REGION
The Journal of ERSA
Powered by WU

Volume 3, Number 2, 2016, 23–45
DOI: 10.18335/region.v3i2.115

journal homepage: region.ersa.org

Tree-based approaches for understanding growth patterns in the European regions*

Paola Annoni[1], Angél Catalina Rubianes[2]

[1] European Commission, Economic Analysis Unit, Directorate General for Regional and Urban Policy, Brussels, Belgium (email: paola.annoni@ec.europa.eu)
[2] European Commission, Economic Analysis Unit, Directorate General for Regional and Urban Policy, Brussels, Belgium (email: angel.catalina-rubianes@ec.europa.eu)

Received: 8 January 2016/Accepted: 21 July 2016

Abstract. We run an empirical analysis to understand the main drivers of economic growth in the European Union (EU) regions in the past decade. The analysis maintains the traditional factors of growth used in the literature on regional growth – stage of development, population agglomeration, transport infrastructure, human capital, labor market and research and innovation – and incorporates the institutional quality and two variables which reflect the macroeconomic conditions in which the regions operate. Given the scarcity of reliable and comparable regional data at the EU level, the starting point of the analysis was devoted to build reliable and consistent panel data on potential factors of growth. Two non-parametric, decision-tree techniques, randomized Classification and Regression Tree and Multivariate Adaptive Regression Splines, are employed for their ability to address data complexities such as non-linearity and interaction effects, which are generally a challenge for more traditional statistical procedures such as linear regression. Results show that the dependence of growth rates on the factors included is clearly non-linear with important factor interactions. This means that growth is determined by the simultaneous presence of multiple stimulus factors rather than the presence of a single area of excellence. Results also confirm the critical importance of the macroeconomic framework together with human capital as major drivers of economic growth. This is overall in line with most of the economic literature, which has persistently underlined the major role of these factors on economic growth but with the novelty that the macroeconomic conditions are here incorporated. Human capital also has an important role, with low-skilled labor supply having a higher detrimental effect than the conducive one of high-skilled labor supply. Other important factors are the quality of governance for most developed economies and, in line with the neoclassical growth theory, the stage of development in particular for less developed economies. The evidence given by the model about the impact of other factors on economic growth such as those on the quality of infrastructure or the level of innovation is more limited and inconclusive.
The analysis conclusions support the reinforcement of the EU economic governance and the conditionality mechanisms set in the new architecture of the EU regional funds

*The authors are grateful to the members of the Economic Analysis Unit of the Directorate General for Regional and Urban Policy, European Commission, for scientific guidance and helpful discussion throughout the analysis. A special thank goes to Beatriz Torighelli, for her essential input in data collection and preliminary data manipulation, and Marina Mastrostefano, who provided insight and expertise that greatly assisted the research and the interpretation of results. The information and views set out in the publication are those of the authors and do not reflect the official opinion of the European Commission.

2014-2020 whose rationale is that the effectiveness of the expenditure is conditional to good institutional quality and sound economic policies.

Key words: Regional economic growth, European Union regions, non-parametric statistics, decision trees, multivariate adaptive regression splines.

1 Introduction

Understanding what triggers economic growth is difficult and controversial (Cristelli et al. 2015). Factors of growth form a complex system where identifying causality is not trivial as contradictions arise where variables are positively associated at some times but appear unrelated or even negatively associated depending on the system state. Such state-dependent behavior is an earmark of complex nonlinear systems and creates problems when fitting models to observational data (Sugihara et al. 2012). Proper methodologies are in need.

The focus of the paper is to identify the main determinants of economic growth in the European Union (EU) regions in the most recent years. The analysis covers the 2003-2013 period and includes as many indicators as possible at the regional, NUTS2 level[1] for all the regions of the 28 EU Member States. The interest is in the NUTS2 geographical level as being the main territorial level for the application of EU regional policies. Socio-economic data availability at the regional level in the EU is unfortunately limited. Two reasons determined the time period spanned by the analysis: first, the availability of reliable and comparable data at the regional level and, second, the inclusion of all the EU28 Member States. Given that the last, big EU enlargement took place in 2004 and that the set of explanatory indicators are one year time lagged with respect to the dependent, we set the starting year at 2003. The ending year is the one with the most recent data on regional GVA growth rates at the time of the analysis.

A variety of empirical studies are available in the econometric literature which explore the effectiveness of European and national policies in stimulating economic growth. They are mostly based on linear regression with growth rate as the response and a set of (sometimes non-linearly transformed) explanatory factors which can include interactions and/or spatial effects. Some relevant examples are discussed in (Rodríguez-Pose, Fratesi 2004, Dall'erba, Le Gallo 2008, Ramajo et al. 2008, Rodríguez-Pose, Garcilazo 2013). These approaches are all model-based with strong underlying assumptions. From the methodological point of view, our analysis differs from most of the studies on economic growth by being non-parametric: we employ data-driven approaches that learn nonlinearities and interactions directly from the data without the need to explicitly model them (Grömping 2009). Two non-parametric statistical methods, Classification and Regression Trees – CART – and Multivariate Adaptive Regression Splines – MARS, are employed in a complementary way. A huge literature is available on the two techniques that have been introduced respectively by the seminal works of (Breiman et al. 1984) and (Friedman 1991). Both techniques belong to the wide family of decision tree techniques, with MARS being the evolution of CART. In the past two decades or so, CART and MARS have been used in a wide range of applications from astronomy (Weir et al. 1995) to biology (Leathwick et al. 2005), from finance (Mezrich 1994) to medicine (Austin 2007), showing their versatility and usefulness. To our knowledge, though, the employment of decision-tree techniques to econometrics and, in particular to the study of economic growth is very limited and specific. An early contribution in this vein uses CART to explore nonlinearities in the process of cross-country output growth (Durlauf, Johnson 1995), while a more recent example can be found in (Curtis, Kokotos 2009) for tourism-based regional development. Varian (2014) recently advocated the use of these techniques as they 'may allow for more effective ways to model complex relationships'. This is exactly why we use them.

[1] The NUTS classification (Nomenclature des Unités Territoriales Statistiques) is a hierarchical system that the European Statistical Office – EUROSTAT – employs for dividing the economic territory of the EU for the collection, development and harmonization of European regional statistics, for socio-economic analyses of the regions and for framing EU regional policies.

The paper is structured as follows: Section 2 provides the description of the two non-parametric techniques employed to identify the main determinants of regional growth while the theoretical framework and the data used is detailed in Section 3. Results are discussed in Section 4 and Section 5 summarizes main results in the European Union context.

2 Methods

For understanding the main determinants of growth we want to capture all the possible non-linearities and interactions that may be, and usually are, present in datasets of this kind. To this aim two non-parametric, data-driven techniques, randomized CART and MARS, are employed.

CART and MARS are tree-based regression techniques which solve the problem of fitting a response y to a set of predictors (explanatory factors) $\mathbf{x} \in \mathcal{R}^n$ from observed data in high dimensions. Methods based on polynomial approximation are generally unsuccessful due to the instability of the polynomial for high n. Methods which locally approximate y in a neighborhood of a point $\mathbf{x}$ are also unsuccessful in high dimensions due to the 'curse of dimensionality', that is too many observations are necessary to get a reliable approximation of the response y (De Veaux et al. 1993). CART and its successor MARS have been designed to overcome these limitations. Their common starting point is a tree-based regression.

2.1 Randomized CART

CARTs are classification-type techniques where the dependent y and the factors $\mathbf{x}$ can be categorical, either nominal or ordinal, or continuous (Breiman et al. 1984, Hastie et al. 2001). Many statistical methods are available for analyzing classification-type problems. Regression approaches, such as logistic regression (Agresti 1990), or classification approaches, such as linear discriminant analysis (Mardia, Kent 1979), play an important role but are both linear and parametric. Although attractively simple, traditional parametric linear models may fail in empirical analysis as in real life effects are most often non linear and highly interactive (Hastie et al. 2001).

Non parametric methods are more flexible and provide a powerful approximation for any type of relationships among any type of variables. Being non-parametric, they do not rely on any explicit or implicit assumptions on the data structure, meaning that linearity or monotonicity of the relationship between the dependent variable and factors are not required. This makes tree-based techniques particularly suitable for empirical analyses as non-linearity is ubiquitous in the socio-economic science.

CART is the most popular tree-based method. It is a stepwise top-down algorithm performing binary splits of the starting population P into smaller and smaller sub-populations less impure than the parent population. Impurity levels are based on the dependent variable y whose type determines the impurity measure to be adopted. The Gini index, G, is used here: G is a non-negative real number with 0 representing perfect purity (Breiman et al. 1984). It is easy to compute, more sensitive to changes in node probabilities and has a twofold meaning: misclassification rate and node variance. At each step one explanatory factor at a time is analyzed and the one which mostly decreases the impurity level of the parent population is selected. When a certain stopping rule is achieved, terminal sub-populations are called leaves. The starting node comprising the whole population is the root. Other tree algorithms allow for multiway splits instead of binary ones but they are generally not recommended (Hastie et al. 2001).

To avoid overfitting trees should be optimized. This can be done either by fixing the minimum number of observations in each leaf or by using the cost-complexity approach, which is a way to cope with trade-off between tree size (number of leaves) and its goodness of fit (Breiman et al. 1984). The choice is here to set the minimum leaf size because we are more interested in having non-irrelevant terminal sub-populations rather than having the best possible tree. Different minimum leaf sizes have been tested with robust results.

After one (optimal) tree is set-up, a class of the response category is assigned to each leaf according to the so called plurality rule: each node is classified into one class k of

the categorical response y ($k = 1, \ldots K$) if k the most frequent class of that node. The goodness of fit of the tree can be measured via the misclassification rate (MR) that can be estimated in various ways. The most common one is the re-substitution estimate of MR, defined as the proportion of misclassified cases (Breiman et al. 1984). If the data set is large enough, the validation process can be undertaken with either split-sample or cross-validation misclassification rates. In the split-sample case, the tree is generated using a training sample and the misclassification rate is computed on the remaining sample, called the validation sample. In cross-validation, the sample is divided into a number of sub-samples, usually 10-20, and the tree is generated excluding one sub-sample at a time. The split-sample approach is employed here to estimate MR for each tree, with the training set including 90% of the sample points, randomly selected (the remaining points are included in the validation set).

In general results from different randomized sample splits differ from each other. In order to get stable results, 1000 different 90%/10% randomized splits are computed and average results together with their estimated confidence intervals are provided.

Explanatory factors are ranked according on their use in the final tree. Different metrics can be used, from the simple count of the number of times a factor is used for splitting to more sophisticated measures based on the purity level improvements in all the splits where the factor is used. Apart from the count of splits, which is a naive variable importance measure, the other options are almost equivalent unless the set of predictors includes interval-scale and categorical variables or when categorical predictors are measured with different number of categories (Strobl et al. 2007). The sum of squared errors (SSE) is here used as variable importance measure (SAS 2014) and, given that predictors are all measured on an interval scale, this choice is not going to substantially affect the results. The SSE for a classification tree is defined as the sum across all the leaves Λ of the square of the number of misclassified observations in each leaf λ_i, $i = 1, \ldots \Lambda$:

$$SSE = \sum_{\lambda=1}^{\Lambda} \sum_{k=1}^{K} \left(N_\lambda - N_\lambda^k\right)^2 \tag{1}$$

where N_λ is the number of observations in leaf λ predicted to be in category k and N_λ^k is the number of observations in λ in category k.

The SSE-based importance of factor x_i, $IMP_{(SSE)}(x_i)$, normalized with respect to the maximum factor importance is then defined as:

$$IMP_{SSE}(x_i) = \frac{\left(\sum_m^M \Delta_m\right)^{0.5}}{\max_{x_i}\left\{IMP_{(SSE)}(x_i)\right\}} \tag{2}$$

where M is the total number of nodes in the tree and Δ_m is the change in SSE at node m.

CART is not without limitations. First, it has difficulties in describing smoothly varying responses. Indeed, the sharp nature of the splits in CART generates discontinuities at the edge of each data sub-region produced by the split. Secondly, the splits are all dependent from another split. This induces a model with high-order interactions among predictors and makes it difficult to interpret results. Thirdly, large sample sizes, typically of the order of magnitude of 10^3, are usually needed to provide stable results[2].

2.2 Multivariate Adaptive Regression Splines

MARS is a successor of CART (Friedman 1991). Similarly to CART, it is non-linear and 'almost' non-parametric. MARS cannot be considered completely non-parametric for at least two reasons: it transforms predictors with specific functional forms (the 'basis') and estimates the model with an ordinary least square regression of the transformed predictors. Being empirical and very flexible, it nevertheless preserves many advantages

[2] Sample sizes required to obtain stable results always depend on the complexity of the relationships to be uncovered. Please consider this indication just as a rule of thumb.

of non-parametric techniques (De Veaux et al. 1993). MARS attempts to remedy the limitations of CART relieving the split discontinuity by means of piece-wise linear splitting functions and reducing the interaction order, because subsequent splits are not necessarily dependent on previous splits.

While decision trees use step functions to model the dependent, MARS uses piecewise linear functions, called basis. This makes it more effective in dealing with model non-linearities and smoothly varying responses (De Veaux et al. 1993, Deichmann et al. 2002). By adding the basis, MARS is capable of uncovering non-linear relationships and interaction effects. In the classical MARS model explanatory factors can be either categorical or continuous while the dependent is continuous. Binary variables can be used as response if the model is run in binary mode with the logistic regression MARS.

The algorithm consists of a two-step analysis that firstly builds a collection of functions called 'basis' – Bs – and automatically selects the best regression model based on a selection of basis functions and their interactions. Bs are piece-wise linear transformations of the explanatory factors $\mathbf{x} = x_1, x_2, \ldots, x_n$ and are used to represent the information contained in one or more x_i. Bs are defined as:

$$B(x_i, \tau) = \max\{0, x_i - \tau\} + \max\{0, \tau - x_i\} \tag{3}$$

where τ is an inflection point along the range of a given predictor x_i, the 'knot' of the basis. More than one knot τ may be used for the same predictor. A MARS model is built using a subset of all such possible piecewise linear functions. Smoother curves can be used as well by allowing for higher order terms in the functional form of Bs, like quadratic or cubic terms. The simplest version of MARS using piecewise linear splines is adopted because it keeps the model simple while giving enough degrees of freedom for fitting the data, like for example through higher-order interactions. Products of Bs can be included in the model to account for different order interactions. Contrary to CART, in MARS the maximum number of interactions allowed is a parameter of the model chosen by the analyst. It is then possible to have full control on the model complexity, in terms of number of interactions allowed. Optimal combinations of basis are used to estimate a least-square model with the Bs as new independent variables.

To determine the optimal number of terms in the model the generalized cross-validation GCV criterion is employed. GCV is the average of the squared residuals times a penalty to take into account the model complexity, in terms of number of basis functions included. The algorithm then involves a backward stage which eliminates Bs that unnecessarily complicate the model. Parsimony on the number of basis entering the model has indeed the desirable effects of limiting spurious interaction effects caused by collinearity, an ever present problem when modeling observational data, while facilitating interpretation. To reinforce parsimony, an additional penalty factor γ can be introduced to reduce the model improvement for any new variable that is introduced at each iteration of the model forward selection. Commonly used values for γ lie between 0 (no penalty on complexity) to 0.15 (high penalty on complexity)[3].

The importance of a variable in the model is computed on the basis of GCV. For variable ranking, GCV is computed with and without each variable in the model and the difference is computed. These differences are then normalized into a 0 (not important) to 100 (most important) scale.

An interesting feature of MARS is the possibility to get graphical representation of the modeled relationships between the dependent and the transformed predictors x_i, model components. The contribution to the response of individual explanatory factors can be shown explicitly, enabling local interpretation of the underlying model (De Veaux et al. 1993).

From our perspective, one of the advantages of MARS over CART is that smaller sample sizes are necessary with MARS, generally of the order of 10^2. In the seminal work by Friedman (1991), the accuracy of different MARS models is assessed using sample sizes going from 50, considered as a small sample, to 200. Another important advantage of MARS is the possibility of better understanding the impact of the predictors on the

[3]For more details on MARS technicalities see Friedman (1991).

response variable in terms of order of interactions and type of non-linear relations. The case study on regional growth will help in elucidating these points.

3 Understanding regional economic growth

3.1 Theoretical framework

The theoretical framework of the analysis starts from the Solow-type growth framework to control for the regional initial GVA per capita as a proxy for its initial capital endowment (Solow 1956, Barro, Sala-i-Martin 1992). But this is only the basic model because it assumes that all the regions feature the same structural characteristics, which is clearly an implausible assumption. Other explanatory factors are then included in the model. Following the main literature contributions and data availability across the EU regions, the regional factors included go from human and physical capital to population density, from levels of employment to the quality of institutions (Mankiw et al. 1992, Rodrik et al. 2004, Kwok, Tadesse 2006, Crescenzi, Rodríguez-Pose 2008, Mohl, Hagen 2010, Rodríguez-Pose 2013, Rodríguez-Pose, Garcilazo 2013, Pescatori et al. 2014).

Two macroeconomic variables are also added with the aim of capturing the impact of debt, both public and private, which is currently considered as a major constraint to economic growth. The quick accumulation of both private and public debt in the Member States with the poorest trends in economic development over the last 10-15 years, the significant correlation of regional development trends with national trends in the EU and the low attention paid to these factors in analysis of regional development in the EU explains the inclusion of these factors

3.2 Data

The empirical analysis uses panel data from 2003 to 2013 available or estimated for the EU regions at the NUTS2 level. Please note that the attribute 'regional' is used as synonymous of 'NUTS2 level' hereafter. The selected time period allowed for including the highest possible number of indicators at the regional level and to carry out an analysis of all the regions in the 28 EU Member States. The task is in general particularly demanding due to the scarcity of reliable and comparable data at the EU level and the complex interaction between the factors of growth.

As described in Sections 2.1 and 2.2, the statistical models chosen for the analysis are both non-parametric: they let the data speak without superimposing any assumptions. This means that preliminary data-handling becomes even more than usual an essential ingredient for reliable results.

Limiting the analysis time span to a decade (2003-2013) allowed us to include a relatively rich set of basic indicators from official sources (Eurostat, World Bank, World Economic Forum, Quality of Government Institute).

As the global financial and economic crisis hit almost worldwide in 2008, the period under analysis captures both pre- and post-crisis years. In the EU the crisis emerged in 2008 and unfolded over the following years revealing long-term problems especially in southern countries. The presence of a structural breakpoint in GVA growth has been statistically tested using the Analysis of Variance – ANOVA (Moore 2004) on every two consecutive growth periods (3-year average periods). Significant differences with p-values < 0.001, are found between the periods $2006-2008; 2007-2009$ and $2007-2009; 2008-2010$. That said, the authors think neither that the crisis substantially changed the main drivers of growth nor that there is the need to split the analysis into the pre- and post-crisis. The reason for this is both economical and statistical. Some economists (Botta 2014, Constantinescu et al. 2015) defend that what we are seeing is not just a cyclical downturn but the result of many macroeconomic and structural imbalances built over time. Consequently, we cannot assume that the crisis is just cyclical and believe that both the European and the global economy will return to its previous levels without major costs. The 2008 events unveiled some major economic imbalances that were underlying the significant economic growth observed in some Member States already before 2008. The fast increase in consumption and investment happened in parallel with

a significant increase in access to credit and indebtedness of those economies. This is reflected by the negative current account deficits, the deterioration of their International investment positions and/or the increasing levels of Government debt. From the statistical perspective, the drivers of growth are likely to be the same both before and after the crisis, even assuming that an after-crisis is already in place. In the post-crisis period the drivers' effect on growth is likely to be amplified. Instead of hampering the statistical analysis, the inclusion of pre- and post-crisis years highlights the positive and negative effects on growth helping in identifying most and least resilient regions. In this sense the crisis "statistically helped" in separating the signal from the noise. For these reasons, we did not consider the crisis as a structural break in the time series but fully incorporated it into the analysis with the entire 2003-2013 period considered as a whole.

The dependent variable y is based on real growth rates of regional gross value added – GVA – per capita. To allow for a time lag and to smooth out sharp changes in yearly growth rates, y at year t is computed as the geometric mean of y in the following 3 years $(y_{t+1}, \ldots, y_{t+3})$. Real growth rates of GVA are available in EUROSTAT at the regional level for most of the EU member states. For some countries the time series of GVA real growths are not available or not complete at the regional level. In these cases growth rates have been estimated on the basis GVA series by economic activity according to EU NACE codes ('Nomenclature Generale des Activites Economiques', revision 2).

The set of explanatory variables is described in the following:

Stage of development GVA per capita in constant prices is chosen to describe the stage of economic development of the regions. According to the neo-classical growth theory (Solow 1956) the growth rate of poor economies is higher than that of more developed economies and, consequently, their income per head (or equivalent) would catch up with that of richer economies. The level of GVA per capita are therefore expected to be one of the most important growth factors. GVA per capita values used in the analysis are computed from the reference year and are consistent with the GVA per capita growths in constant price used for the response y.

Urban areas The level of agglomeration is included in order to test whether more agglomerated regions perform better as advocated by the new economic geographers (Krugman 1998). The shares of people living in metropolitan areas or commuting belts are included as a proxy for the presence of dense urban areas. Cities commuting belts follow the definition of Functional Urban Areas according to the methodology jointly developed by the Organization for Economic Co-operation and Development (OECD) and the European Commission (Dijkstra, Poelman 2012). The year of reference of this indicator is 2006 as time series are not available so far. It is worth noting that this is a very slow moving indicator. Given the context of the analysis, it is expected to play the role of a static control variable more than a dynamic factor of growth[4].

Road infrastructure The road infrastructure indicator depends both on the availability of roads and the spatial distribution of the population. It does not simply measure the number or density of road kilometers but takes into account the population density of the areas connected by the road network. The level of road infrastructure is then a 'potential accessibility' indicator based on the assumption that the attraction of a destination increases with size, represented by population, and declines with distance, travel time or costs. The indicator is estimated by EC-DG for Regional and Urban Policy on the basis of results of the project described in (Stelder 2013). It is available for one year only, 2012. In a similar vein as for the Urban areas indicator, an indicator of this type serves here more as a static control variable than as a dynamic factor of growth.

Quality of governance Institutions have been recently emphasized as playing a key role in explaining the causes of economic growth/stagnation (Rodrik et al. 2004, Kwok, Tadesse 2006, Rodríguez-Pose 2013). Recent analyses at the regional level in the EU

[4]The limited time span of the analysis and the fact the no time series is available for this indicator makes its role differ from the others explanatory factors

uncovered an important sub-national dimension that can partly explain the observed within-country divergences in economic performance (Charron et al. 2012, Charron, Lapuente 2013, Rodríguez-Pose 2013). It is then important to include in the analysis a measure of quality of governance at the regional level. The regional data used in the paper is computed on the basis of the regional Quality of Government index – QoG – by the University of Gothenburg (Charron et al. 2014) and a composite index of national indicators yearly published by the World Bank and the World Economic Forum. The regional values of the QoG index are used to compute the regional/national ratio of the perceived quality of institutions within each country. These ratios are then applied to the national aggregated index computed from the World Bank and the World Economic Forum selection of indicators for the whole period under analysis. This national index is based on a total of 14 indicators, 6 coming from the World Bank-Worldwide Governance database and 8 from the World Economic Forum-Global Competitiveness Index database[5]. This approach allowed us to set up a time series of an indicator measuring the quality of institutions at the regional level.

Macroeconomic conditions Two indicators which provide a proxy for the macroeconomic context of the regions are included in the analysis: the *Net foreign position* – NFP – and *Government debt*. NFP is measured with the Net International Investment Position indicator available in EUROSTAT at the national level as the difference between national assets and liabilities of the country with respect to the rest of the world, expressed as a percentage of national Gross Domestic Product, GDP. The indicator records the net financial position of the domestic sectors of the economy versus the rest of the world, as the share of GDP. It is frequently used in economic analysis and research focusing on external vulnerability of countries and the risk of crises (DG ECFIN 2012). NFP is also highly correlated with the level of indebtedness of the households and the financial sector. Typically, highly negative values of net foreign position result from persistently high current account deficits and this is why the indicator is used as a measure of country vulnerability: the lower (or more negative) its values, the more vulnerable the country. The government debt, available in EUROSTAT as percentage of national GDP, is the second macroeconomic indicator included in the analysis. The relation between economic growth and government indebtedness is still an open and controversial issue, especially regarding the minimum level where government debt starts to be significantly detrimental to economic growth (Pescatori et al. 2014). However, the importance of public debt as one of the factors of growth is not disputable, especially for the time period of the analysis. Both indicators, NFP and government debt, are available at the country level only. A straightforward regionalization method is adopted that firstly redistributes the national value across the regions according to their population share, and secondly normalizes the regional values as shares of regional GDP. The approach is simpler than the one employed for the quality of governance but its rational is clear: it assumes that the level of national government debt, for instance, is distributed across the regions according to their population and their GDP. This can be seen as a rescaling procedure rewarding highly productive regions (where few people produce a high GDP) and, symmetrically, penalizes those with high population levels and low GDP that are assigned a higher debt share. This implicitly assumes that these regions (with low GDP and high population) are more affected by a deteriorated macroeconomic environment because more vulnerable.

Human capital Two human capital related indicators are included in the analysis, namely *lowly-* and *highly-educated workforce*. They are available at the regional level in EUROSTAT and are defined respectively as: 1. share population aged 25-64

[5]Selected indicators from the World Bank: 1. Voice & Accountability, 2. Political stability, 3. Government effectiveness, 4. Regulatory quality, 5. Rule of law, 6. Control of corruption. Selected indicators from the World Economic Forum: 7. Property rights, 8. Intellectual property protection, 9. Efficiency of the legal framework in settling disputes, 10. Efficiency of the legal framework in challenging regulations, 11. Transparency of government in policy making, 12. Business costs of crime and violence, 13. Organized Crime, 14. Reliability of police services.

with at most secondary education attainment and 2. share of population aged 25-64 with completed tertiary education.

Labour market Similarly to the human capital component, the labor market is described by two classical indicators available in EUROSTAT at the regional level: *long-term unemployment* and *employment* rates. The former is the 12-month or more unemployment rates as % of active population; the latter is the percentage of employed persons aged 20-64 with respect to the population cohort 20-64.

Research & Innovation As a proxy for the research and innovation potential of a region a composite index is computed from seven indicators available in EUROSTAT at the regional level: 1. Total patent applications, 2. Core creative class employment, 3. Knowledge workers, 4. Total intramural R&D expenditures, 5. Human resources in science and technology, 6. High-tech patents and 7. ICT patents.

All the explanatory factors listed above are included in the analysis at the NUTS2 level and for the time period 2003-2010. As aforementioned, the only exceptions are Urban areas and Road infrastructure that refer to the years 2006 and 2012 respectively. Descriptive statistics are shown in Table A.1 of the Appendix for all the factors of growth included in the analysis. Statistics are separately computed for the EU28, EU15 and EU13 groups of countries.

4 Results

Randomized CART and MARS are used in a complementary way to understand data on regional growth in the EU as a whole (EU28), with a sample size of 2144 observations (268 regions x 8 years). The EU is an interesting mix of Member States with a long EU membership (EU15) and ones which joined the EU after 2004 (EU13)[6]. Is it meaningful to consider the two groups all together? The t-test (Mood et al. 1974) carried out for all the variables in the analysis shows indeed a significant difference between the two groups (with p-values always < 0.0001). The EU13 averages are always significantly lower than the EU15 ones, apart from growth rate and long-term unemployment. The EU13 group indeed grows faster than the EU15 and has higher average levels of long-term unemployment.

Given t-test results, the EU28 scenario is integrated, when feasible, with additional analyses carried out separately for the EU15 group and the EU13 one. The sample size of the two groups of countries is 1680 and 464 respectively. This does not allow for getting reliable results from CART for the EU13 case. CART results are then discussed only for the EU28 and the EU15 scenarios. MARS is instead run for the three groups and provides some insights into the different mechanisms of growth across different areas.

All the analyses are run using SAS® ver. 9.4.

4.1 Randomized CART

The dependent variable y used in all the simulations is the real regional GVA per capita growth rate. For CART analysis y is categorized into three classes based on yearly quartiles: low growth rate if $y < P_{25\%}$; intermediate if $P_{25\%} \leq y < P_{75\%}$; high if $y \geq P_{75\%}$. The purpose is in fact to identify the most important factors driving high or low regional growth. A sensitivity analysis has been carried out to assess the robustness of results with respect to different types of categorization: 4 classes, with thresholds $\{P_{25\%}, P_{50\%}, P_{75\%}\}$, and 5 classes, with thresholds $\{P_{20\%}, P_{40\%}, P_{60\%}, P_{80\%}\}$. Variable importance ranking is stable for both the EU28 and the EU15 scenarios,especially for the most important factors (See Table A.2 in the Appendix for results).

Table 1 shows the parameters used in the randomized CART analysis. Two different analyzes are carried out: one for the EU as a whole (EU28) and one for the EU15 group.

[6]EU15 includes: Belgium, Denmark, Germany, Ireland, Greece, Spain, France, Italy, Luxembourg, Netherlands, Austria, Portugal, Finland, Sweden and United Kingdom; EU13 includes: Bulgaria, Czech Republic, Estonia, Croatia, Cyprus, Latvia, Lithuania, Hungary, Malta, Poland, Romania, Slovenia and Slovakia.

Table 1: Parameters used for the CART analysis

CART parameters	**Selected option**
Dependent variable	Real GVA growth rate categorized into three classes: *low* if below $P_{25\%}$ *medium* if between $P_{25\%}$ and $P_{75\%}$ *high* if above $P_{75\%}$
Type of split	Binary
Criterion	Gini index
Pruning	none
Limit on the leaf dimension	Yes minimum number of observations = 20
Cross-validation	Yes sample partition into 90% (training) and 10% (validation)
Randomization	Yes 1000 randomized $90 \div 10$ partitions
variable importance criterion	*SSE*

Main results are shown in Table 2. The two randomized CARTs are characterized by average misclassification rates *MR* of 0.29 and 0.32 respectively for the EU28 and EU15 case. As expected, the average accuracy of the models improves with larger sample sizes. Factors are reordered according to the normalized variable importance IMP_{SSE}, as defined in (2). IMP_{SSE} and vary between 0 (no importance) and 1 (highest importance). Estimated standard deviation for the mean, coefficient of variation and lower and upper limits of the 95% confidence intervals for the mean are also shown in Table 2. A factor is considered having an important effect on growth if $IMP_{SSE} \geq 0.6$, a medium effect if $0.2 \leq IMP_{SSE} < 0.6$ and non-relevant effect if $IMP_{SSE} < 0.2$.

The two cases share some common features, as some variables remain important or unimportant in both cases, but interesting specificities can be identified. The share of poorly educated workforce and the Net foreign position are the two most important factors in both cases. The former is negatively (partially) correlated with growth rate (Table 3), meaning that the growth rate is higher where the share of poorly educated people is lower; the latter is positively (partially) correlated with growth rate, showing that the lower the vulnerability level of the region, measured by the Net foreign position, the higher its growth rate (Table 3). The stage of development is the third most relevant factor for the EU28, with negative orientation which conforms to the neoclassical theory of growth. However the stage of development becomes a medium-impact factor for the EU15 group, with the Quality of Governance being the third most important growth factor. The EU15 group includes all the most developed economies and the role of institutions assumes more importance. This is the effect of the interdependency between different factors of growth. The quality of governance is unlikely to have a purely autonomous effect on economic growth; it rather interacts with other factors to play a role in relative terms. Institutions remain silent till a certain level of level of development is reached, above which they start to make the difference and become more important than other basic aspects like the initial regional endowment. Accordingly, the Quality of Governance is not detected as important by MARS in the least developed economies of the EU13 group (Section 4.2). The two least important factors are in both cases the share of Employment and Research and Innovation, all the others being medium-impact factors. It is interesting to note that Government debt has a higher, negative effect for the EU15 model.

In results interpretation it is important to remember that the analysis is not capable of capturing factors' impact on long-term growth, due to the limited time-span of the

Table 2: CART results: Variable ranking based on their average normalized importance(from 1000 randomized CARTS)

EU28 model	**Average $MR = 0.29$**				
Factor	**Mean**	**Std Dev**	**CoV** $\left(\frac{\sigma}{\mu} \cdot 100\right)$	**CI for Mean** lower 95%	upper 95%
Lowly educated workforce	1.00	0.00	0.01	1.00	1.00
Net foreign position	0.83	0.08	9.14	0.82	0.83
Stage of development	0.66	0.07	10.37	0.65	0.66
Quality of governance	0.37	0.07	18.60	0.37	0.37
Urban areas	0.35	0.10	28.60	0.34	0.35
Long-term unemployment	0.25	0.10	39.31	0.25	0.26
Road infrastructure	0.24	0.08	34.16	0.24	0.25
Government debt	0.24	0.10	39.84	0.23	0.25
Highly educated workforce	0.17	0.12	70.29	0.16	0.18
Employment	0.12	0.11	91.02	0.11	0.12
Research and Innovation	0.09	0.09	96.64	0.09	0.10
EU15 model	**Average $MR = 0.32$**				
Factor	**Mean**	**Std Dev**	**CoV** $\left(\frac{\sigma}{\mu} \cdot 100\right)$	**CI for Mean** lower 95%	upper 95%
Net foreign position	0.98	0.04	3.97	0.98	0.98
Lowly educated workforce	0.98	0.03	3.41	0.97	0.98
Quality of governance	0.46	0.08	18.26	0.46	0.47
Government debt	0.35	0.09	25.99	0.34	0.35
Stage of development	0.32	0.10	32.63	0.31	0.32
Urban areas	0.30	0.09	30.54	0.30	0.31
Long-term unemployment	0.23	0.13	57.26	0.23	0.24
Road infrastructure	0.22	0.12	55.95	0.21	0.23
Research and Innovation	0.20	0.14	68.28	0.19	0.21
Employment	0.20	0.13	64.90	0.19	0.20
Highly educated workforce	0.18	0.13	72.54	0.17	0.18

available regional data time series.

4.2 Multivariate Adaptive Regression Splines

Table 4 compares the accuracy of different MARS models for the EU28, the EU15 and the EU13. In all the cases the penalty γ is set to 0.05 which corresponds to a moderate penalty (Section 2.2). The goodness of fit of the models, measured by GCV and adjusted R^2, generally increases as higher-order interactions are included. A simple additive model is not suitable to investigate growth patterns as interacting effects are important elements of the analysis. Nevertheless, third or higher-order models do not substantially increase the model accuracy so second-order models are considered as the best ones (only third-order interaction models are shown in Table 4). R^2 values are overall pretty low, ranging from a minimum of 0.39, for the EU15-additive model, to a maximum of 0.62, for the second and third-order interactive model for the EU13. This means that the noise-to-signal ratio is high and, consequently, the margin of error cannot be considered as negligible. Any interpretation of results must then be tempered by these considerations. The EU15 accuracy is the lowest one among the three, well below 50% of variance accounted for, and for this reason this model has been discarded from further analysis.

The second-order EU28 model is able to account for 50% of the variance (Table 4). The three most important factors are Net foreign position, Stage of development and share of Lowly educated workforce (Table 5). These are the same as the ones identified by CART analysis for the EU-28 model, even if with a different order of importance (Table

Table 3: Partial correlation coefficients between real GVA growth rate and its explanatory factors (p-values in brackets)

	Stage of dev.	**Urban areas**	**Road infr.**
EU28	-0.195	0.011	-0.041
	(**<0.0001**)	(0.628)	(0.066)
EU15	-0.128	0.000	-0.044
	(**<0.0001**)	(0.9980)	(0.0774)
	Quality of gov.	**Net foreign position**	**Government debt**
EU28	0.040	0.288	-0.323
	(0.0681)	(**<0.0001**)	(**<0.0001**)
EU15	0.122	0.257	-0.180
	(**<0.0001**)	(**<0.0001**)	(**<0.0001**)
	Lowly ed. workforce	**Highly ed. workforce**	**Long-term unempl.**
EU28	-0.143	-0.134	0.113
	(**<0.0001**)	(**<0.0001**)	(**<0.0001**)
EU15	-0.052	-0.148	0.043
	(0.0381)	(**<0.0001**)	(0.084)
	Employment	**Research and Innovation**	
EU28	-0.092	0.074	
	(**<0.0001**)	(**0.0008**)	
EU15	-0.118	0.071	
	(**<0.0001**)	(**0.0045**)	

2). To get further insight into the relationships between growth and its factors, Figure 1 shows the contribution to the prediction, which is the estimated growth rate, of the four most important factors in the additive model[7]. For the Net foreign position factor, which is used as a proxy for the level of vulnerability or resilience of the region (Lau et al. 2003), the dependence is positive when net foreign position tends to the balanced level from negative values. After this level, the factor does not influence growth rate any longer as the curve almost levels off (Figure 1a). The dependence of growth rate on the stage of development is in line with the neoclassical growth theory, as in the CART case: at the increase of the stage of development, the economic growth rate slows down. The pace is however different: the decrease of growth rate is steeper for low levels and slower for higher levels of development (Figure 1b). The dependence of growth rate on the share of Lowly educated workforce is instead somewhat surprising at least in the right-hand side of the curve with higher shares of lowly educated people positively associated with growth rate (Figure 1c). A possible explanation can be found in the effects of interaction between Lowly educated workforce and other important factors, as can be seen in Table 4. It is also true that in poor economies, with a low-tech job environment, lowly educated people boost the first stage of the economy. This assumption would nevertheless need further investigation. Government debt shows instead an interesting relationship with growth rate with an almost neutral effect up to a certain level of debt, around 130% of GDP, above which growth rate decreases steeply as debt increases (Figure 1d). Even if some of the components show interesting and conceptually reasonable relationships with growth, any interpretations must be taken with caution, given the relative low level of accuracy of the model. Also, the existence of collinearity among the explanatory factors generally causes problems in the interpretation of results (Friedman 1991).

The EU13 model reaches the highest accuracy, explaining 62% of the variance and Table 5 lists the most important factors. The most important factor of growth is the Net foreign position with a positive effect on growth rate as for the EU28 model (Figure 2a).

[7] For interpretation purposes only the additive model components are shown.

Table 4: Comparison of different MARS models and their interacting factors

Model	*GCV* based R^2	Adj. R^2	Relevant interactions
EU28 additive	0.44	0.45	none
EU28 second-order	0.48	0.50	Stage of dev. – Lowly ed. workforce Stage of dev. – Net foreign position Net foreign position – Gov. debt Net foreign position – Long-term unempl. Net foreign position – Lowly ed. workforce Lowly ed. workforce – Highly ed. workforce
EU28 third-order	0.49	0.50	Net foreign position – Gov. debt Net foreign position – Long-term unempl. Stage of dev. – Lowly ed. workforce Net foreign position – Lowly ed. workforce – Stage of dev. Net foreign position – Gov. debt – Employment Net foreign position – Gov. debt – Highly ed. workforce
EU15 additive	0.39	0.39	none
EU15 second-order	0.39	0.42	Net foreign position – Quality of governance Net foreign position – Long-term unempl. Net foreign position – Road infrastructure Lowly ed. workforce – Highly ed. workforce Stage of dev. – Gov. debt
EU15 third-order	0.42	0.43	Net foreign position – Quality of Governance Net foreign position – Employment Net foreign position – Lowly ed. workforce Stage of dev. – Quality of Governance Net foreign position – Lowly ed. workforce – Quality of gov. Net foreign position – Lowly ed. workforce – Employment Net foreign position – Stage of dev. – Quality of gov.
EU13 additive	0.51	0.56	none
EU13 second-order	0.56	0.62	Net foreign position – Stage of dev. Net foreign position – Highly ed. workforce Net foreign position – Lowly ed. workforce Highly ed. workforce – Urban areas Long-term unemployment – Gov. debt
EU13 third-order	0.56	0.62	Net foreign position – Stage of dev. Net foreign position – Highly ed. workforce Net foreign position – Lowly ed. workforce Long-term unemployment – Gov. debt Highly ed. workforce – Urban areas

The shape of the curve is however different, only slightly increasing for low and negative values and a steep increase as the Net foreign position reaches the parity (zero level). Having a highly educated workforce is important in this case with a constant and positive effect (Figure 2b). Long-term unemployment is an important factor for the EU13 case but with a surprising positive effect (Figure 2c). Apart from model accuracy considerations, this may be due to the interaction of long-term unemployment with other important factors as can be seen in Table 4. The stage of development is important for the EU13 group as well but to a lesser extent than in the EU28 case as its normalized importance is 38% against the 54% for EU28 (Table 5). In both cases the higher the GVA levels the lower the growth rate, with a minor anomaly in the EU13 case for very low GVA values (Figure 2d).

Finally, Government debt is only slightly less important than the stage of development, in the EU13 case and its relationships with growth is the opposite as that of the EU28 case, with a positive effect until values of around 25% of the GDP above which the effect levels off (Figure 3). This is in line with the assumption that there is a certain level of government debt beyond which economic growth starts to be impeded.

Table 5: Second-order MARS models: important variables for the EU28 and the EU13 models

Model	Variable	Number of Basis	Normalized importance
EU28	Net foreign position	8	100.00
	Stage of development	6	54.07
	Lowly ed. workforce	9	38.66
	Government debt	1	13.19
	Long-term unemployment	1	9.51
	Highly ed. workforce	2	6.61
	Quality of Governance	2	5.97
EU13	Net foreign position	8	100.00
	Highly ed. workforce	5	67.35
	Long-term unemployment	4	62.66
	Stage of development	2	38.32
	Government debt	1	34.01
	Urban areas	1	21.28
	Lowly ed. workforce	2	6.69

5 Conclusive remarks

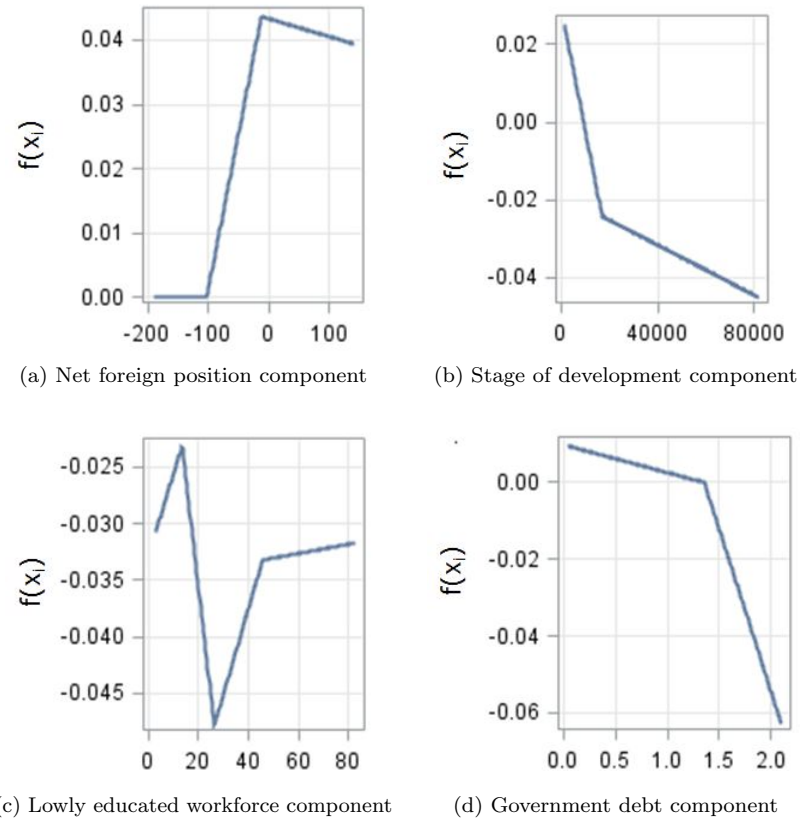
In the search of the main determinants of regional economic growth, in our view this paper features some novelties.

First it employs non-parametric, data-driven statistical models as an alternative to more classical regression techniques. They are more suitable to deal with complex data which feature non-linearities and interaction effects and proved to be rather informative with respect to the type of relationships between growth and its main factors.

Second, it reaches the regional, sub-national level across the whole EU. Reaching the regional level proved to be particularly demanding due to scarcity of reliable and comparable regional data at the EU level. A large part of the analysis has been then devoted to the building of comparable panel data but longer time series and richer datasets would be needed to overcome the short-run perspective of the analysis and the omission of important factors of growth.

Third, the analysis maintains the traditional factors used in the literature on growth but also incorporates the institutional quality and two variables which aim to reflect the macroeconomic conditions in which the regions operate.

The main results of the analysis can be summarized as follows. Macroeconomic conditions are found to be important to explain the economic growth of regions. They are typically national variables and have been here broken down the regional level with a straightforward approach based on regional population shares. The macroeconomic framework has been generally ignored in the analysis of regional growth trends and convergence, traditionally focused on factors of production (typically infrastructure and education) and the drivers behind Total Factor Productivity (quality of institutions, technological progress, research and innovation). The economic crisis has however shown that the macroeconomic conditions of the economies in which European regions operate are critical and our results actually show the importance of macroeconomic factors in explaining regional growth. The macroeconomic framework is approximated by two variables of the Scoreboard of Indicators of the Macroeconomic Imbalances Procedure, the Net International Investment Position and the Government Debt. While they are highly correlated to other variables such as the current account balance, private sector debt or the financial sector liabilities, other important variables such as the unit labor costs or the export market shares are not captured. The net foreign position is intended to capture the degree of vulnerability (negative values) or resilience (positive values) of the national economy in which the region operates. Government debt is also a very



(a) Net foreign position component

(b) Stage of development component

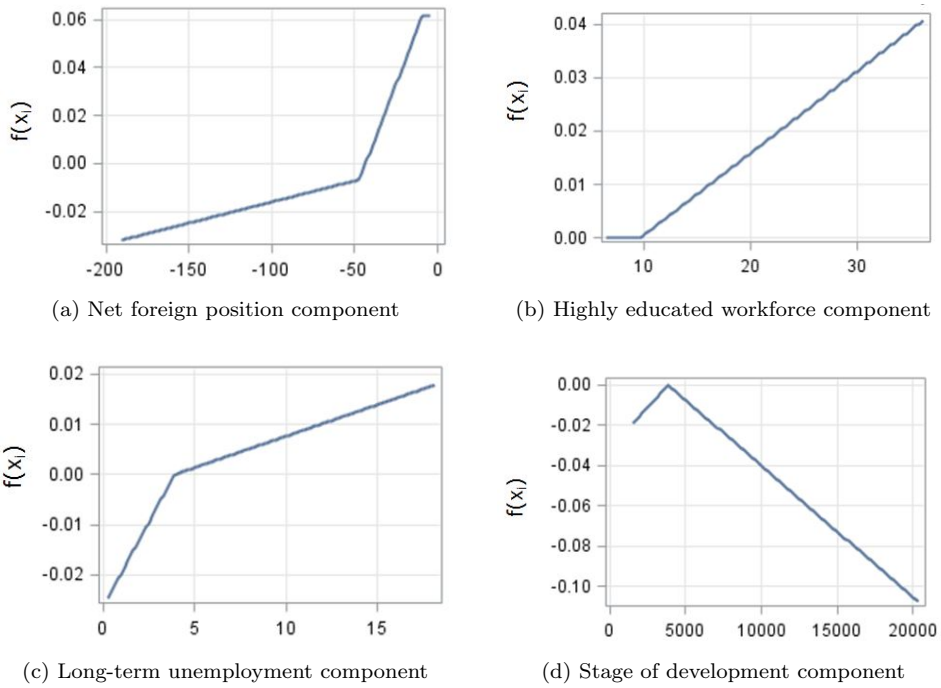
(c) Lowly educated workforce component

(d) Government debt component

Figure 1: MARS EU28 additive model: dependence of growth on the main four factors, as detected by the model.

relevant indicator of vulnerability of a Member State and its capability to deal with economic crisis and asymmetric shocks. The crisis has shown that the adverse effects of the uncertainties regarding the sustainability of the Government debt spreads to the whole economy through a contraction of the financing supply by the market and an increase in the risk premium faced by public and private operators. What we observe is that positive levels of net foreign position are always fostering growth while for government debt the picture is more diverse. Results suggest that there are good debt levels which have a positive effect on growth (see the EU13 scenario). However, above a certain threshold, around 130% of GDP, debt has a clear counteracting effect. A high level of general government debt is therefore a problem 'per se' and its consequences cannot be compensated by a low level of private debt. Further discussion and investigation may be required to improve the availability of regional statistics on macroeconomic variables, to analyze the impact of the macroeconomic framework on regional economic growth.

Human capital is another relevant factor driving economic growth of the EU regions. The paper confirms the wide consensus of the economic literature about the major importance of human capital for economic growth, in line with a number of studies (Solow 1956, Mankiw et al. 1992, Lucas 1988, Barro 1989). Human capital is measured at both ends of the scale, in terms of lowly and highly educated people of working age. The analysis interestingly indicates that higher shares of poorly educated people are more detrimental than lower shares of highly educated ones, as also highlighted by a recent study at the regional level in OECD countries (OECD 2012). Human capital is likely to be transmitted into higher economic growth through higher productivity of the labor force but also through technological progress, increasing therefore the Total Factor Productivity of countries and regions. The importance of the institutional quality is also confirmed by the analysis, fully in line with a vast number of economic analyses such as in (Knack, Keefer 1995, Acemoglu et al. 2003, Rodrik et al. 2004) which have identified the quality



(a) Net foreign position component

(b) Highly educated workforce component

(c) Long-term unemployment component

(d) Stage of development component

Figure 2: MARS EU13 additive model: dependence of growth on the main four factors, as detected by the model.

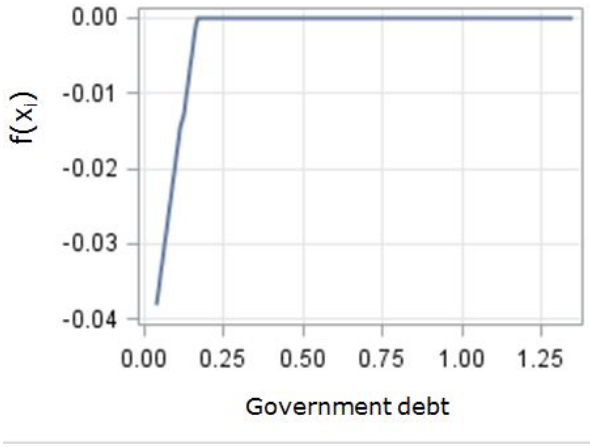


Figure 3: MARS EU13 additive model: dependence of growth on Government debt, as detected by the model.

of institutions as a major explanatory factor of growth and regional disparities. Good institutions may lead to higher economic growth through a higher productivity of the factors of production, lower rent-seeking behaviors, more room for technological progress and innovation, lower administrative costs and corruption, etc. There has been however a wide debate on the literature on what does institutional quality mean and what the most relevant indicators that capture this notion are. This paper uses the Quality of Government Index, published twice (in 2010 and 2013) by the University of Gothenburg, to regionalize a longer time series of well-known governance indicators by the World Bank and the World Economic Forum.

Finally, the evidence given by the model about the impact of other factors on economic growth such as those on the population agglomeration, infrastructure or the level of innovation is more limited and inconclusive. As striking as it may seem, these findings are however in line with what was found in a recent OECD analysis on regional growth

(OECD 2012). It is beyond the purpose of this paper to dig into why these factors are detected as non-influential. One of reasons is surely related to the comparatively short-term perspective of the analysis due to data availability constraints. An interesting explanation of the little support for the link between innovative activities and growth at the regional level – *the innovative puzzle* – is for example provided in the recent OECD analysis just mentioned (OECD 2012).

The conclusions of the paper underpin the rationale behind the reinforcement of the European economic governance and the conditionality mechanisms set in the new architecture of the EU regional funds 2014-2020. In 2011 the European institutions adopted a new economic surveillance procedure for the prevention and correction of macroeconomic imbalances which strengthens the economic surveillance powers at the EU level. The reason behind is the recognition that significant factors influencing economic performance and stability had overall been ignored by the EU economic surveillance, which was limited to the monitoring of the fiscal and budgetary positions of Member States until the advent of the economic crisis. The allocation of regional funds is now made conditional to (i) compliance with a number of ex-ante conditionalities which aim to ensure a minimum level of framework conditions related to institutional quality and to (ii) compliance with the fiscal and macroeconomic procedures enshrined in the EU primary and secondary legislation. The rationale behind this conditionality is that the effectiveness of the expenditure is reinforced by good institutional quality and sound economic policies as suggested by the results of this study.

References

Acemoglu D, Johnson S, Robinson J, Thaicharoen Y (2003) Institutional causes, macroeconomic symptoms: Volatility, crises, and growth. *Journal of Monetary Economics* 50: 49–123. CrossRef.

Agresti A (1990) *Categorical Data Analysis*. New York: John Wiley & Sons. CrossRef.

Austin PC (2007) A comparison of regression trees, logistic regression, generalized additive models, and multivariate adaptive regression splines for predicting AMI mortality. *Statistics in medicine* 26: 2937–2957. CrossRef.

Barro R (1989) Economic growth in a cross section of countries. NBER working paper no. 3120. CrossRef.

Barro R, Sala-i-Martin X (1992) Convergence. *Journal of Political Economy* 100: 223–251. CrossRef.

Botta A (2014) Structural asymmetries at the roots of the eurozone crisis: What's new for industrial policy in the EU? Levy Economics Institute working paper no. 794. CrossRef.

Breiman L, Friedman J, Olshen R, Stone C (1984) *Classification and regression trees*. Wadsworth & Brooks

Charron N, Dijkstra L, Lapuente V (2014) Regional governance matters: Quality of government within European Union member states. *Regional Studies* 48: 68–90. CrossRef.

Charron N, Lapuente (2013) Why do some regions in Europe have higher quality of government? *The Journal of Politics* 75: 567–582. CrossRef.

Charron N, Lapuente V, Dijkstra L (2012) Regional governance matters: A study on regional variation in quality of government within the EU. DG Regional Policy working papers WP01/2012

Constantinescu C, Mattoo A, Ruta M (2015) The global trade slowdown: Cyclical or structural? International Monetary Fund, Strategy, Policy and Review Department, working paper no. 15/6. CrossRef.

Crescenzi R, Rodríguez-Pose A (2008) Infrastructure endowment and investment as determinants of regional growth in the European Union. European Investment Bank Paper 13(2)

Cristelli M, Tacchella A, Pietronero L (2015) The heterogeneous dynamics of economic complexity. *PLOS ONE* 10: 1–15. CrossRef.

Curtis P, Kokotos P (2009) A decision tree application in tourism based regional economic development. *Tourismos: an international multidisciplinary journal of tourism* 4: 169–178

Dall'erba S, Le Gallo J (2008) Regional convergence and the impact of European structural funds 1989-1999: A spatial econometric analysis. *Papers in Regional Science* 87: 219–244. CrossRef.

De Veaux R, Psichogios D, Ungar L (1993) A comparison of two nonparametric estimation schemes: MARS and neural networks. *Computers & Chemical Engineering* 17: 819–837. CrossRef.

Deichmann J, Eshghi A, Haughton D, Sayek S, Teebagy N (2002) Application of multiple adaptive regression splines (MARS) in direct response modeling. *Journal of Interactive Marketing* 16: 15–27. CrossRef.

DG ECFIN (2012) Scoreboard for the surveillance of macroeconomic imbalances. Technical Report 92, European Economy Occasional Papers

Dijkstra L, Poelman H (2012) Cities in Europe. the new OECD-EC definition. European Commission - DG for Regional and Urban Policy WP 01/2012

Durlauf S, Johnson P (1995) Multiple regimes and cross-country growth behaviour. *Journal of Applied Econometrics* 10: 365–384. CrossRef.

Friedman J (1991) Multivariate adaptive regression splines. *Annals of Statistics* 19: 1–141. CrossRef.

Grömping U (2009) Variable importance assessment in regression: linear regression versus Random Forest. *The American Statistician* 63: 308–319. CrossRef.

Hastie T, Tibshirani R, Friedman J (2001) *The elements of statistical learning. Data mining, inference and prediction.* Springer. CrossRef.

Knack S, Keefer P (1995) Institutions and economic performance: cross-country tests using alternative institutional measures. *Economic and Politics* 7: 207–227. CrossRef.

Krugman P (1998) What's new about the new economic geography? *Oxford Review of Economic Policy* 14: 7–11. CrossRef.

Kwok C, Tadesse S (2006) National culture and financial systems. *Journal of International Business Studies* 37: 227–247. CrossRef.

Lau F, Yung S, , Yong I (2003) Introducing a framework to measure resilience of an economy. *Hong Kong Monetary Authority Quarterly Bulletin* 35: 28–34

Leathwick JR, Rowe D, Richardson J, Elith J, Hastie T (2005) Using multivariate adaptive regression splines to predict the distributions of New Zealand's freshwater diadromous fish. *Freshwater Biology* 50: 2034–2052. CrossRef.

Lucas R (1988) On the mechanics of economic development. *Journal of Monetary Economics* XXII: 3–42. CrossRef.

Mankiw N, Romer P, Weill D (1992) A contribution to the empirics of economic growth. *Quarterly Journal of Economics* 107: 407–437. CrossRef.

Mardia K, Kent, J.T.and Bibby J (1979) *Multivariate Analysis.* Academic Press, San Diego, U.S.

Mezrich J (1994) When is a tree a hedge? *Financial Analysts Journal* 50: 75–81. CrossRef.

Mohl P, Hagen T (2010) Do EU structural funds propmote regional growth? New evidence from various panel data approaches. *Regional Science and Urban Economics* 40: 353–365. CrossRef.

Mood AM, Graybill FA, Boes DC (1974) *Introduction to the Theory of Statistics* (3rd ed.). McGraw-Hill, New York

Moore DS (2004) *The basic practice of statistics* (3rd ed.). W.H. Freeman and Co., New York and Basingstoke

OECD (2012) *Promoting growth in all regions*. OECD publishing

Pescatori A, Sandri D, Simon J (2014) Debt and growth: Is there a magic threshold? Working Paper WP/14/34, International Monetary Fund. CrossRef.

Ramajo J, Marquez M, Hewings G, Salinas M (2008) Spatial heterogeneity and interregional spillovers in the European Union: Do cohesion policies encourage convergence across regions? *European Economic Review* 52: 551–567. CrossRef.

Rodríguez-Pose A (2013) Do institutions matter for regional development? *Regional Studies* 47: 1034–1047. CrossRef.

Rodríguez-Pose A, Fratesi U (2004) Between development and social policies: The impact of european structural funds in objective 1 regions. *Regional Studies* 38: 97–113. CrossRef.

Rodríguez-Pose A, Garcilazo E (2013) Quality of government and the returns of investment: Examining the impact of cohesion expenditure in european regions. OECD Regional development working papers 2013/12. CrossRef.

Rodrik D, Subramanian A, Trebbi F (2004) Institutions Rule: The Primacy of Institutions over Geography and Integration in Economic Development. *Journal of Economic Growth* 9: 131–165. CrossRef.

SAS (2014) The HP-SPLIT procedure, on-line material. Technical report, SAS Institute Inc., Cary, NC, USA

Solow R (1956) A contribution to the theory of economic growth. *Quarterly Journal of Economics* 70: 65–94. CrossRef.

Stelder D (2013) Changes in road infrastructure and accessibility in europe since 1960. Tender reference nr 2012.ce.16.bat.040, European Commission, DG for Regional and Urban policy

Strobl C, Boulesteix A, Zeileis A, Hothorn T (2007) Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics* 8: 1–21

Sugihara G, May R, Ye H, Hsieh C, Deyle E, Fogarty M, Munch S (2012) Detecting causality in complex ecosystems. *Science* 338: 496–500. CrossRef.

Varian H (2014) Big data: New tricks for econometrics. *Journal of Economic Perspectives* 28: 3–28. CrossRef.

Weir N, Fayyad UM, Djorgovski S (1995) Automated star/galaxy classification for digitized POSS-II. *The Astronomical Journal* 109: 2401–2414. CrossRef.

A Appendix

Table A.1: Descriptive statistics of indicators included in the analysis

GVA per capita

year	*Country group*	*Average*	*SD*	*Min*	*Max*
2003	**EU_13**	5036	3137	1498	15838
	EU_15	21537	7072	9462	66904
2003 Total		17966	9354	1498	66904
2004	**EU_13**	5283	3235	1564	16672
	EU_15	22017	7177	9511	68453
2004 Total		18396	9493	1564	68453
2005	**EU_13**	5525	3424	1583	17564
	EU_15	22324	7486	9583	73020
2005 Total		18688	9711	1583	73020
2006	**EU_13**	5851	3599	1649	18570
	EU_15	22913	7664	9714	74094
2006 Total		19221	9909	1649	74094
2007	**EU_13**	6171	3823	1754	20023
	EU_15	23468	8005	10063	78926
2007 Total		19725	10204	1754	78926
2008	**EU_13**	6386	3878	1814	20334
	EU_15	23426	8134	10152	81903
2008 Total		19739	10215	1814	81903
2009	**EU_13**	6078	3646	1720	18882
	EU_15	22387	7629	9847	75123
2009 Total		18857	9674	1720	75123
2010	**EU_13**	6193	3737	1661	19464
	EU_15	22754	7969	10109	76813
2010 Total		19170	9964	1661	76813

Urban Areas

year	*Country group*	*Average*	*SD*	*Min*	*Max*
2006	**EU_13**	32	29	0	100
	EU_15	42	33	0	100

Road infrastructure

year	*Country group*	*Average*	*SD*	*Min*	*Max*
2012	**EU_13**	48	26	7	115
	EU_15	105	81	1	311

Quality of governance*

year	*Country group*	*Average*	*SD*	*Min*	*Max*
2003	**EU_13**	-1.14	0.83	-2.94	0.32
	EU_15	0.33	0.72	-1.85	2.87
2003 Total		0.01	0.96	-2.94	2.87
2004	**EU_13**	-1.23	0.75	-2.87	0.11
	EU_15	0.41	0.75	-1.87	2.95
2004 Total		0.06	1.01	-2.87	2.95
2005	**EU_13**	-1.17	0.76	-2.83	0.04
	EU_15	0.32	0.73	-1.94	2.73
2005 Total		0.00	0.96	-2.83	2.73
2006	**EU_13**	-1.17	0.74	-2.64	0.13
	EU_15	0.33	0.81	-1.98	2.60
2006 Total		0.01	1.01	-2.64	2.60
2007	**EU_13**	-1.25	0.69	-2.67	0.23
	EU_15	0.47	0.87	-2.14	2.66
2007 Total		0.10	1.09	-2.67	2.66
2008	**EU_13**	-1.17	0.64	-2.63	0.33
	EU_15	0.33	0.84	-2.17	2.63
2008 Total		0.00	1.01	-2.63	2.63
2009	**EU_13**	-1.13	0.59	-2.66	0.24
	EU_15	0.15	0.86	-2.30	2.51
2009 Total		-0.13	0.97	-2.66	2.51
2010	**EU_13**	-1.03	0.58	-2.63	0.23
	EU_15	0.22	0.87	-2.26	2.52
2010 Total		-0.05	0.96	-2.63	2.52

* Note: the Quality of governance index is by construction expressed in z-scores

Net foreign position* (regionalized)

year	*Country group*	*Average*	*SD*	*Min*	*Max*
2003	**EU_13**	-39.8	27.7	-120.4	38.9
	EU_15	-12.9	27.1	-79.9	140.3
2003 Total		-18.7	29.4	-120.4	140.3
2004	**EU_13**	-43.9	30.1	-130.3	39.4
	EU_15	-15.6	29.0	-91.9	113.1
2004 Total		-21.7	31.5	-130.3	113.1
2005	**EU_13**	-49.1	33.3	-147.1	36.0
	EU_15	-14.7	34.8	-106.9	127.8
2005 Total		-22.1	37.3	-147.1	127.8
2006	**EU_13**	-56.6	36.9	-161.3	37.8
	EU_15	-17.0	40.1	-120.1	131.6
2006 Total		-25.5	42.7	-161.3	131.6
2007	**EU_13**	-65.6	37.8	-167.1	17.7
	EU_15	-19.7	42.5	-134.5	95.5
2007 Total		-29.6	45.6	-167.1	95.5
2008	**EU_13**	-72.5	38.3	-169.7	2.6
	EU_15	-16.2	41.7	-120.3	100.1
2008 Total		-28.4	47.1	-169.7	100.1
2009	**EU_13**	-79.8	41.9	-190.9	12.6
	EU_15	-17.6	50.7	-138.5	86.7
2009 Total		-31.1	55.3	-190.9	86.7
2010	**EU_13**	-81.5	39.2	-186.0	8.1
	EU_15	-17.1	51.9	-139.1	98.0
2010 Total		-31.1	56.1	-186.0	98.0

Continued on the next page ...

Government debt (regionalized)

year	Country group	Average	SD	Min	Max
2003	**EU_13**	0.42	0.19	0.06	0.87
	EU_15	0.69	0.29	0.06	1.67
2003 Total		0.64	0.29	0.06	1.67
2004	**EU_13**	0.45	0.21	0.05	0.93
	EU_15	0.70	0.29	0.06	1.64
2004 Total		0.64	0.29	0.05	1.64
2005	**EU_13**	0.42	0.22	0.05	0.94
	EU_15	0.71	0.30	0.06	1.67
2005 Total		0.65	0.31	0.05	1.67
2006	**EU_13**	0.42	0.25	0.04	1.09
	EU_15	0.70	0.31	0.07	1.68
2006 Total		0.64	0.32	0.04	1.68
2007	**EU_13**	0.40	0.25	0.04	1.06
	EU_15	0.67	0.31	0.07	1.63
2007 Total		0.62	0.32	0.04	1.63
2008	**EU_13**	0.37	0.25	0.05	1.10
	EU_15	0.72	0.32	0.14	1.68
2008 Total		0.64	0.33	0.05	1.68
2009	**EU_13**	0.49	0.29	0.07	1.35
	EU_15	0.85	0.33	0.16	1.85
2009 Total		0.77	0.35	0.07	1.85
2010	**EU_13**	0.52	0.28	0.07	1.34
	EU_15	0.93	0.36	0.20	2.10
2010 Total		0.84	0.38	0.07	2.10

Lowly educated workforce

year	Country group	Average	SD	Min	Max
2003	**EU_13**	22.4	11.0	5.6	80.2
	EU_15	35.5	16.8	3.8	84.7
2003 Total		32.7	16.6	3.8	84.7
2004	**EU_13**	21.5	10.7	5.1	76.4
	EU_15	34.2	16.2	3.5	82.8
2004 Total		31.4	16.0	3.5	82.8
2005	**EU_13**	20.4	10.6	4.5	74.8
	EU_15	33.2	15.4	3.6	81.3
2005 Total		30.4	15.4	3.6	81.3
2006	**EU_13**	19.2	10.1	4.6	73.5
	EU_15	32.7	15.1	3.8	80.8
2006 Total		29.8	15.3	3.8	80.8
2007	**EU_13**	18.5	10.0	4.3	73.4
	EU_15	32.1	15.1	3.4	81.4
2007 Total		29.2	15.2	3.4	81.4
2008	**EU_13**	17.9	10.0	4.4	72.2
	EU_15	31.4	15.0	3.0	82.0
2008 Total		28.5	15.1	3.0	82.0
2009	**EU_13**	17.3	9.9	4.2	69.2
	EU_15	30.6	14.8	4.0	79.6
2009 Total		27.7	14.9	4.0	79.6
2010	**EU_13**	16.7	9.8	3.3	67.0
	EU_15	29.8	14.6	3.6	78.4
2010 Total		27.0	14.7	3.3	78.4

Highly educated workforce

year	Country group	Average	SD	Min	Max
2003	**EU_13**	14.6	5.8	6.5	30.1
	EU_15	22.3	7.7	6.1	43.0
2003 Total		20.6	8.0	6.1	43.0
2004	**EU_13**	15.5	5.9	6.8	31.1
	EU_15	23.5	7.6	6.6	43.8
2004 Total		21.8	8.0	6.6	43.8
2005	**EU_13**	16.2	6.0	7.5	33.2
	EU_15	23.9	7.8	7.7	45.5
2005 Total		22.3	8.1	7.5	45.5
2006	**EU_13**	16.8	6.1	8.0	33.2
	EU_15	24.4	7.8	8.2	45.8
2006 Total		22.8	8.1	8.0	45.8
2007	**EU_13**	17.3	6.4	7.3	33.3
	EU_15	24.8	8.0	7.4	47.6
2007 Total		23.2	8.3	7.3	47.6
2008	**EU_13**	18.1	6.6	6.8	34.5
	EU_15	25.4	7.9	7.2	48.3
2008 Total		23.8	8.2	6.8	48.3
2009	**EU_13**	19.0	6.6	8.4	36.1
	EU_15	26.3	8.2	8.2	51.5
2009 Total		24.7	8.5	8.2	51.5
2010	**EU_13**	19.9	6.8	9.0	35.7
	EU_15	27.0	8.5	9.9	53.1
2010 Total		25.4	8.6	9.0	53.1

Long-term unemployment

year	Country group	Average	SD	Min	Max
2003	**EU_13**	6.9	4.1	1.0	16.8
	EU_15	3.1	2.8	0.1	15.4
2003 Total		3.9	3.5	0.1	16.8
2004	**EU_13**	6.6	3.7	1.2	17.4
	EU_15	3.2	2.8	0.3	13.8
2004 Total		4.0	3.3	0.3	17.4
2005	**EU_13**	6.4	3.7	1.2	18.1
	EU_15	3.3	2.8	0.3	13.4
2005 Total		4.0	3.2	0.3	18.1
2006	**EU_13**	5.5	3.0	0.9	15.9
	EU_15	3.1	2.5	0.3	12.0
2006 Total		3.6	2.8	0.3	15.9
2007	**EU_13**	4.1	2.2	0.7	11.8
	EU_15	2.7	2.2	0.3	11.1
2007 Total		3.0	2.3	0.3	11.8
2008	**EU_13**	3.0	1.9	0.5	9.6
	EU_15	2.5	2.0	0.1	9.4
2008 Total		2.6	2.0	0.1	9.6
2009	**EU_13**	3.1	1.8	0.4	8.8
	EU_15	2.8	1.9	0.3	11.7
2009 Total		2.9	1.9	0.3	11.7
2010	**EU_13**	4.3	2.4	0.2	12.3
	EU_15	3.5	2.3	0.2	12.2
2010 Total		3.7	2.4	0.2	12.3

Continued on the next page ...

Employment

year	*Country group*	*Average*	*SD*	*Min*	*Max*
2003	**EU_13**	62.6	6.7	51.2	77.0
	EU_15	69.4	7.0	46.1	86.5
2003 Total		67.9	7.5	46.1	86.5
2004	**EU_13**	62.7	6.5	50.7	75.9
	EU_15	69.4	6.6	47.8	81.9
2004 Total		67.9	7.1	47.8	81.9
2005	**EU_13**	63.2	6.2	53.9	76.9
	EU_15	70.2	6.4	48.1	82.1
2005 Total		68.7	7.0	48.1	82.1
2006	**EU_13**	64.6	6.0	54.5	77.2
	EU_15	70.9	6.3	48.3	82.5
2006 Total		69.6	6.8	48.3	82.5
2007	**EU_13**	66.0	5.7	55.9	77.2
	EU_15	71.7	6.4	47.9	86.7
2007 Total		70.5	6.7	47.9	86.7
2008	**EU_13**	66.9	5.7	54.7	78.0
	EU_15	72.1	6.6	46.4	88.8
2008 Total		71.0	6.8	46.4	88.8
2009	**EU_13**	65.5	5.1	53.2	76.9
	EU_15	71.0	6.9	44.8	84.0
2009 Total		69.8	6.9	44.8	84.0
2010	**EU_13**	64.5	4.8	53.7	76.0
	EU_15	70.6	7.0	43.7	83.3
2010 Total		69.3	7.1	43.7	83.3

Innovation index*

year	*Country group*	*Average*	*SD*	*Min*	*Max*
2003	**EU_13**	-0.86	0.53	-1.67	1.09
	EU_15	-0.03	0.79	-1.87	2.43
2003 Total		-0.21	0.82	-1.87	2.43
2004	**EU_13**	-0.81	0.54	-1.65	1.22
	EU_15	0.08	0.78	-1.79	2.62
2004 Total		-0.11	0.82	-1.79	2.62
2005	**EU_13**	-0.74	0.59	-1.70	1.36
	EU_15	0.13	0.79	-1.64	2.78
2005 Total		-0.06	0.83	-1.70	2.78
2006	**EU_13**	-0.67	0.60	-1.51	1.43
	EU_15	0.19	0.79	-1.59	2.80
2006 Total		0.00	0.83	-1.59	2.80
2007	**EU_13**	-0.65	0.55	-1.39	1.34
	EU_15	0.19	0.79	-1.45	2.75
2007 Total		0.01	0.82	-1.45	2.75
2008	**EU_13**	-0.56	0.64	-1.52	1.56
	EU_15	0.24	0.80	-1.67	3.03
2008 Total		0.06	0.84	-1.67	3.03
2009	**EU_13**	-0.53	0.57	-1.36	1.28
	EU_15	0.28	0.81	-1.33	2.79
2009 Total		0.10	0.83	-1.36	2.79
2010	**EU_13**	-0.36	0.88	-1.51	2.73
	EU_15	0.37	0.93	-1.53	3.08
2010 Total		0.21	0.96	-1.53	3.08

* Note: the Innovation index is by construction expressed in z-scores

3-year average of annual real GVA per capita growth

3-y average period	*Country group*	*Average*	*SD*	*Min*	*Max*
2004_2006	**EU_13**	1.052	0.022	1.013	1.111
	EU_15	1.021	0.011	0.987	1.064
2004_2006 Total		1.027	0.019	0.987	1.111
2005_2007	**EU_13**	1.051	0.026	1.002	1.110
	EU_15	1.021	0.012	0.974	1.062
2005_2007 Total		1.027	0.021	0.974	1.110
2006_2008	**EU_13**	1.051	0.025	1.004	1.113
	EU_15	1.015	0.013	0.971	1.044
2006_2008 Total		1.023	0.022	0.971	1.113
2007_2009	**EU_13**	1.016	0.026	0.956	1.068
	EU_15	0.992	0.013	0.961	1.027
2007_2009 Total		0.997	0.019	0.956	1.068
2008_2010	**EU_13**	1.003	0.023	0.950	1.041
	EU_15	0.989	0.016	0.947	1.048
2008_2010 Total		0.992	0.019	0.947	1.048
2009_2011	**EU_13**	1.000	0.021	0.966	1.038
	EU_15	0.992	0.019	0.918	1.038
2009_2011 Total		0.994	0.019	0.918	1.038
2010_2012	**EU_13**	1.021	0.022	0.974	1.059
	EU_15	1.006	0.029	0.907	1.089
2010_2012 Total		1.009	0.028	0.907	1.089
2011-2013	**EU_13**	1.029	0.023	0.964	1.085
	EU_15	1.003	0.030	0.892	1.076
2011-2013 Total		1.009	0.031	0.892	1.085

Table A.2: CART robustness analysis for the EU28 (left column) and the EU15 (right column) scenarios

	EU28 scenario				EU15 scenario			
	3 classes (thresholds= P25%, P75%)				3 classes (thresholds= P25%, P75%)			
	Mean	Lower 95% CI	Upper 95% CI	Rank	Mean	Lower 95% CI	Upper 95% CI	Rank
Stage of development	0.66	0.65	0.66	3	0.32	0.31	0.32	5
Urban areas	0.35	0.34	0.35	5	0.30	0.30	0.31	6
Net foreign position	0.83	0.82	0.83	2	0.98	0.98	0.98	1
Governmt debt	0.24	0.23	0.25	8	0.35	0.34	0.35	4
Transport infrastructure	0.24	0.24	0.25	7	0.22	0.21	0.23	8
Quality of governance	0.37	0.37	0.37	4	0.46	0.46	0.47	3
Lowly educated workforce	1.00	1.00	1.00	1	0.98	0.97	0.98	2
Highly educated workforce	0.17	0.16	0.18	9	0.18	0.17	0.18	11
Long-term unemployment	0.25	0.25	0.26	6	0.23	0.23	0.24	7
Employment	0.12	0.11	0.12	10	0.20	0.19	0.20	10
Research and Innovation	0.09	0.09	0.10	11	0.20	0.19	0.21	9
MR	0.29				0.32			
	4 classes (thresholds=P25%, P50%, P75%)				4 classes (thresholds=P25%, P50%, P75%)			
	Mean	Lower 95% CI	Upper 95% CI	Rank	Mean	Lower 95% CI	Upper 95% CI	Rank
Stage of development	0.48	0.48	0.49	3	0.15	0.15	0.16	10
Urban areas	0.22	0.22	0.22	8	0.26	0.25	0.27	9
Net foreign position	0.63	0.63	0.63	2	0.95	0.94	0.95	2
Governmt debt	0.39	0.39	0.40	5	0.48	0.47	0.49	4
Transport infrastructure	0.19	0.18	0.19	10	0.32	0.32	0.33	6
Quality of governance	0.43	0.43	0.44	4	0.66	0.65	0.66	3
Lowly educated workforce	1.00	.	.	1	0.99	0.99	0.99	1
Highly educated workforce	0.26	0.26	0.26	7	0.30	0.29	0.30	7
Long-term unemployment	0.28	0.28	0.29	6	0.38	0.38	0.39	5
Employment	0.19	0.18	0.19	9	0.29	0.29	0.30	8
Research and Innovation	0.13	0.13	0.14	11	0.10	0.09	0.11	11
MR	0.39				0.42			
	5 classes (thresholds=P20%, P40%, P60%, P80%)				5 classes (thresholds=P20%, P40%, P60%, P80%)			
	Mean	Lower 95% CI	Upper 95% CI	Rank	Mean	Lower 95% CI	Upper 95% CI	Rank
Stage of development	0.56	0.55	0.57	3	0.20	0.19	0.21	10
Urban areas	0.19	0.18	0.20	10	0.31	0.30	0.31	8
Net foreign position	0.67	0.66	0.68	2	0.94	0.93	0.94	2
Governmt debt	0.38	0.38	0.39	5	0.54	0.53	0.55	4
Transport infrastructure	0.20	0.20	0.21	8	0.32	0.31	0.32	7
Quality of governance	0.56	0.55	0.56	4	0.71	0.71	0.72	3
Lowly educated workforce	1.00	.	.	1	0.99	0.99	0.99	1
Highly educated workforce	0.31	0.30	0.32	6	0.44	0.43	0.44	5
Long-term unemployment	0.24	0.24	0.25	7	0.36	0.35	0.36	6
Employment	0.20	0.19	0.20	9	0.26	0.26	0.27	9
Research and Innovation	0.16	0.15	0.17	11	0.14	0.13	0.15	11
MR	0.45				0.48			