(F_2) = (1-w)I_1 + wI_2$$

Remark 1. Notice that the right-hand side is a convex combination of $I1 = \Phi(F1)$ and $I2 = \Phi(F2)$ and weights $w \in [0,1]$ can be used. The constraint $w \leq 1/2$ implies $(1-w) > w$ and expresses the idea that F_1 is the primer factor to define the ordering and F_2 plays a secondary role, the role being partially if $w = 1/2$. Some immediate relations can be easely shown.

Properties.

1. For any $w \in [0,1], I_{12}^w \in [0,1]$.
2. For any $w \in [0,1/2], if \; F_2 \leq F_1, then \; I_{12}^w \leq I_1$.

3. For any $w \in [0,1/2], if\ F_1 \leq F_2, then\ I_{12}^w \geq I_1$.

In a more general way, more than two factors can be combined to define a weighted bounded index, taking values in [0,1].

Definition 3. Given random variables $F_1, \ldots, F_k$ taking real values, consider the index Iw 1…k defined by the distribution functions $\Phi_1, \ldots, \Phi_k$ and scalars $w = (w1, \ldots, wk)$ sorted in increasing order $w_1 \geq w_2 \geq \cdots \geq w_k$ and satisfying $\sum_{j=1}^{k} w_j = 1$:

$$I_{1\ldots k}^w := \sum_{j=1}^{k} w_j \Phi_j(F_j) = \sum_{j=1}^{k} w_j I_j \quad (5)$$

Remark 2. Notice that the index in (4) is a particular case of this general index, with $k = 2$, $w1 = 1 - w$ and $w2 = w$. Another example is given by

$$I_{123}^w := 1/3(\Phi_1(F_1) + \Phi_2(F_2) + \Phi_3(F_3)) = 1/3(I_1 + I_2 + I_3). (6)$$

Illustration. To better understand the use of indices in (3) and (4), two standard Gaussian random samples of size 100 with some amount of correlation (ρ = 0.3) are generated (see Table 18). We divide the sample into two subsamples, the first one characterized by cases satisfying F2 < F1 and the second one composed by cases satisfying F1 ≤ F2. Ordered values of both subsamples are shown in Table 19.

F1	F2
0.90	0.97
0.31	-1.00
-0.36	1.15
-0.12	0.45
…	….

Table 18 Header of a b-variate standard Gaussian sample of size 100 and correlation 0.3

F1	F2	F1	F2
-1.82 -	0.84	-0.89	-1.87
-1.81	-1.73	-0.65	-1.82
-1.81	-0.39	-0.63	-0.67
-1.53	-1.16	-0.62	-2.75
…	….	…	….

Table 19 Header of subsample F1 ≤ F2 on the left-hand side and subsample F1 ≥ F1 on the right. Cases are sorted by F1 in increasing order.

In the top left corner of Figure 34, we represent the transformation on F_1 giving rise to index I_1, and left to right, top to bottom we show the effect of penalizing cases using index Iw 12, for weights w = 0.1 to w = 0.5, by 0.1. In Figure 35, the same indices are shown together with the two subsamples behavior: the red points corresponding to cases with F2 ≤ F1 are negatively penalized, and the blue points to cases F2 > F1 and are positively penalized. In both figures, we appreciate clearly the differences between taking a low weight w = 0.1 on the top-left and the highest weight w = 0.5 on the dawn-right.

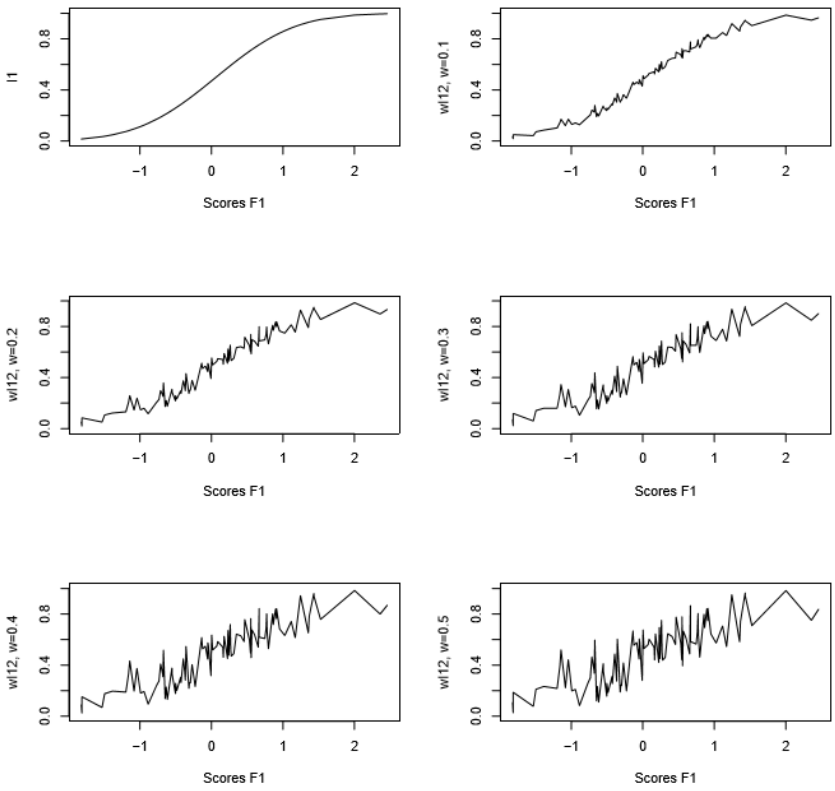

Figure 34 Effect of different values of penalization in F1

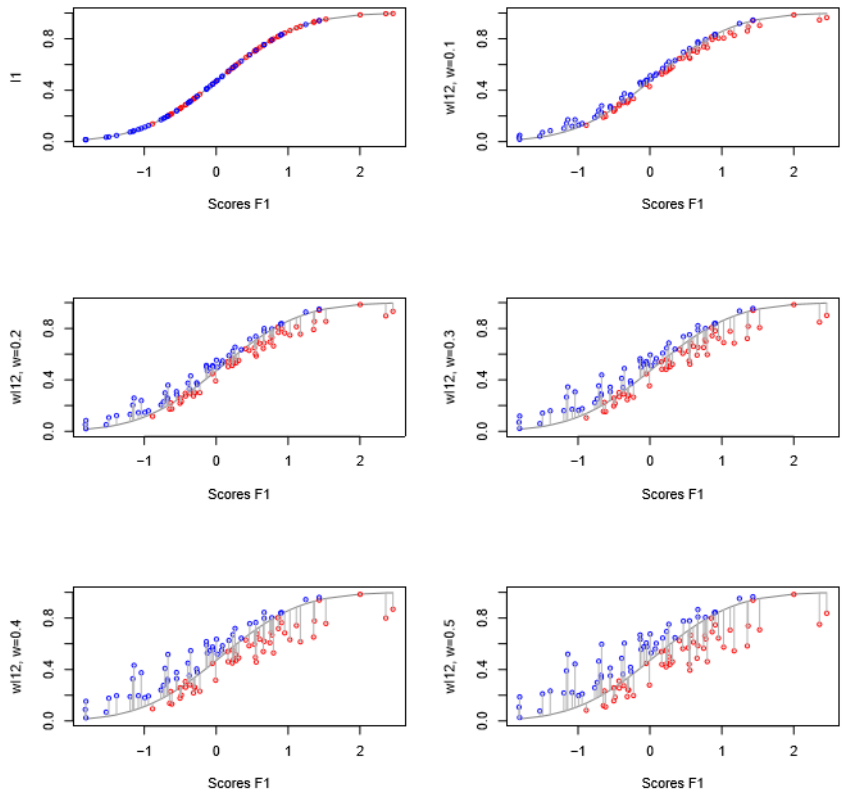

Figure 35 Effect highlighted values that increase (in blue) and that decrease (in red)

Scenarios definition

According to different criteria on what means "urban sustainable progress", four conceptual scenarios (S_n) using the three complementary indicators (I_1 –*Economic Growth*, I_2 –*Urban Ecology*, I_3 –*Social Cohesion*) are modelled:

- S_1 *Economic Development* (I_1, penalizing very low values of I_2 and I_3)
- S_2 *Social Sustainability* (I_3, without I_1 and penalizing low values of I_2)
- S_3 *Environmental Sustainability* (I_2, without I_1 and penalizing low values of I_3)
- S_4 *Inclusive Growth* ($S_{4.1}$: balance in I_1, I_2 and I_3; $S_{4.2}$: penalizes unbalance)

For the four suggested scenarios, several penalizations are used (Table 20). It is important to note that in all the scenarios the weights can be modified according to the user criteria.

S_1 *Economic Development*			S_2 *Social Sustainability*		
I_1	*Economic Growth*	$w_1 = 0.8$	I_1	*Economic Growth*	$w_1 = 0$
I_2	*Urban Ecology*	$w_2 = 0.1$	I_2	*Urban Ecology*	$w_2 = 0.2$
I_3	*Social Cohesion*	$w_3 = 0.1$	I_3	*Social Cohesion*	$w_3 = 0.8$
S_3 *Environmental Sustainability*			S_4 *Inclusive Growth*		
I_1	*Economic Growth*	$w_1 = 0$	I_1	*Economic Growth*	$w_1 = 1/3$
I_2	*Urban Ecology*	$w_2 = 0.8$	I_2	*Urban Ecology*	$w_2 = 1/3$
I_3	*Social Cohesion*	$w_3 = 0.2$	I_3	*Social Cohesion*	$w_3 = 1/3$

Table 20 Scenarios of urban network sustainable progress

Scenario S_1 – *Economic Development*

The more mainstream of all, this is the scenario of neoclassical economic theory, were the economic growth must be maximum and practically unrestricted by other factors. To be able to visualize the regions that comply with this standard, the compound index S_1 has its weights: $w_1 = 0.8$, $w_2 = 0.1$, $w_3 = 0.1$, set to maximize the economic factor (without neglecting the others).

$$S_1 = w_1 I_1 + w_2 I_2 + w_3 I_3;$$

$$w_1 + w_2 + w_3 = 1$$

Scenario S_2 – *Social Sustainability*

In this case, we maximize the social equality factor. The economic factor does not play explicitly any role (however, could be an important variable in the expression of the partial indicators). Consequently, the index take into account the indicators I_2 and I_3 in the following way:

$$S_2 = w_2 I_2 + w_3 I_3;$$

$$w_2 + w_3 = 1$$

The way that this scenario is constructed is by set $w_2 = 0.2$ and $w_3 = 0.8$.

Scenario S_3 – *Environmental Sustainability*

In this case, we maximize the urban ecology factor. The economic factor does not play explicitly any role (only as a variable in the expression of the other indicators). Accordingly, the index take into account the indicators I_2 and I_3:

$$S_3 = w_2 I_2 + w_3 I_3;$$

$$w_2 + w_3 = 1$$

This scenario is constructed by set $w_2 = 0.8$ and $w_3 = 0.2$.

Scenario S_4 – *Inclusive Growth*

In this scenario are the regions that are over the average in all three indicators I_1, I_2 and I_3. We propose two indices: $S_{4.1}$ (inclusive equal weigh), and $S_{4.2}$ (inclusive and balanced):

$S_{4.1}$ is an inclusive equal weigh index:

$$S_{4.1} = \frac{1}{3} I_1 + \frac{1}{3} I_2 + \frac{1}{3} I_3;$$

$$w_1 = \tfrac{1}{3}, w_2 = \tfrac{1}{3}, \mathrm{w}_3 = \tfrac{1}{3}$$

$S_{4.2}$ is an inclusive and also balanced index, that penalizes the unbalance between I_1, I_2 and I_3:

$$\mathrm{I} = \left(\frac{1}{3} + 2\beta\right) \mathrm{Min}\{I_1, I_2, I_3\} + \frac{1}{3} \mathrm{Med}\{I_1, I_2, I_3\} + \left(\frac{1}{3} - 2\beta\right) \mathrm{Max}\{I_1, I_2, I_3\}$$

$\beta \leq \frac{1}{6}$. In particular, $S_{4.2}$ corresponds to $\beta = \frac{1}{12}$, therefore:

$$S_{4.2} = \frac{1}{2} \mathrm{Min}\{I_1, I_2, I_3\} + \frac{1}{3} \mathrm{Med}\{I_1, \mathrm{I}_2, I_3\} + \frac{1}{6} \mathrm{Max}\{I_1, I_2, I_3\}$$

Model results

Territorial units at regional scale

The next plots show each region (NUTS 3) in a color scale according to the urban network sustainable progress conceptual scenario in consideration: S_1 (*Economic Development*; Figure 36), S_2 (*Social Sustainability*; Figure 37), S_3 (*Environmental Sustainability*; Figure 38), $S_{4.1}$ (*Inclusive Growth* –equal weigh; Figure 39), $S_{4.2}$ (*Inclusive Growth* –balanced; Figure 40). The regions with no assigned value (white color in the maps) are not included in the analysis.

In general, the maps represent consistently the economic development (Figure 36), the social sustainability (Figure 37) and the environmental sustainability (Figure 38) of the urban network dynamics in the European NUTS 3 regions from 1995 to 2010. The first two indices (S_1 and S_2) clearly improve over time (S_2 decreases in the last period probably as consequence of the financial crisis), but this trend is not evident for the third index (S_3). European regions seem to advance more in economic terms than in socio-environmental ones.

More interesting are the behavior of the inclusive growth indices –in the equal weigh version (Figure 39) and the balanced version (Figure 40). Both approaches reflect the regions with more equilibrated sustainable progress (in social, ecological and economic terms), and the resilience of urban networks (for instance, against perturbations like the financial crisis) in comparison with less urbanized regions. We will analyze in more detail these indices in the next section.

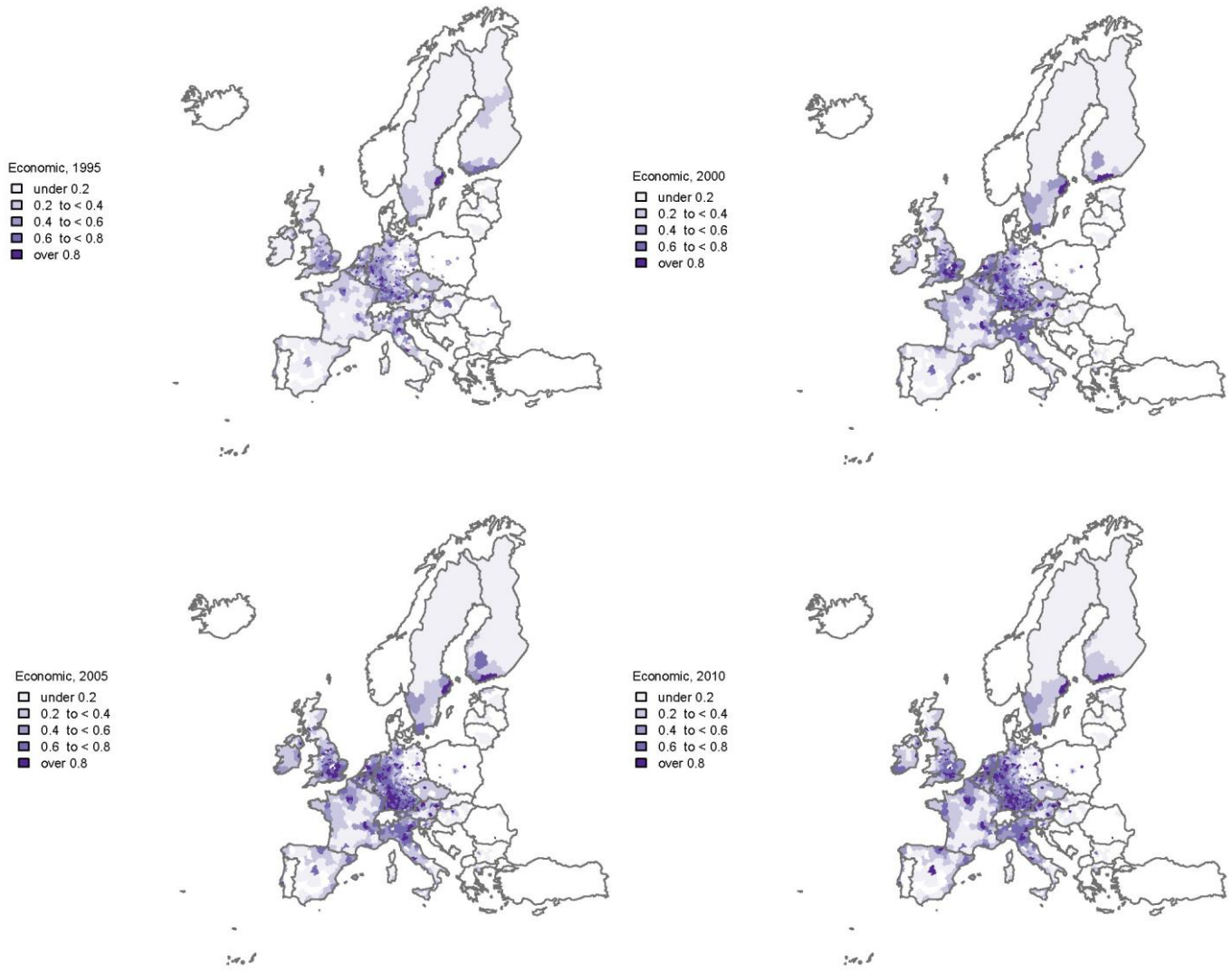

Figure 36 Scenario *S_1 –Economic Development* at NUTS3 level (1995-2010)

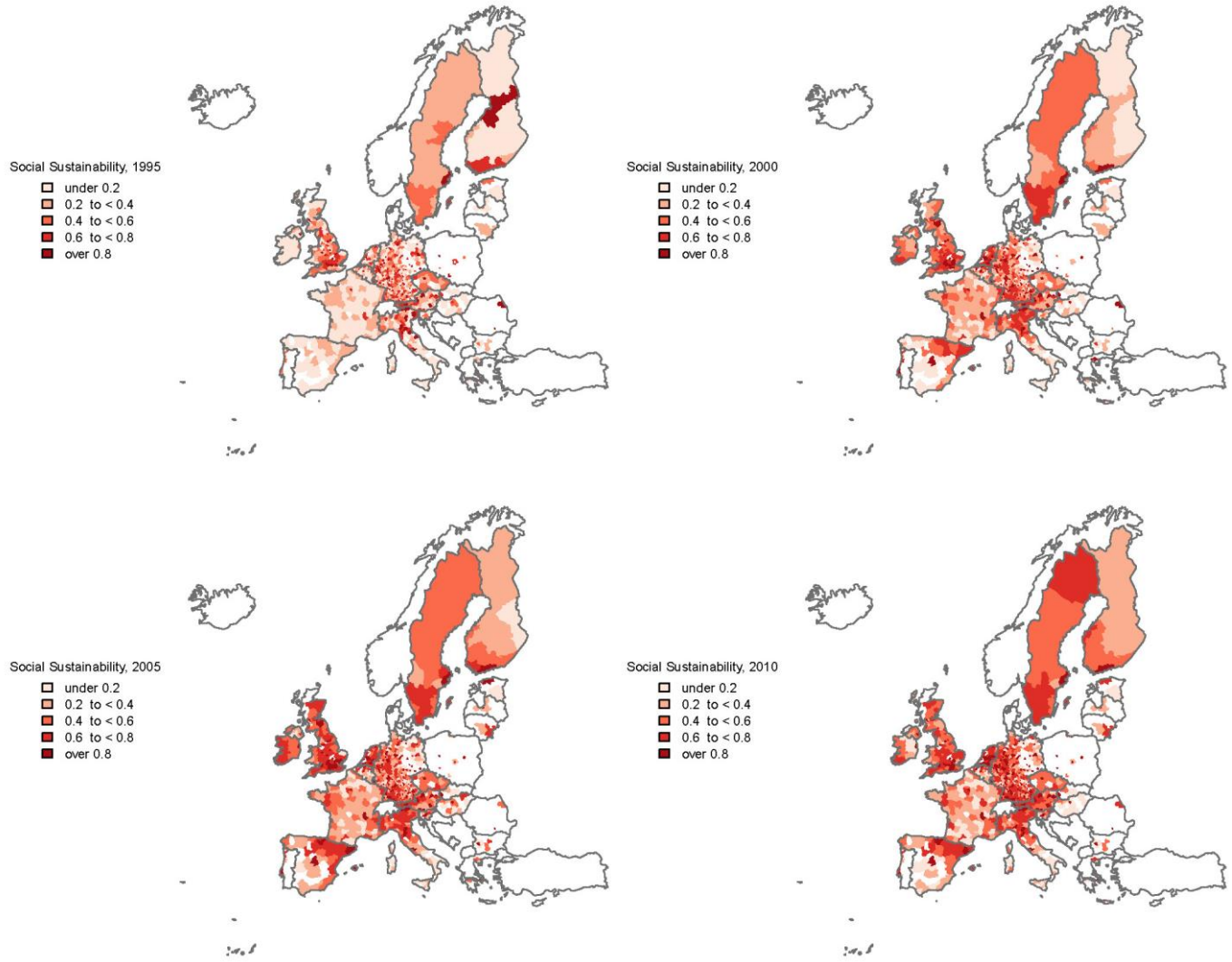

Figure 37 Scenario S_2 –*Social Sustainability* at NUTS3 level (1995-2010)

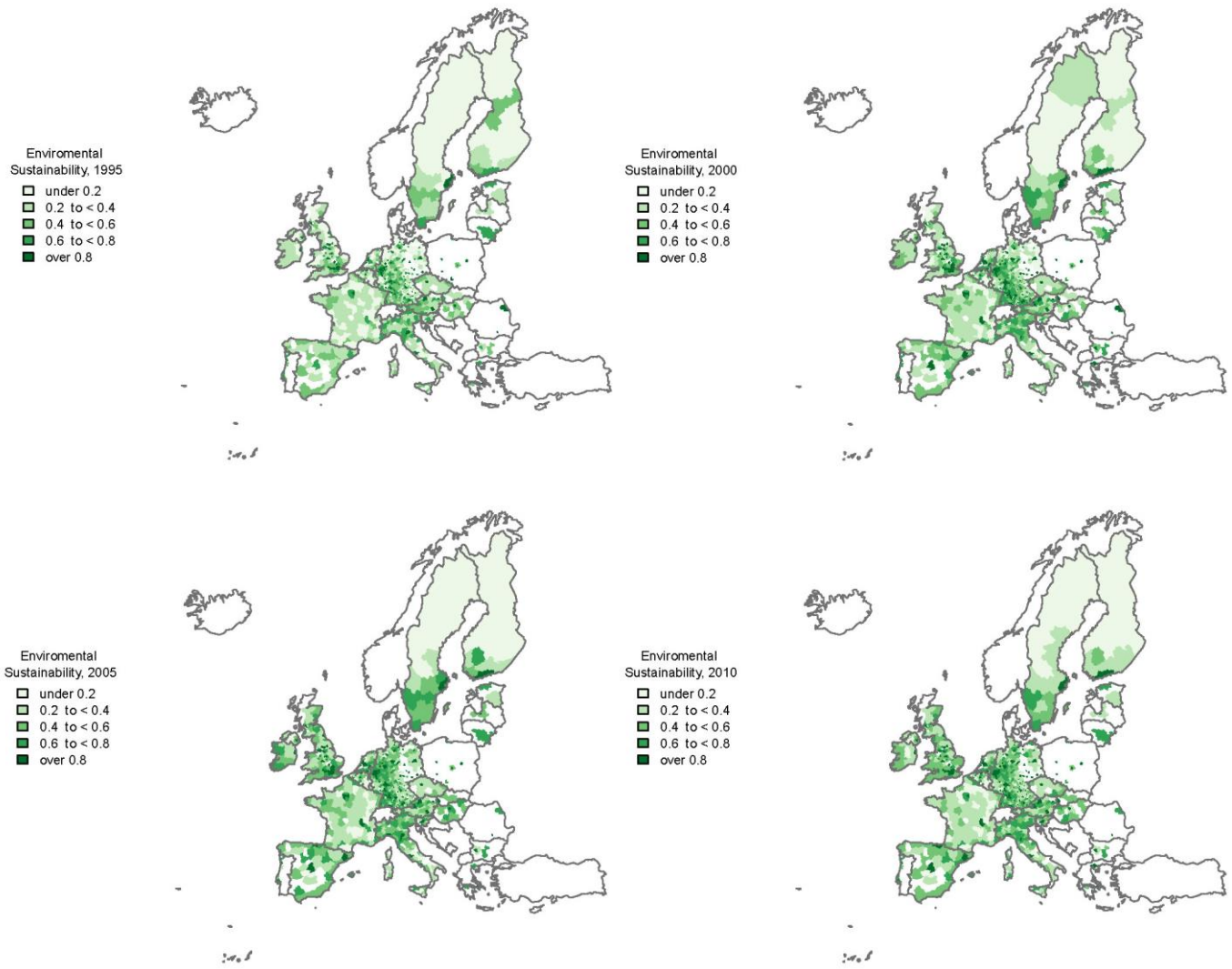

Figure 38 Scenario S_3 –*Environmental Sustainability* at NUTS3 level (1995-2010)

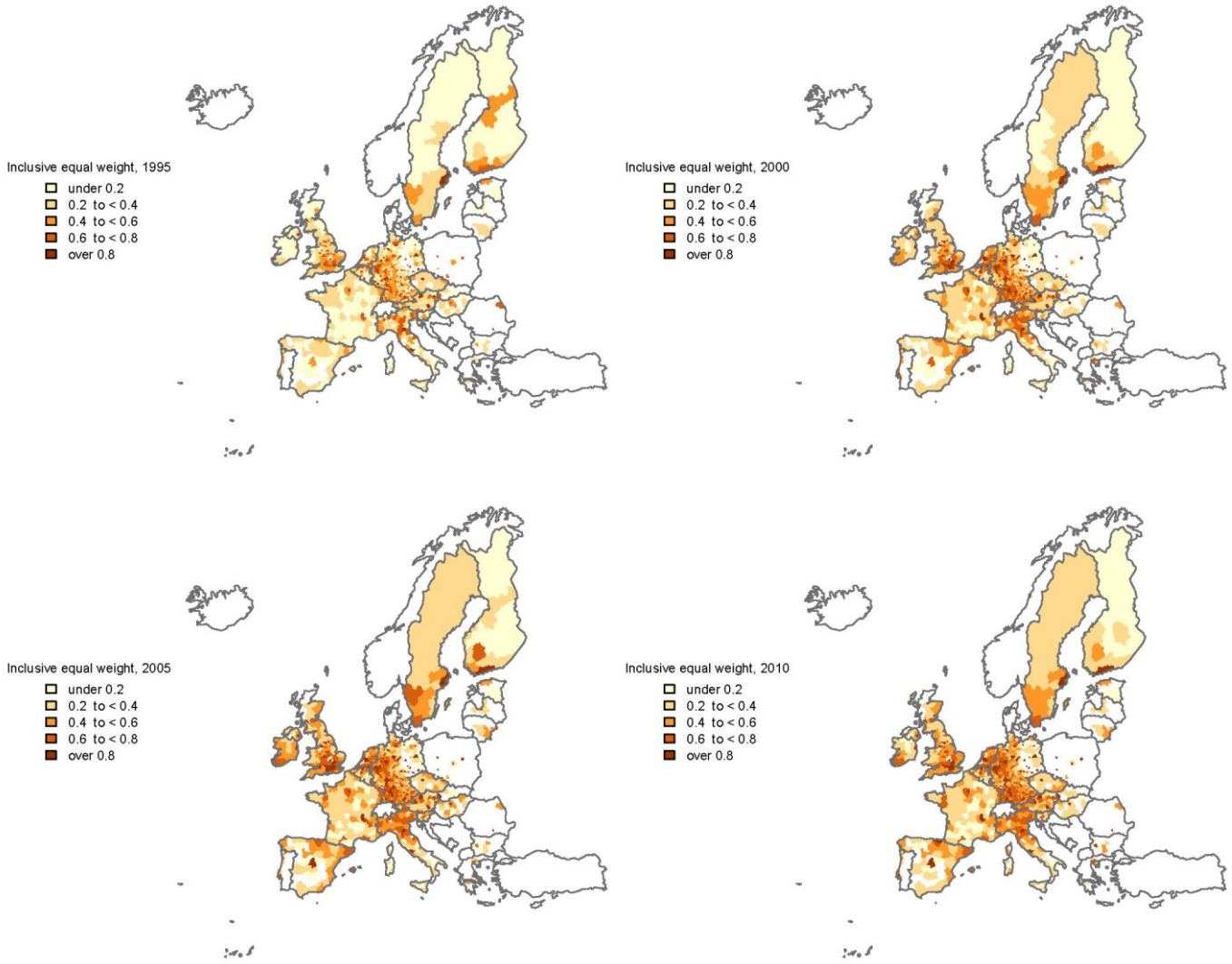

Figure 39 Scenario $S_{4.1}$ –*Inclusive Growth* (equal weigh) at NUTS3 level (1995-2010)

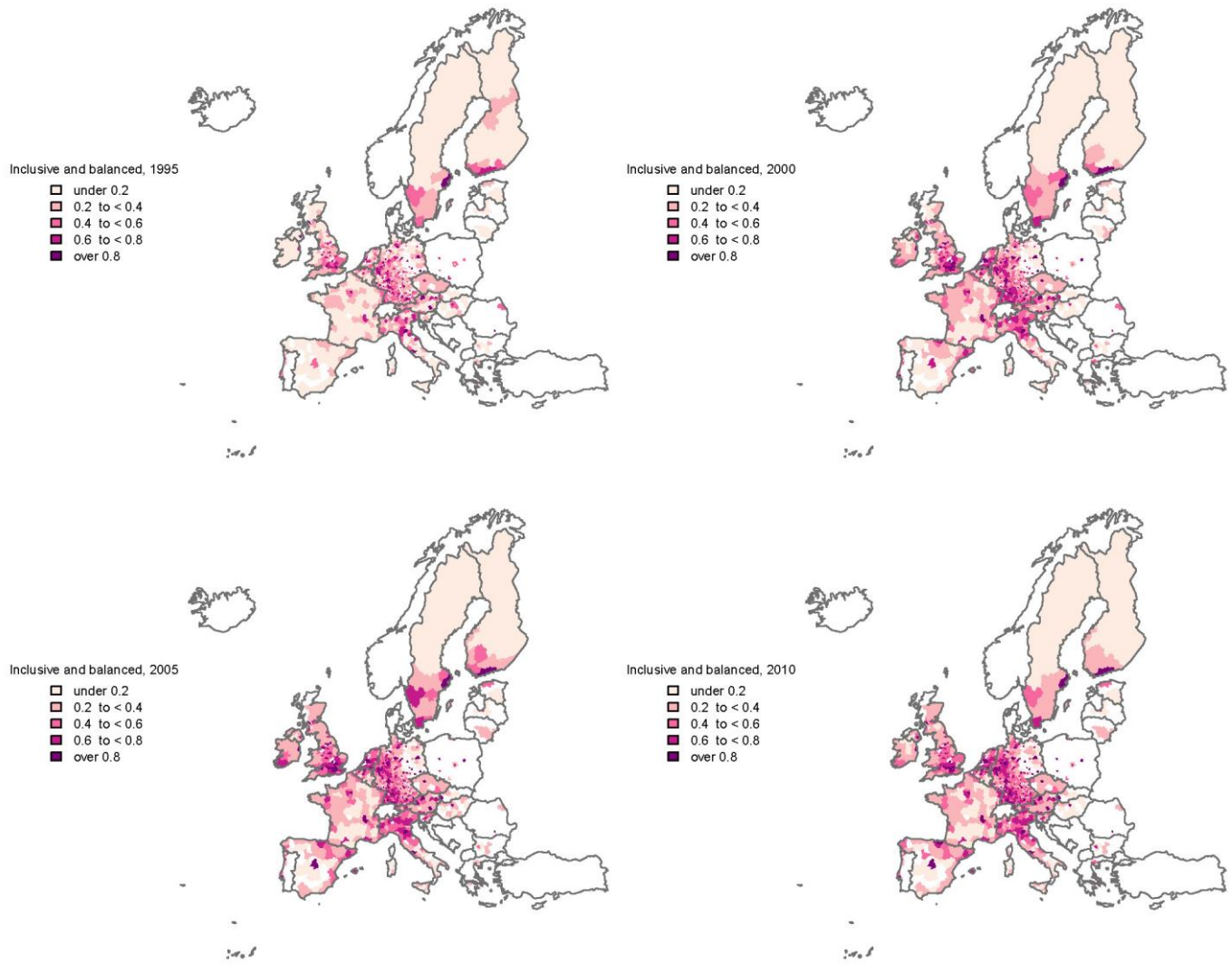

Figure 40 Scenario $S_{4.2}$ –*Inclusive Growth* (balanced) at NUTS3 level (1995-2010)

Urban networks at megaregional scale

The application of the indicators I_1 *–Economic Growth,* I_2 *–Urban Ecology, and* I_3 *–Social Cohesion* at megaregional level (Table 21) in the period of analysis (1995-2010), show higher economic values in Paris (PAR; from I_1 = 0,76 to I_1 = 0,89), Frankfurt-Stuttgart (FRG; from I_1 = 0,24 to I_1 = 0,85) and Amsterdam-Brussels-Antwerp (AMB; from I_1 = 0,54 to I_1 = 0,76). The results also show better ecological behavior in Berlin (BER; I_2 = 0,89), Madrid (MAD; I_2 = 0,84) and Paris (PAR; I_2 = 0,82), mainly due to their urban density; and higher social cohesion in FRG (I_3 = 0,89) and PAR (I_3 = 0,87). No-megaregions (NMR) present the lowest values in the three indicators of urban sustainability (I_1 = 0,05; I_2 = 0,34; I_3 = 0,02).

In 2010, Barcelona-Lyon (BAL) was the 8th megaregion in economic growth (I_1 = 0,42), the 9th megaregion in urban ecology (I_2 = 0,65), and the 11th megaregion in social cohesion (I_3 = 0,58) – only NMR and Lisbon (LIS; I_3 = 0,53) have lower values of this indicator (Table 21).

The scenarios S_1 *–Economic Development*, S_2 *–Social Sustainability*, and S_3 *–Environmental Sustainability* at megaregional level (Table 22) confirm a general values increase in the period of analysis (1995-2010). In 2010, the results show higher economic development and social cohesion in PAR (S_1 = 0,88; S_2 = 0,86) and FRG (S_1 = 0,85; S_2 = 0,87); and more environmental sustainability in BER (S_3 = 0,87), MAD (S_3 = 0,84) and PAR (S_3 = 0,83). No-megaregions (NMR) show the lowest values in economic development and social sustainability (S_1 = 0,17; S_2 = 0,38), but LIS the lowest value in environmental sustainability (S_3 = 0,49).

BAL is the 9th megaregion in economic development (S_1 = 0,46), the 11th megaregion in social sustainability (S_2 = 0,59), and the 10th in environmental sustainability (S_3 = 0,63). The scenarios calculation (Table 22) include the penalization of different factors (as described in Table 20) and, in our opinion, are more precise measures than the single indicators.

The scenarios $S_{4.1}$ *–Inclusive Growth* (equal weigh), and $S_{4.2}$ *–Inclusive Growth* (balanced) calculated at megaregional level (Table 23) in the period of analysis (1995-2010), show higher values in the megaregions PAR ($S_{4.1}$ = 0,86; $S_{4.2}$ = 0,85) and FRG ($S_{4.1}$ = 0,84; $S_{4.2}$ = 0,82); and lower values in no-megaregions –NMR ($S_{4.1}$ = 0,33; $S_{4.2}$ = 0,26). There is also a general increase of the scenarios values overtime.

In 2010, BAL is the 10th ($S_{4.1}$ = 0,55) or 9th ($S_{4.2}$ = 0,51) megaregion in terms of inclusive growth (coming from $S_{4.1}$ = 0,34 and $S_{4.2}$ = 0,27 in 1995). Although $S_{4.1}$ and $S_{4.2}$ present similar results (lower values in $S_{4.2}$ –due to the fact that in this scenario the imbalance is penalized) (Table 23), we consider $S_{4.2}$ as the better approximation to the "urban network sustainable progress", according to the principle of "inclusive growth", that is de development of urban systems towards social equality, ecological efficiency and economic competiveness.

Year	Identification 1	I_1	Identification 2	I_2	Identification 3	I_3
1995	NMR 1995	0.0480	NMR 1995	0.3392	NMR 1995	0.0186
	VIB 1995	0.1359	VIB 1995	0.6180	VIB 1995	0.3317
	FRG 1995	0.2368	FRG 1995	0.4063	FRG 1995	0.2969
	AMB 1995	0.5416	AMB 1995	0.2068	AMB 1995	0.0075
	PRA 1995	0.3267	PRA 1995	0.5539	PRA 1995	0.8172
	LIS 1995	0.2168	LIS 1995	0.4530	LIS 1995	0.1485
	MAD 1995	0.4525	MAD 1995	0.8269	MAD 1995	0.2733
	BAL 1995	0.2661	BAL 1995	0.5573	BAL 1995	0.1843
	PAR 1995	0.7643	PAR 1995	0.8448	PAR 1995	0.6494
	LON 1995	0.4091	LON 1995	0.5249	LON 1995	0.4611
	GLB 1995	0.1447	GLB 1995	0.5859	GLB 1995	0.4552
2000	NMR 2000	0.0798	NMR 2000	0.4798	NMR 2000	0.2050
	VIB 2000	0.3076	VIB 2000	0.4675	VIB 2000	0.1272
	FRG 2000	0.8535	FRG 2000	0.8387	FRG 2000	0.8360
	AMB 2000	0.7430	AMB 2000	0.6368	AMB 2000	0.4413
	PRA 2000	0.5293	PRA 2000	0.6076	PRA 2000	0.7796
	BER 2000	0.5253	BER 2000	0.9246	BER 2000	0.6960
	LIS 2000	0.2381	LIS 2000	0.4471	LIS 2000	0.3506
	MAD 2000	0.6926	MAD 2000	0.8592	MAD 2000	0.8025
	BAL 2000	0.4419	BAL 2000	0.6067	BAL 2000	0.5445
	PAR 2000	0.8756	PAR 2000	0.8760	PAR 2000	0.8528
	RMT 2000	0.5566	RMT 2000	0.6385	RMT 2000	0.5140
	LON 2000	0.6042	LON 2000	0.5724	LON 2000	0.6791
	GLB 2000	0.2168	GLB 2000	0.5573	GLB 2000	0.5652
2005	NMR 2005	0.0940	NMR 2005	0.5176	NMR 2005	0.2705
	VIB 2005	0.3871	VIB 2005	0.5671	VIB 2005	0.2501
	FRG 2005	0.8937	FRG 2005	0.8238	FRG 2005	0.8476
	AMB 2005	0.7679	AMB 2005	0.7015	AMB 2005	0.7491
	PRA 2005	0.6029	PRA 2005	0.6038	PRA 2005	0.7933
	BER 2005	0.3752	BER 2005	0.8784	BER 2005	0.5515
	LIS 2005	0.2911	LIS 2005	0.4687	LIS 2005	0.5622
	MAD 2005	0.4337	MAD 2005	0.8368	MAD 2005	0.8617
	BAL 2005	0.5042	BAL 2005	0.6545	BAL 2005	0.6599
	PAR 2005	0.8585	PAR 2005	0.8381	PAR 2005	0.8344
	RMT 2005	0.5870	RMT 2005	0.6487	RMT 2005	0.6050
	LON 2005	0.5409	LON 2005	0.6593	LON 2005	0.7170
	GLB 2005	0.2937	GLB 2005	0.6453	GLB 2005	0.7387
2010	NMR 2010	0.1045	NMR 2010	0.5449	NMR 2010	0.3478
	VIB 2010	0.3244	VIB 2010	0.6960	VIB 2010	0.6626
	FRG 2010	0.8507	FRG 2010	0.7729	FRG 2010	0.8882
	AMB 2010	0.7584	AMB 2010	0.6838	AMB 2010	0.8124
	PRA 2010	0.5983	PRA 2010	0.5714	PRA 2010	0.8056
	BER 2010	0.4419	BER 2010	0.8936	BER 2010	0.7534
	LIS 2010	0.3864	LIS 2010	0.4848	LIS 2010	0.5324
	MAD 2010	0.4859	MAD 2010	0.8445	MAD 2010	0.8304
	BAL 2010	0.4162	BAL 2010	0.6470	BAL 2010	0.5783
	PAR 2010	0.8868	PAR 2010	0.8206	PAR 2010	0.8710
	RMT 2010	0.4933	RMT 2010	0.6932	RMT 2010	0.5977
	LON 2010	0.4119	LON 2010	0.6585	LON 2010	0.6397
	GLB 2010	0.2312	GLB 2010	0.6462	GLB 2010	0.6761

Table 21 Indicators I_1 *–Economic Growth*, I_2 *–Urban Ecology*, and I_3 *–Social Cohesion* at megaregional level (1995-2010)

Year	Identification 1	S_1	Identification 2	S_2	Identification 3	S_3
1995	NMR 1995	0.0742	NMR 1995	0.0827	NMR 1995	0.2751
	VIB 1995	0.2037	VIB 1995	0.3890	VIB 1995	0.5607
	FRG 1995	0.2598	FRG 1995	0.3188	FRG 1995	0.3844
	AMB 1995	0.4547	AMB 1995	0.0474	AMB 1995	0.1670
	PRA 1995	0.3985	PRA 1995	0.7645	PRA 1995	0.6066
	LIS 1995	0.2336	LIS 1995	0.2094	LIS 1995	0.3921
	MAD 1995	0.4721	MAD 1995	0.3841	MAD 1995	0.7162
	BAL 1995	0.2870	BAL 1995	0.2589	BAL 1995	0.4827
	PAR 1995	0.7609	PAR 1995	0.6885	PAR 1995	0.8058
	LON 1995	0.4259	LON 1995	0.4738	LON 1995	0.5121
	GLB 1995	0.2198	GLB 1995	0.4814	GLB 1995	0.5598
2000	NMR 2000	0.1323	NMR 2000	0.2600	NMR 2000	0.4248
	VIB 2000	0.3056	VIB 2000	0.1953	VIB 2000	0.3994
	FRG 2000	0.8503	FRG 2000	0.8366	FRG 2000	0.8382
	AMB 2000	0.7022	AMB 2000	0.4804	AMB 2000	0.5977
	PRA 2000	0.5621	PRA 2000	0.7452	PRA 2000	0.6420
	BER 2000	0.5823	BER 2000	0.7417	BER 2000	0.8789
	LIS 2000	0.2703	LIS 2000	0.3699	LIS 2000	0.4278
	MAD 2000	0.7203	MAD 2000	0.8138	MAD 2000	0.8478
	BAL 2000	0.4686	BAL 2000	0.5569	BAL 2000	0.5942
	PAR 2000	0.8734	PAR 2000	0.8575	PAR 2000	0.8713
	RMT 2000	0.5605	RMT 2000	0.5389	RMT 2000	0.6136
	LON 2000	0.6085	LON 2000	0.6578	LON 2000	0.5938
	GLB 2000	0.2857	GLB 2000	0.5636	GLB 2000	0.5588
2005	NMR 2005	0.1540	NMR 2005	0.3199	NMR 2005	0.4682
	VIB 2005	0.3914	VIB 2005	0.3135	VIB 2005	0.5037
	FRG 2005	0.8821	FRG 2005	0.8429	FRG 2005	0.8285
	AMB 2005	0.7594	AMB 2005	0.7396	AMB 2005	0.7110
	PRA 2005	0.6220	PRA 2005	0.7554	PRA 2005	0.6417
	BER 2005	0.4432	BER 2005	0.6169	BER 2005	0.8130
	LIS 2005	0.3360	LIS 2005	0.5435	LIS 2005	0.4874
	MAD 2005	0.5168	MAD 2005	0.8567	MAD 2005	0.8418
	BAL 2005	0.5348	BAL 2005	0.6588	BAL 2005	0.6555
	PAR 2005	0.8541	PAR 2005	0.8351	PAR 2005	0.8374
	RMT 2005	0.5950	RMT 2005	0.6137	RMT 2005	0.6399
	LON 2005	0.5703	LON 2005	0.7054	LON 2005	0.6708
	GLB 2005	0.3734	GLB 2005	0.7200	GLB 2005	0.6640
2010	NMR 2010	0.1729	NMR 2010	0.3872	NMR 2010	0.5055
	VIB 2010	0.3954	VIB 2010	0.6692	VIB 2010	0.6893
	FRG 2010	0.8467	FRG 2010	0.8651	FRG 2010	0.7959
	AMB 2010	0.7564	AMB 2010	0.7867	AMB 2010	0.7095
	PRA 2010	0.6164	PRA 2010	0.7588	PRA 2010	0.6182
	BER 2010	0.5182	BER 2010	0.7814	BER 2010	0.8656
	LIS 2010	0.4109	LIS 2010	0.5229	LIS 2010	0.4943
	MAD 2010	0.5562	MAD 2010	0.8332	MAD 2010	0.8417
	BAL 2010	0.4555	BAL 2010	0.5920	BAL 2010	0.6333
	PAR 2010	0.8786	PAR 2010	0.8609	PAR 2010	0.8307
	RMT 2010	0.5237	RMT 2010	0.6168	RMT 2010	0.6741
	LON 2010	0.4594	LON 2010	0.6434	LON 2010	0.6547
	GLB 2010	0.3172	GLB 2010	0.6701	GLB 2010	0.6522

Table 22 Scenarios S_1 *–Economic Development*, S_2 *–Social Sustainability*, and S_3 *–Environmental Sustainability* at megaregional level (1995-2010)

Year	Identification 1	$S_{4.1}$	Identification 2	$S_{4.2}$
1995	NMR 1995	0.1353	NMR 1995	0.0818
	VIB 1995	0.3619	VIB 1995	0.2815
	FRG 1995	0.3133	FRG 1995	0.2851
	AMB 1995	0.2520	AMB 1995	0.1630
	PRA 1995	0.5659	PRA 1995	0.4842
	LIS 1995	0.2728	LIS 1995	0.2220
	MAD 1995	0.5176	MAD 1995	0.4253
	BAL 1995	0.3359	BAL 1995	0.2737
	PAR 1995	0.7528	PAR 1995	0.7203
	LON 1995	0.4650	LON 1995	0.4457
	GLB 1995	0.3953	GLB 1995	0.3217
2000	NMR 2000	0.2549	NMR 2000	0.1882
	VIB 2000	0.3008	VIB 2000	0.2441
	FRG 2000	0.8428	FRG 2000	0.8398
	AMB 2000	0.6070	AMB 2000	0.5567
	PRA 2000	0.6388	PRA 2000	0.5971
	BER 2000	0.7153	BER 2000	0.6488
	LIS 2000	0.3453	LIS 2000	0.3104
	MAD 2000	0.7848	MAD 2000	0.7570
	BAL 2000	0.5310	BAL 2000	0.5035
	PAR 2000	0.8681	PAR 2000	0.8643
	RMT 2000	0.5697	RMT 2000	0.5489
	LON 2000	0.6186	LON 2000	0.6008
	GLB 2000	0.4464	GLB 2000	0.3883
2005	NMR 2005	0.2940	NMR 2005	0.2234
	VIB 2005	0.4014	VIB 2005	0.3486
	FRG 2005	0.8550	FRG 2005	0.8434
	AMB 2005	0.7395	AMB 2005	0.7284
	PRA 2005	0.6667	PRA 2005	0.6349
	BER 2005	0.6017	BER 2005	0.5178
	LIS 2005	0.4406	LIS 2005	0.3955
	MAD 2005	0.7107	MAD 2005	0.6394
	BAL 2005	0.6062	BAL 2005	0.5802
	PAR 2005	0.8437	PAR 2005	0.8396
	RMT 2005	0.6136	RMT 2005	0.6033
	LON 2005	0.6390	LON 2005	0.6097
	GLB 2005	0.5592	GLB 2005	0.4851
2010	NMR 2010	0.3324	NMR 2010	0.2590
	VIB 2010	0.5610	VIB 2010	0.4991
	FRG 2010	0.8373	FRG 2010	0.8181
	AMB 2010	0.7515	AMB 2010	0.7301
	PRA 2010	0.6584	PRA 2010	0.6194
	BER 2010	0.6963	BER 2010	0.6210
	LIS 2010	0.4679	LIS 2010	0.4435
	MAD 2010	0.7203	MAD 2010	0.6605
	BAL 2010	0.5472	BAL 2010	0.5087
	PAR 2010	0.8595	PAR 2010	0.8484
	RMT 2010	0.5947	RMT 2010	0.5614
	LON 2010	0.5700	LON 2010	0.5289
	GLB 2010	0.5178	GLB 2010	0.4437

Table 23 Scenarios $S_{4.1}$ *–Inclusive Growth* (equal weigh), and $S_{4.2}$ *–Inclusive Growth* (balanced) at megaregional level (1995-2010)

Conclusions

General remarks

Over the last two centuries, the boundaries of the city have been constantly redefined. Trullén et al. (2013) explain that the real force behind the city's change of scale has been the liberating effect of so called "spatially mobile external economies" which are not constrained to a single place by agglomeration forces. This driving force is able to create what Lang and Nelson (2009) call "large-scale trans-metropolitan urban structures", such as urban regions and megaregions.

The development of these urban networks is cause and consequence of the densification and acceleration of socioeconomic processes, resulting in increasing levels of complexity. From an economic point of view, the megaregion scale of organization appears to be accelerating global change (Grazi et al., 2008), concentrating a huge amount of world production and innovation, and is associated with higher levels of per capita income and creativity (Florida et al.,2008; Ross, 2009; Marull et al., 2013). However, an issue that has received less attention in the literature (exceptions are Wheeler, 2009 and Campbell, 2009) is that once formed megaregions also become efficient in resource consumption and promote well-being (Marull and Boix, 2017).

The question we raise is whether, once formed, the subsequent dynamics of urban networks are sustainable or not. A positive expectation of inclusive growth of existing megaregions (integrating social, economic and ecological dimensions) is a reason to facilitate the conditions for the formation of new ones. On the other hand, evidence that existing megaregions are evolving towards positions of reduced sustainability provides arguments for preventing new ones being formed, while for existing urban networks although it could be difficult to dissolve them there could nevertheless be attempts to manage them through pro-active policy.

Are the dynamics of urban networks sustainable? We explore the hypothesis that increasing complexity in regions and megaregions implies less demand on resources needed to generate organized information and social cohesion, thereby making the urban systems more efficient and stable. This study proposes new structural indices for measuring sustainable urban network progress according to different conceptual scenarios, at the regional and megaregional scale.

Proposed model

We use night-time light (NLT) data from the broadband near-visible infrared channel of the DMSP-OLS satellite sensor to monitor the dynamics of urbanization. We propose four indices for sustainable progress of networks of cities, according with the following conceptual scenarios: S_1 –*Economic development* (mainstream economic model), S_2 –*Social sustainability* (based on social equality), S_3 -*Environmental sustainability* (based on resource consumption) and S_4 – *Inclusive growth* (with two variations: $S_{4.1}$: equal factor weigh –economic, social and ecological; and $S_{4.1}$: balanced). These indices are derived from the integration of different indicators obtained by structural equation models.

The statistics used are component analysis, factor analysis, cluster analysis, structural analysis, and a probabilistic method for the indices development. The models and indices approximate the problem raised over the necessity of a standard way to determine the sustainable progress within a given urban region, beyond the GDP. It has accomplished the goal of providing tools to undercover the hidden factors of interest, to model its relationships as a measure of urban system progress. In the study, we apply the integration of economic, social and ecological indicators into indices (according to four conceptual scenarios) to the entire European NUTS 3 regions and the 12 existing megaregions from 1995 to 2010 –period with available satellite data.

Preliminary results suggest the index $S_{4.2}$ as the better measure of the urban network sustainable progress, according to the principle of inclusive growth, which is the balanced development of urban systems towards social equality, ecological efficiency and economic competiveness.

The results also prove the decoupling of economic growth and the social and environmental development by showing regions that have improved at it without an economic growth above average. The main conclusion is that regional and megaregional urban systems respond to increasing complexity by adapting their relational structures to become more efficient and stable, and become more sustainable forms of organization. Consequently, it could be necessary to re-direct land use policies towards improving sustainability at the level of the megaregion.

Policy implications

The functioning of agglomeration economies, in the form of urbanization economies and network economies, and the transformation of the economic model towards a knowledge-based economy, allows rising levels of GDPpc while, at the same time, reduce energy intensity –lowering impact in terms of entropy, and increase urban organized information –promoting social equality. The experience of the best performing megaregions shows that this is possible. It follows that the change in the economic model, with increasing importance of agglomeration economies where knowledge becomes the key productive factor, should be the driver of change in the sustainable urban progress.

It is known that as a system becomes more complex it reduces its dependence on energy but increases its organized information and knowledge. By analogy with systems of cities, it means that in the future, urban competitiveness will rely on economic models based more on knowledge than on consumption of resources. Trying to change the economic model without considering the role played by agglomeration economies, could have unexpected negative impacts and cause discontinuities in economic but also in social and environmental terms.

The results obtained in this paper can also be interpreted as a contribution to the dialogue between two lines of research, the ecological economy and the urban ecology, at a new spatial scale to explore sustainability: the emerging megaregions. Moreover, our results could be put in relation with the Strategy Europe 2020, which aims at a “smart, sustainable and inclusive growth”, highlighting the role that urban networks can play in achieving these objectives.

Further research

In developed countries, the existing economic model emphasizes the GDP growth, which implies social and environmental instability. It also shades developing countries to alternatives for more sustainable models of urban progress. Development based on economic growth has down natural resources. Much of the generated wealth has been unequally distributed (Wilkinson and Pickett, 2009). New research should analyze the contribution of urban networks to the solution of the prevailing economic model crisis and its social and ecological consequences.

There is a claim to defeat GDP as the official measure of development. The UN announced the Sustainable Development Goals, a set of international objectives to improve global well-being. Developing integrated measures of urban sustainable progress attached to these goals could offer the opportunity to define what well-being means, how to measure and achieve it. Lasting the present economic model would tolerate the increase of social inequality and would ignore the continued destruction of the natural capital. The current recognition that GDP is an inadequate measure of sustainable progress (Constanza et al., 2014), suggests to admit the

complexity of the interaction between social, ecological and economic factors in global urban systems.

The main objective of the first phase of this line of research has been to create integrated indices of urban network sustainable progress according to four conceptual scenarios, able to measure the performance and dynamics of urban systems at regional and megaregional levels, based on official (Eurostat) and satellite (NASA) data, and given standardized socio-economic-ecological factors. However, it is necessary to improve and validate the methodology, and to apply the models and indices at metropolitan level as well.

Improvements in the scenarios formulation could be made. Larger data sets with larger periods should be studied, and an easy way to relate different scales of observations in terms of region size have to be developed to truly understand the conceptual variables of interest at different levels. More variables and factors, as long as they comply with the conceptual framework, must be analyzed to account for the variance that was not explained in the three-factor model. With the availability of larger periods, time series analysis could be implemented to create an autoregressive model to predict the variation of the variables. Last, an aspect that was not studied was the causality between variables.

Finally, it would be interesting to model the temporal variation of night-time light (NTL) intensities obtained from the satellite database. A better understanding of this NTL territorial matrix variation of a given area (i.e. using cellular automata modeling), and the possibility of making useful predictions about urban development scenarios related to social-economic-ecological models, could be applied in regional planning and land use policies.

Bibliography

Batabyal, A., Nijkamp, P., 2009. Sustainable development and regional growth. In: Capello, R., Nijkamp, P. (Eds.), Handbook of Regional Growth and Development Theories. Edward Elgar, Cheltenham, UK/Northampton, MA, pp. 282–301.

Borowy, I., 2014. Defining Sustainable Development: the World Commission on Environment and Development. Brundtland Commission. Routledge.

Box, G.E.P., Cox, D.R., 1964. An analysis of transformations. Journal of the Royal Statistical Society Series B 26 (2), 211–252.

Breheny, M., 1992. Sustainable development and urban form. In: European Research in Regional Science. Pion Limited, London.

Camagni, R., 2005. Economia Urbana. Antoni Bosch Editor, Barcelona.

Campbell, S., 2009. Mega-regions and sustainability. In: Ross, C. (Ed.), Megaregions: Planning for Global Competitiveness. Island Press, pp. 127–139.

Christaller, W., 1933. Die Zentralen Orte in Süddeutschlandland (Central Places in Western Germany)., pp. 1968.

Costanza, R., Kubiszewski, I., Giovannini, E., Lovins, H., McGlade, J., Pickett, K.E., Ragnarsdóttir, K.V., Roberts, D., De Vogli, R., Wilkinson, R., 2014. Time to leave GDP behind. Nature, 283-285.

Costanza, R., Hart, M., Posner, S. & Talberth, J., 2009. Beyond GDP: The Need for New Measures of Progress. Boston University, 2009.

Dewar, M., Epstein, D., 2007. Planning for megaregions in the United States. Journal of Planning Literature 22, 118–124.

Doll, C., 2008. Thematic Guide to Night-time Light Remote Sensing and its Applications. Centre for International Earth Science Information Network (CIESIN), Columbia University, New York.

Elkington, J., 1999. Cannibals with forks : the triple bottom line of 21 century business. Oxford.

Florida, R., Gulden, T., Mellander, C., 2007. The Rise of the Mega Region. J.L. Rotman School of Management, University of Toronto, The Martin Prosperity Institute.

Griggs, D. et al. 2013. Nature 495, 305–307.

Grazi, F., van den Bergh, J.C., van Ommeren, J.N., 2008. An empirical analysis of urban form, transport, and global warming. Energy Journal 29 (4), 97–122.

Hasmath, R., (ed.) 2015. Inclusive Growth, Development and Welfare Policy: A Critical Assessment. Routledge.

Kubiszewski, I. et al. 2013. Ecol. Econ. 93, 57–68.

Kuznets, S., 1934. National Income, 1929–1932. 73rd US Congress, 2d session, Senate document no. 124, 5-7.

Lang, R.E., Dhavale, D., 2005. Beyond Megalopolis: Exploring America's New Megapolitan Geography. Metropolitan Institute at Virginia Tech, Census Report Series.

Lang, R.E., Nelson, A.C., 2009. Defining megapolitan regions. In: Ross, C.B., Contant,C. (Eds.), Megaregions in America. Island Press, Washington, DC.

Marull, J., Galletto, V., Domene, E., Trullén, J., 2013. Emerging megaregions: a new spatial scale to explore urban sustainability. Land Use Policy 34, 353–366.

Marull, J., Font, C., Boix, R., 2015. Modelling urban networks at mega-regional scale: Are increasingly complex urban systems sustainable? Land Use Policy 43, 15–27.

Marull, J., Boix, R. (eds.), 2016. Megaregions i desenvolupament urbà sostenible. Factors estratègics per a l'àrea metropolitana de Barcelona en el context europeu. Papers 58, 164 pp.

OECD, 2008. Handbook on constructing composite indicators: Methodology and user guide. http://www.oecd.org/std/42495745.pdf.

Pred, A., 1977. City-systems in Advanced Economies. Hutchinson, London.

Ross, C.L., 2009. Megaregions: Planning for Global Competitiveness. Island Press, Washington.

Rubin, D.B., 1987. Multiple Imputation for Nonresponse in Surveys. John Wiley and Sons.

Seligman, M.E.P., 2012. Flourish: A Visionary New Understanding of Happiness and Well-being.

Atria.

Small, C., Pozzi, F., Elvidge, C.D., 2005. Spatial analysis of global urban extent from DMSP-OLS night lights. Remote Sensing of Environment 96 (3/4), 277–291.

Stiglitz, J.E., Sen, A., Fitoussi, J.-P., 2009. Report by the Commission on the Measurement of Economic Performance and Social Progress Vol. 12.

Thomson, G.H., 1951. The Factorial Analysis of Human Ability. London University Press.

Trullén, J., Boix, R., Galletto, V., 2013. An insight on the unit of analysis in urban research. In: Kresl, P.K., Sobrino, J. (Eds.), Handbook of Research Methods and Applications in Urban Economies. Edward Elgar, Northampton, MA, pp.235–264.

Wheeler, S., 2009. Regions, megaregions, and sustainability. Reg. Stud. 43, 863–876.

Van den Bergh, J. C. J. M. J., 2009. Econ. Psychol. 30, 117–135.

Williams, K., Burton, E., Jenks, M., 2000. Achieving Sustainable Urban Form. E & FN Spon, London.

Wilkinson, R.G., Pickett, K., 2009. The Spirit Level: Why Greater Equality Makes Societies Stronger. Bloomsbury.

Zhang, Q., Seto, K.C., 2011. Mapping urbanization dynamics at regional and global scales using multi-temporal DMSP/OLS nighttime light data. Remote Sensing of Environment 115, 2320–2329.

Tables

Figures