This is a pre-print of an article published in the journal **Socio-Economic Planning Sciences**. The final authenticated version is available online at:

Bertoletti, A.; **Berbegal-Mirabent, J.**; Agasisti, T. (2022). Higher education systems and regional economic development in Europe: A combined approach. *Socio-Economic Planning Sciences*, 2022. DOI: 10.1016/j.seps.2022.101231.

Higher education systems and regional economic development in Europe: A combined approach using econometric and machine learning methods

Alice Bertoletti*, School of Management, Politecnico di Milano, Milan 20156, Italy

Jasmina Berbegal-Mirabent, Universitat Politècnica de Catalunya, Department of Management

Vilanova i la Geltrú 08800, Spain

Tommaso Agasisti, School of Management, Politecnico di Milano, Milan 20156, Italy

*Corresponding Author: e-mail: alice.bertoletti@polimi.it

Abstract:

This study deals with the analysis of regional economic development in Europe. Specifically, it examines the extent to which the performance and characteristics of higher education systems (HESs) influence regional economic development. The analysis employs data at the regional level, examining 649 NUTS-3 in 29 European countries, from 2014 to 2016. The empirical analysis, based on an original dataset that we developed, employs a novel methodological strategy that combines a traditional econometric approach with random forest. The findings detect the existence of nonlinear relationships between regional GDP per capita and HES indicators, which could have been overlooked by previous studies in the literature. Furthermore, the empirical results demonstrate the importance of comprehensively modelling the diversity of HESs, since distinct characteristics and performance can contribute differently to the economy of the regions. In particular, the most important factors for regional economic development are the size of HESs, the internationalisation of the students and research productivity. Finally, this paper provides useful insights for policymakers by suggesting new instruments for driving and fostering the economic development of their regions.

Keywords: higher education; regions; economic development; machine learning.

1. Introduction

Recent studies challenge the assumption that investments in advanced human capital and higher education (HE) can be effective for fostering social and economic development. For example, both the review by Holmes (2013) and the study by Hanushek (2016) indicate that skills and competences acquired at the end of compulsory education seem to encourage growth much more than pure measures of years of education. In times of tight public budgets and when policymakers are planning policies for sustaining economic recovery and development, this issue is of paramount importance for supporting decisions about the allocation of public resources. While existing knowledge provides many suggestions about the potential effects of HE (and human capital more generally) on country-level economic growth (see Wolf, 2002; Keller 2006; Yusuf and Nabeshima, 2006; Pelinescu, 2015; Hanushek and Woessman, 2020) there are two areas of investigation that need further substantial attention: (i) the role of universities for the local economy and (ii) the need to take better account of the heterogeneity of universities.

Firstly, a specific viewpoint that has policy relevance is the relationship between universities and local economy. Empirical evidence on the relationships between regional economic development and university systems is limited (Valero and Van Reenen, 2019; Amendola et al., 2020). At the country level, human capital (measured, for example, in educational attainment) can be generally associated with the national educational system. However, at regional level, where there are significant migration flows, this assumption does not hold anymore. It therefore becomes essential to understand the role of the universities per se in fostering local economic development. The extant literature presents still limited results that address this problem, and these studies are often focused on specific activities (or missions) of universities, without providing an overall picture of their effects (i.e. considering teaching, research and third mission activities).

Secondly, while the literature tends to overlook Higher Education Systems (HESs) heterogeneity in their empirical models, differences in terms of size, funding, and performance of HESs (within countries) are expected to affect the extent and the modes in which HESs influence the local economy (Santoalha et al., 2018). Indeed, HES diversity has grown increasingly during recent decades (Teichler, 2008; Rossi, 2010), driven by higher education policies, responding to labour-market needs, or following competitive and strategic behaviours (Dill and Teixeira, 2000). In this vein, modelling the role of HES in impacting economic growth makes little sense, while describing the specific features of universities

populating the HE domain is more appropriate for understanding its effects. In detail, the empirical analysis should describe the specific features of universities operating in each territory (for example, dedication to teaching, research and knowledge transfer) for exploring the channels through which they affect the economic development of the territory, rather than concentrating on single, unidimensional indicators of universities' activities. Some studies already supported the argument that microeconomic features of universities clearly reveal a high degree of heterogeneity, which must be properly considered (Daraio et al., 2011) also because this heterogeneity can be the result of specific strategic choices and background characteristics that can drive the economic impact of universities on the surrounding context.

The present paper addresses the two topics described above with the specific aim of comprehensively studying the relationships between HESs and economic development at regional level. More specifically, we examine the following research question: *How do the performance and the specific characteristics of HESs influence economic development at the regional level, in Europe?*

We address the above research question by studying the GDP per capita (i.e. the measure of economic development) of more than 650 NUTS-3 regions in 29 European countries, between 2014 and 2016. Based on the conceptual discussion formulated in this introduction, the paper tests the hypothesis that the specific, heterogeneous characteristics and activities of universities located in a region/territory have a statistically significant effect in explaining its economic development. The empirical analysis employs the GDP per capita as the indicator of economic development, as traditionally suggested in literature. This choice also makes high availability of data possible for measuring the dependent variable. On the other hand, we are fully aware that GDP per capita is an indicator of productive flows and measures market activities only (Stiglitz et al., 2009). Instead, the regional HESs studied here consist of all the academic and applied science universities operating in a given region (based on the definition of ETER database). Art academies, research institutes, military academies and conservatories have been excluded so that a homogeneous sample, which can be compared between one country and another, can be obtained. In this sense, when interpreting the results, we must be aware that the impact of the whole higher education sector could be higher than the one presented in the paper.

The analyses presented here develop the existing knowledge in four crucial directions. Firstly, the paper brings new empirical evidence into the debate on the role of universities in influencing the local economy, evidence which still scarcely offers quantitative estimates on

these relationships (Valero and Van Reenen, 2019; Amendola et al., 2020). Secondly, the analysis provides a more comprehensive approach in characterising HESs, making an explicit analysis of the universities' heterogeneity across HESs, by developing a framework of 15 indicators. Thirdly, the paper adopts a new methodology strategy that combines the robustness of econometric models with the flexibility of a machine learning model (random forest). These techniques represent an interesting solution to model complex contexts with many highly correlated factors, as is the case with the HESs analysed here. The use of machine learning in modelling the complex relationship between HE systems and local economic growth, moving beyond the more traditional approaches used in the recent contribution by Agasisti and Bertoletti (2020), explicitly considers the possible non-linearities of the relationship and the interactions with other contextual factors (see the discussion of machine learning use in the econometric framework presented by Mullainathan and Spiess, 2017). Lastly, the empirical analyses are based on a new dataset which is the result of substantial efforts in collecting and processing information from multiple sources (i.e. ETER, InCites, Eurostat and OECD Regional Database). Moreover, the dataset is more detailed than those already employed in previous studies because here the level of analysis is at NUTS-3 regions. Given the high heterogeneity of HESs, this level of analysis is particularly suitable as it provides precise estimates of the relationships between regional HESs and local economic performance.

The paper is structured as follows. Section 2 contains a review of the key literature examining the relationships between HESs and regional economic development. This section also includes a discussion on the main contributions in the field of higher education diversity, which is considered of central importance here for modelling HESs. Section 3 presents the methodological approach and the empirical framework adopted, while the data and the choice of variables are described in Section 4. Finally, Section 5 contains the main results, which are then discussed (along with research and policy implications) in Section 6.

2. Literature review

2.1. The relationships between HESs and regional economic development

A large number of contributions have studied the role of knowledge in influencing economic development. Most of these works have focused on a specific channel through which educational systems can contribute to economic outputs and the creation of human capital. These studies have been mainly developed in the field of Human Capital Theory and are

based, in particular, on the endogenous economic model presented by the New Growth Theory (see Lucas, 1988; Romer, 1990; Mankiw et al., 1992), which identifies knowledge as one of the three main determinants of economic development (together with capital and labour). This framework has motivated numerous macroeconomic works to empirically investigate the effect of human capital on the economic outputs of nations and, more recently, of regions. The evidence in the field of regional development confirms the results found at national level, revealing a positive effect between human capital and regional economic development (see, for instance, Mankiw et al., 1992; de la Fuente and Doménech, 2006).

The role of universities in fostering local economic development has also been addressed by works in the field of regional studies. Indeed, despite the great attention given to the role of human capital, HESs can influence economic outputs through two additional main mechanisms (Agasisti and Bertoletti, 2020). By providing R&D spillovers, universities can foster local innovation that, in turn, is considered to have a positive impact on the economic performance of regions (Bramwell and Wolfe, 2008; Diebolt and Hippe, 2019; Horváth and Berbegal-Mirabent, 2020). At the same time, universities are known to act as labour-intensive enterprises (Yen et al., 2015) which produce direct economic effects on the local economy by creating a new demand for local goods and services (Hermannsson et al, 2017; Steinacker, 2005). However, the literature studying the overall economic contribution of universities is only recent and presents partial evidence (Valero and Van Reenen, 2019; Amendola et al., 2020).

The few contributions that empirically analyse the relationships between regional higher education systems and local economic outputs focus mainly on regions belonging to one country (Agasisti and Bertoletti, 2020). Focusing on single countries, these works can exploit the greater availability of data of national databases, but they are lacking in providing evidence valid in an international framework. In particular, considerable attention has been given to USA regions. Lendel (2010) examined the effect associated with the very presence of research universities on the economic outputs of US metropolitan statistical areas. The results of the work revealed that HESs with at least one research university generate positive and significant effects on regional economic outputs (in terms of employment and GDP per capita). Similar results were found by Cermeño (2019), who showed that the establishment of a new university produced an increase between 1% and 3% in the annual GDP of US counties, from 1930 to 2010. Besides the link between the presence of universities and economic outputs, qualitative dimensions of US universities have started to be investigated by

Goldstein and Drucker (2006). By adopting a quasi-experimental approach, a key contribution of this study is a deliberate separation of the regional economic impacts of different university activities. The empirical results validate the hypothesis that the university activities of teaching (proxied by the number of degrees awarded), research (operationalised by R&D expenditures), and technology development (number of patents) help to raise regional economic progress (measured in terms of regional average earnings). These effects were found to be particularly important in small- and medium-sized regions, while for larger regions economic growth was more dependent on non-university factors such as business services and starting employment level.

A multidimensional framework of education indicators has been also used by Schubert and Kroll (2016) in their study on the relationship between German universities and both GDP per capita and the unemployment of NUTS-3 regions. Exploiting a national statistical database, the authors modelled universities through six indicators: number of students, investments, number of staff, number of publications, number of graduates and third party funds. The results indicated that not only regional characteristics play a role in explaining economic growth, but also that the universities' characteristics exert a decisive influence. On average, universities contributed €8,300 to regional GDP per capita in the period 2000 to 2011. All indicators were found to be significant; yet, universities that have well-established links with the business ecosystem, have a greater focus in science-related disciplines and are located in technologically strong regions make a greater contribution to the economic well-being of the region.

Some papers focus on specific characteristics of universities in contributing to local economic performance. Amendola et al. (2020) investigated the effects of the graduate human capital of universities on regional economic development of Italian provinces. Controlling for the mobility of students and endogenous regressors, the authors found that HESs positively affect the local economy, with stronger effects for graduates in technological fields and universities with a high reputation and located in the most-developed areas. Italian provinces were also investigated by Agasisti et al. (2019), who focused on the link between the efficiency of universities (their capacity to convert inputs into outputs) and regional GDP per capita. Their results showed that the presence of efficient universities in the regions positively influences local economic development. Finally, the link between UK universities performance and regional economic outputs was discussed by Guerrero et al. (2015), who empirically studied the gross value added of 74 NUTS-3 regions in the United Kingdom. Employing a structural

equation modelling approach, the authors considered three aggregated indicators of academic performance (i.e. teaching, research and entrepreneurial performance), finding larger effects associated with research and entrepreneurial activities.

Finally, the literature presents some contributions analysing the impact of universities on the labour market and economic development of peripheral regions, even if the results are based on single case studies (see Evers, 2019; Rossi and Goglio, 2020). The work of Evers (2019) highlights the role of universities in increasing the qualified human capital and wage growth of the peripheral regions, while satellite universities have been found to foster local economic development through research, community engagement and demand for knowledge-intensive services.

Moving the attention to multi-country studies, the literature offers a very limited number of papers. Lille and Roigas (2017) focused on the human capital produced by universities in European NUTS-2 regions, measured as the share of tertiary students, finding however limited effects associated with this indicator. Instead, Valero and Van Reenen (2019) analysed the effect of the number of universities operating in the HESs on GDP per capita of 1,500 NUTS-2 regions across 78 countries. Their results showed that a 10% increase in the number of universities in the region was able to produce on average a growth in GDP per capita of around 0.4%. Only the work of Agasisti and Bertoletti (2020) represents a first attempt at using a multidimensional set of indicators for modelling HESs in an international analysis of regional development. Studying 284 NUTS-2 regions in Europe, the authors proved the importance of employing a suitable model of indicators for measuring higher education systems. Indeed, the results show that the presence of universities in the regions captures only partially the economic impact of HESs, which is strongly influenced by the size, research outputs and subject specialisation of the universities in the region. However, a comprehensive characterisation of HESs has not yet been fully achieved. Although the employment of a rich set of HESs variables, the role of some key features remains unexplored, such as mission-orientation, sectoral typology, university internationalization and resources. The purpose of the present paper is to fill this gap by focusing on the role of the heterogeneity of the universities in influencing regional GDP, rather than modelling the interaction between HESs and economic growth (as in Agasisti and Bertoletti, 2020). In pursuit of this purpose, a primary element of novelty is the adoption of a methodology that relaxes the linearity assumption and can handle a high number of covariates.

The review presented above underlines the central issue of the literature which lacks empirical proofs on the overall contribution of universities to local economic performance. Even if some empirical studies provide partial evidence on the relevance of the heterogeneity of universities in influencing local economic outputs, a comprehensive picture is still missing. This lack is more evident when an international context is adopted since cross-country data is particularly limited. It follows that literature is also limited in providing a general model explaining the relationships between HESs and economic development in an international context.

In view of this, we argue that the extant literature only provides a partial representation of the relationships between regional HES and economic development of the territory (Chatterton and Goddard, 2000), requiring new research efforts able to effectively deal with and address these shortcomings, particularly as regards how to adequately capture the heterogeneity of HESs in this analysis. The present paper aims at overcoming these limitations by developing a framework of 15 indicators that refers directly to the literature on higher education diversity (see Section 4.1). The relationships of this comprehensive set of HESs and local economic development variables are studied for 29 European countries. In this way, we estimate a communal model for Europe and thus provide international evidence. The analysis is supported by the use of a novel methodological approach that allows the high-dimensional problem introduced by including a large number of the HES covariates to be addressed.

2.2.The importance of modelling HES diversity

The review of the literature presented in Section 2.1 highlights the importance of fully characterising HESs. The extant works in the literature tend to provide evidence generally on the average contribution of universities in the regions, without fully considering differences in the quality of HESs and without indicating which characteristics and performance are more likely to be associated with larger economic outputs[1]. However, the quality and way through which universities carry out their main activities, teaching, research and third mission (Martin, 2012), are likely to significantly affect the economic output of regions and cannot be ignored. Omitting differences in the quality of education may provide distorted estimates of the economic impact of universities (Hanushek and Wößmann, 2010). For instance, HESs

[1] As presented in Section 2.1, it is worth noticing that multidimensional frameworks of university indicators are reported in the paper of Goldstein and Drucker (2006), Schubert and Kroll (2016), and Agasisti and Bertoletti (2020).

with excellent teaching performance are likely to generate highly qualified human capital[2] that, in turn, is expected to produce larger effects on the economic outputs. Indeed, several studies (see for instance Hanushek and Kimko, 2000; Jamison et al., 2007) found that the quality of human capital has a much larger effect on economic development than the quantity of schooling. Moreover, the research performance of HESs, in terms of quantity and quality of publications, can significantly affect the economic development of regions, sparking knowledge spillovers and innovation processes (Denti, 2010; Barra et al., 2021). In fact, empirical evidence proving the positive effect of research quality on regional innovation was found by Hegde (2005) and Malva and Carree (2013). Third-mission performance, involving academic activities aimed at engaging with the society and generating knowledge transfer, may also contribute considerably to local economic development. In particular, the intensity of university-industry collaboration is seen as a core driver of innovation (Mueller, 2006; Diebolt and Hippe, 2019).

The quality through which universities conduct their activities is not the only element of HES heterogeneity that should be considered. Besides their performance in teaching, research and third mission, HESs can differ from each other in particular characteristics that define their uniqueness in terms of programmes, missions, funding, etc. (Agasisti and Berbegal-Mirabent, 2021). In recent decades, HESs have continued to increase their diversity (Teichler, 2008; Rossi, 2010), fostered by higher education policies, responding to the needs of the labour market, or as an effect of competitive and strategic behaviours (Dill and Teixeira, 2000). Accordingly, the literature on higher education diversity has been extensively developed (Teichler, 2010), covering both national and regional contexts (Santoalha et al., 2018). Based on these studies, we can distinguish between internal diversity, which represents differences within higher education institutions (HEIs), and external diversity, referring to the differentiation between institutions (Birnbaum, 1983; and Kivinen and Rinne, 1996). In modelling the diversity between HESs, we are therefore interested in representing the external diversity that characterises each university in the system. According to Kivinen and Rinne (1996), there are 7 main dimensions through which external diversity can be realised: (i) sectoral diversity, in particular between vocational oriented colleges and universities; (ii) differentiation in their missions, such as orientation to research, vocational courses or postgraduate programmes; (iii) curricular and programmatic diversity, which refers to differentiation in terms of subjects, fields of studies and degree levels offered; (iv) diversity in

[2] It is worth noticing that considerable interregional migration flows may produce relevant distortions when a regional framework is adopted (Abel and Deitz, 2009).

the duration of courses; (v) geographical distribution; (vi) diversity in the communities served; and (vii) type of funding.

These features may play a relevant role in increasing the economic development of the territory (Santoalha et al., 2018). For example, studies in the field of science or STEM are more suited to empirical applications and are expected to positively affect local economic performance (see Becher and Trowler, 2001; and Xie et al., 2015; Agasisti and Bertoletti, 2020). Furthermore, the presence of prestigious and research-oriented universities seems to generate positive effects on the economic outputs of UK regions (Guerrero et al., 2015; Amendola et al., 2020). Moreover, local GDP per capita is likely to be influenced by differences in terms of the legislative typology and strategy orientation of higher education systems (Schubert and Kroll, 2016). For instance, medical schools are usually associated with low performance and efficiency in their technology transfer activities (Thursby and Kemp, 2002; Anderson et al., 2007).

Finally, the characteristics and the performance of HESs, not only can significantly influence the impact of universities on the local economy, but they are also likely to interact with the environment in which they operate. Indeed, regional factors can represent important determinants for HES performance (see Agasisti and Bertoletti, 2019), and universities may develop specific features as a response to regional needs (Santoalha et al., 2018).

From the above, it can be concluded that it is of paramount importance to include a wide set of indicators that adequately capture the main features and the performance of HESs, otherwise, any attempt at examining HESs contribution to regional economic development will suffer from being over simplistic and consequently provide an incomplete view. In our opinion, if this picture is still missing, or it is only partially provided by the studies in the literature, it is due to two main limitations. Firstly, the scarce availability of data for measuring HESs, especially for multi-country analyses, could have discouraged the use of complex systems of indicators and secondly, the methodologies employed for investigating the economic impact of HESs. The majority of the studies adopt traditional econometric models to estimate the economic contribution of universities. These approaches seem adequate when a simple set of educational indicators is used. However, they may be inappropriate for modelling complex systems of HES variables, which are likely to interact with each other and with the regional environment.

3. Empirical framework and methodology

The empirical analysis conducted in this paper is based on two complementary approaches. As a first stage, we adopt a traditional econometric strategy (i.e. a regression analysis and a system Generalised Method of Moments, system-GMS hereafter) which is particularly suited to dealing with the issue of causality and the potential endogeneity of the regressors. The results of the econometric model also represent a reference through which to compare the analyses of the second stage, which is instead based on machine learning (ML). The ML approach is employed to represent the complexity (interaction and nonlinearity) of the relationships between regional HESs and the local economic performance. A full modelling of the characteristics and the performance of HES requires a methodology able to handle a high number of covariates, which are likely to interact between each other and that usually co-exist in the same environment (see the correlation analysis in Section A1). ML can meet these needs thanks to its high flexibility. In particular, the random forest model we adopt in the second stage makes it possible to handle the presence of many and high-correlated covariates and to model their respective interactions. Moreover, machine learning techniques are based on a nonparametric approach and do not force any linear relationship between the outcome and the covariates. In this way, we can relax the linearity assumption that is implied by econometric models which impose a parametric functional relationship between the independent variables and the response. This feature of ML is particularly interesting in the case of our study, since there is no strong *a priori* reason to assume the existence of a linear relationship between educational outputs and economic development (Krueger and Lindahl, 1998, Sianesi and Van Reenen, 2003). In this way, the second stage aims at uncovering potential relationships between HES characteristics and regional economic development which may be hidden by the adoption of parametric models. Contrary to the first-stage estimates, it is worth clarifying that the analyses in the second stage do not aim to deal with reverse causality, which would be instead hard to control by employing ML techniques. Thus the findings emerging from the ML approach must be interpreted as correlational, rather than causal. The joint adoption of the econometric and ML approaches allows compensation for the shortcomings of each methodology and provides robust results. While insight on the causal effects is generated by the GMM approach, the random forest technique allows nonlinear relationships to be estimated and a high number of correlated regressors to be handled.

In both stages, the empirical models employ the same set of covariates, examining their effect on the level of regional GDP per capita. We excluded employing an indicator of economic growth as dependent variable due to the short time-span available (and this opens the possibility for further research in the future). To use a wide set of variables on HES characteristics, we had to focus the analyses on a four-year period, in which ETER[3] (i.e. our source for HES indicators) data are available.

3.1. First stage – Econometric model and the Generalized Method of Moments

The parametric model we estimate in the first stage is expressed by equation (1) and is based on the general model of aggregate production function in levels (De la Fuente and Doménech, 2006).

$$(1)\ log(Y_{ic,t}) = \alpha_1(HES_{ic,t-l}) + \alpha_2(Empl_{rate}{}_{ic,t-l}) + \alpha_3(Net_{mig_r}{}_{ic,t-l}) \ \alpha_4 \log(Density_{pop}{}_{ic,t-l}) + \alpha_5\, log(K_{ic,t-l}) + \alpha_6(Hc64_{n2}{}_{ic,t-l}) + \gamma_c + \tau_t + \varepsilon_{ic,t}$$

Where $Y_{ic,t}$ is the GDP per capita in NUTS-3 region *i*, in the country *c* and at year *t; HES* represents a vector of variables on the characteristics of the regional higher education systems and the respective performance (see HES variables in Table 1 for details). Besides the *HES* vector, containing our variables of interest, we include different controls at NUTS-3 level. We use the employment rate of the region ($Empl_{rate}$) to take into account the labour in the area (see for example De la Fuente and Doménech, 2006 and Gennaioli et al., 2014) and we employ the interregional net migration rate (Net_mig_r) as control for the mobility of human capital (see Boucher et al., 2003 and Gennaioli et al., 2014). We also include the population density ($Density_{pop}$), as control for the presence of capital cities or important business cities in the regions (Cuaresma et al., 2014; Fournier, 2016)[4]. Two control factors are employed at NUTS-2 level since data at NUTS-3 are not available: the gross fixed capital formation (*K*), used as control for the stock of physical capital in the region (see for example Marrocu and Paci, 2010; De la Fuente and Doménech, 2006), and regional human capital ($Hc64_{n2}$), expressed as the share of population with higher education (Gennaioli et al., 2014; Valero and Van Reenen, 2019), representing the stock of qualified human capital. The last three variables in equation (1) are the country fixed effects (γ_c), the time dummies (τ_t) and the error terms

[3] European Tertiary Education Register, see section 4.2 for details.
[4] Since NUTS are based on homogenous levels of population, when the population density is high, regions have a small area that can include only one city or even just a part of the city (e.g. London).

($\varepsilon_{ic,t}$). The country fixed effects are included following the work of Gennaioli et al. (2014) and allow control for country-level factors, such as cultural and social characteristics, national economic performance and features deriving from national structure of HESs. However, we do not use regional fixed effects as suggested by De la Fuente and Doménech (2006). The rationale behind this decision is that the influence of cultural and social characteristics is considerably smaller within countries than the one existing between nations (Gennaioli et al., 2014).

Since determinants are supposed to need some time to reveal their effects on the economic performance, all the regressors are included in the model with a 2-year lag. Longer time lags are not considered so as not to reduce the sample size excessively, given the short time-span of our available data.

Equation (1) is estimated by employing regression analyses and the system Generalized Method of Moments (sys-GMM) (Arellano and Bond, 1991). A sys-GMM approach is adopted in order to control for the endogeneity of the independent variable that could affect empirical estimates (Ullah et al., 2018). In fact, the regressors included in equation (1) and, in particular, the variables on regional HESs, can be influenced, in turn, by the economic level of the region. For example, the establishment of new universities in a certain area can be fostered by the demand for high-skilled people in highly developed regions. Moreover, the economic development of regions may also encourage the presence of universities with specific characteristics. For instance, richer regions are likely to foster the establishment of private universities in the area and can more easily finance higher education activities.

GMM allows controlling for the endogeneity of the regressors by differencing equation (1) and using the lags of the independent variable as instruments. System GMM is an augmented version of the GMM estimator developed by Arellano and Bover (1995) and Blundell and Bond (1998). The authors found that the efficiency of the GMM improves when the original equation in levels is added to the model and, therefore, first-difference lagged variables can be included to instrument their own levels. In this way, sys-GMM can be represented through a system of equations (one for each period considered) which uses different sets of "internal" and "external" instruments. The validity of the results can be verified by estimating the Arellano-Bond test for no autocorrelation, which checks the assumption of serial independence in the original errors (the assumption is verified when the differenced residuals do not show autocorrelation of the second order). The Hansen test of over-identifying

restrictions instead allows the assumption of exogenous instruments in the case of heteroscedasticity or autocorrelation in the data (Roodman, 2018) to be verified.

3.2. Second stage – Machine learning approach and Random forest

The second stage of empirical analyses is based on a machine learning approach and, in particular, on the random forest (RF) technique (see James et al., 2013).

RF is grounded on the regression tree method, which consists of a recursive algorithm that starts by splitting the covariates space into two regions (rectangles) and it models the response by the mean of the depending variable in each region (Hastie et al., 2009). The algorithm will choose automatically the dependent variable and the "split point" associated with the best fit. In other words, the parent node is split into two descendent nodes depending on the independent variable selected. Then, this tree-growing process continues by splitting each region into other two regions and choosing the most suited variable among the independent variables that have still not be used in the tree. This algorithm of binary partition is modelled by equation (2):

$$f(x) = \sum_{m=1}^{M} c_m I\ (x \in R_m) \quad (2)$$

Where M is the number of regions through which the space is parted, x is a vector of inputs and c_m is a constant value for each region that models the response (y).

Since our dependent variable is continuous, we estimate a regressor tree and therefore use Mean Squared Error (MSE) as the criterion to decide the binary partition:

$$\sum\ (y_i - f(x_i))^2 \quad (3)$$

In this case, the best c_m is calculated as the mean of the dependent variable in the responses y_i in the region R_m.

The procedure here described stops when the stopping rule is satisfied and the tree is thus pruned. The dimension of the tree is based on minimisation of the cost-complexity criterion that prevents overfitting. At the end of the process, the relationships between regressors and the dependent variable are provided by the analysis of the subgroups of observations in the leaf nodes generated by the last partition (Lemon et al., 2003).

RF is an evolution of trees, since it provided an ensemble predictor based on multiple regression trees. In particular, this machine learning technique starts with generating many training sets deriving from subsets of original data and builds bootstrap samples for each

training set (Breiman, 1996, 2001). A regression tree is then computed for each training set using MSE as a splitting criterion. The final estimation is given by the average of the predictions calculated for each training set separately. At each node, the RF algorithm considers a subsample of regressors, making it possible to reduce the model variance and to handle the presence of many and highly correlated covariates (Breiman, 1996, 2001). By subsampling the predictors, all variables have indeed the possibility of being considered in the tree splits, limiting the risk that the effects of some covariates are covered by ones of more significant variables in the sample (Hastie et al., 2009). Thanks to this feature, ML can handle a high number of highly correlated covariates that represent one of the main advantages of using this particular technique for estimating the influence of multiple HES factors on local economic development.

As output, random forest provides a ranking of the regressors, based on the respective importance in explaining the dependent variables (Shi and Horvath, 2006). In the case of the analysis, we compute %IncMSE as a measure of variables' importance. The indicator is usually employed for continuous dependent variables and represents the percentage increase in the MSE, generated by excluding a specific regressor from the sample. The estimate of the variables' importance represents a relevant feature when nonlinear relationships are modelled. Indeed, since the linearity assumption is relaxed, the interpretation of the relevance of regressors is not straightforward, as in parametric analysis. The possibility of estimating variables' importance was the reason behind choosing the random forest over other nonparametric and semiparametric models which do not provide this information.

Another significant advantage of trees and RF is the availability of graphics tools which facilitate interpretation of the results. Indeed, given that the functional form is not set *a priori*, these representations provide information on the type of relationship (e.g. linear, polynomial or complex) between a regressor and the dependent variable. In this paper, we computed partial dependence plots and joint partial plots. These graphics tools show how the dependent variable partially depends on values of a regressor (or two regressors in case of joint partial plots), averaging out the effects of the other independent variables in the model (Hastie et al., 2009).

RF can also handle the presence of missing data in the predictors, without having to eliminate all the observations with missing values that could lead to a serious depletion of the training set, especially when a large set of variables is studied. When the covariate considered for the

split has missing values, the tree employs only the observations for which the predictor is not missing, building thus a "surrogate variable".

The RF advantages, here described, clarify the choice of random forest for estimating nonlinear relationships in the second step of our analysis over other parametric or semiparametric approaches. Nevertheless, in Section A3 of the Annex, we provide additional estimates generated by employing a semiparametric model (i.e. Generalized Additive Model) to offer evidence on the inference of the nonlinear effects.

4. Data and choice of HES variables

4.1. Modelling and measuring the characteristics and performance of HESs

HESs are not unidimensional entities but represent complex realities in which several factors coexist and interact (Bonaccorsi and Daraio, 2007). Accordingly, the diversity among HESs can occur along several dimensions (Rossi, 2010). Only by fully modelling this multidimensional heterogeneity is it possible to comprehensively characterise HESs and then provide a robust estimate of their local economic contribution.

To this end, this study proposes a framework of dimensions and indicators to model HESs and their heterogeneity (see Table 1). The framework is based on the taxonomy of HES diversity provided by Kivinen and Rinne (1996) (see Section 2.2), which is adapted to the specific aim of this work by incorporating the evidence in the field of regional development. In particular, we identify 10 dimensions and 15 indicators, grouped in three main areas, namely the institutional characteristics and HES size, programmatic characteristics and resources and HESs performance. The first group of dimensions represents the basic characteristics of HESs and refers mainly to the sectoral and funding diversity identified by Kivinen and Rinne (1996). In particular, we consider the typology of the universities with a focus on the share of applied science universities (Schubert and Kroll, 2016) and medical schools (Anderson et al., 2007; Amendola et al., 2020). Moreover, we distinguish between universities funded publicly and privately. The interest in the differential role of public and private universities on the territory has already been explored in the literature (see for instance Casani et al., 2014; Guironnet et al., 2018; Lepori, 2021). In this study, we proxy it by taking into account the share of students enrolled in private universities over the total number of students in the system (Agasisti and Bertoletti, 2020), under the assumption that public funding intensity is lower for higher proportions of students in private institutions (direct and complete financial

information about the different revenue sources are not available at the NUTS-2 level). Nevertheless, the variable (named public-private structure) captures a systematic source of variation in the level of public funding, and also includes some elements of the public/private structure of the HE system as a whole. Therefore, similar to Horváth & Berbegal-Mirabent (2020), this variable also allows us to analyse the knowledge spillover capability of public and private universities in the region. As a basic characteristic, we also consider the size of HESs, measured both in terms of the number of institutions (Valero and Van Reenen, 2019) and the total number of students (Lille and Roigas, 2017).

The second group of dimensions aims at capturing the differentiation in terms of the programmes offered and the resources employed by universities in the HESs. More specifically, we consider the diversity in programmes and subjects by measuring the share of students in STEM disciplines (Xie et al., 2015; Agasisti and Bertoletti, 2020) and the presence of doctoral programmes (Agasisti and Bertoletti, 2019). This area of the framework also captures the differences in terms of 'community served' and 'mission' (Kivinen and Rinne, 1996), with a focus on internationalisation which is measured as the share of students taking part in Erasmus programmes. The Erasmus initiative represents one of the most relevant instruments to support the internationalisation strategies of European HEIs (Teichler, 2009). Therefore, although the indicator is not comprehensive of all mobility initiatives, it represents a sound driver for the internationalisation of students and, indirectly, of academic staff. The intensity of Erasmus mobility is indeed the result of international agreements between higher education institutions, which are based on the professional contacts of professors and thus it indirectly reflects the internationalisation of the academic staff (Restaino, 2020). In this paper, we do not distinguish between the educational levels of the Erasmus students since disaggregated data on this indicator is scarcely available. Nevertheless, we are aware that Erasmus students enrolled on the master programmes could potentially present a different behaviour compared to those in bachelor's programmes. For instance, students attending master's courses could pay more attention to the career opportunities provided by the universities in the region. On the other hand, the measure aims at capturing the general effort of universities in building international networks rather than reflecting students' choices regarding their periods abroad. The resources available for delivering academic activities are instead captured by the mean of student-teacher ratios associated with the universities in the region (Agasisti and Bertoletti, 2019; Amendola et al., 2020). This indicator represents the

physical resources of the HESs. Measures for financial resources are not included due to the absence of a large quantity of data.

The last group of dimensions in Table 1 refers to the performance in the three activities of HESs, namely teaching, research and third mission (Martin, 2012). The indicators used to measure them are based on the framework proposed by Agasisti and Bertoletti (2019). More specifically, we employ a measure of graduation rate for teaching performance, built as the ratio between graduates and enrolled students in a given year[5]. Additional measures on the quality of teaching (such us credits passed or detailed information by cohorts) are unfortunately not available for the context of our analysis. In particular, the international university ranking reporting, an evaluation of teaching activities, refers only to a selected number of institutions and offer data for limited periods. While some institutional experiments aimed at measuring the quality of teaching in a more direct way – see the project AHELO by OECD as described in Dias and Amaral (2014) – they have been abandoned and no standardized information about quality differences across graduates exist. Concerning research performance, the number of publications per researcher has been used to measure the research productivity in quantitative terms, whereas the quality of research activities is approximated by the top 10% of most highly-cited publications (as classified by InCites). The indicator reflects the traditional measures of the quality of research since it is based on the number of citations of the publications (see Table 1). In addition, the literature suggests that the share of top cited publications can be considered as a realistic indicator for measuring research excellence (Tijssen et al., 2002; Barra et al., 2019). As an additional measure of research performance, we employ the share of publications with international co-authors, which leads to larger visibility of the publications (Puuska et al., 2014) and, in general, to a greater citation impact (Polyakov et al., 2017). Nevertheless, even if the indicator can be associated with a higher number of citations in some specific research fields (Sooryamoorthy, 2009), the literature has empirically demonstrated that international collaboration is structurally different from citation-based indicators (Schmoch and Schubert, 2008). In the present paper, international co-authorship aims at capturing the expertise of researchers to create academic networks, reflecting the scholars' ability to analyse data from different international sources (Polyakov et al., 2017). Finally, third-mission performance is

[5] This measure of teaching performance assumes that there are no fluctuations in the size of students' cohorts across years. The choice is justified by limited data comparability and availability in the years before the ones selected for the empirical analysis. Therefore, the present paper considers a short panel (mostly, 2011-14) so the assumption of steady state can be considered reasonable within this time frame.

represented by the percentage of publications with at least one co-author from the industry. The indicator is meant to reflect the knowledge transfer activities of universities, without pretending to be informative on the intangible outcomes ascribed to the third mission. This indicator is extensibly used in worldwide rankings (e.g. U-Multirank, CWTS Leiden) as a measure of knowledge transfer, as it captures the efforts resulting from the co-creation and application of new knowledge derived from a joint collaboration between a university and an industrial partner. Recent studies have also relied on this metric to account for third-mission outputs (see for instance, Agasisti and Bertoletti, 2020; Albats et al., 2018; Tijssen et al., 2016). The assumption behind the use of this indicator is that intense collaboration is behind the publication of joint works between academics and personnel from industries, an intuition that is corroborated by the academic literature in the field (see the interesting and complete discussion in the recent paper by Pohl, 2021). Despite intangible forms of university-industry collaboration which can foster local economic development significantly, these activities are difficult to capture and measure, especially in an international setting (Molas-Gallart, 2002). Similarly, more commercial outputs emerging from university third-mission activities are also not available at NUTS-3 level. However, this does not represent a crucial limitation since patents and licences capture only the smallest value of the overall knowledge transfer of universities (Agrawal and Henderson, 2002; Perkmann et al., 2011).

HES Areas	HES Dimensions	Indicators
Institutional characteristics and HES size	Sectoral and typology	Applied science universities; Medical Universities
	HES size	Number of universities; Students enrolled
	Public-private structure	Share of students in private universities
Programmatic characteristics and resources	Programmes and subjects	Doctoral students enrolled; % STEM
	Students internalization	Erasmus students incoming; Erasmus students outgoing
	Resources	Students/teacher ratio
HESs performance	Teaching performance	Graduation rate
	Research performance	Publications per researcher; Top 10% documents; International collaboration
	Third-mission performance	Industry collaboration

Table 1. Framework of indicators and dimensions of HES diversity./ *Source:* Produced by the authors from contributions in literature.

4.2. Data and descriptive statistics

The paper employs data at the regional level, examining more than 649 NUTS-3 in 29 European countries of which 27 are EU members (Romania is excluded due to a problem of data availability) plus Norway and the UK. The 649 regions analysed in the paper are about 45% of the total NUTS 3 in the 29 countries studied. The other 782 regions do not contain any universities (see the specific definition below) and therefore are not investigated in the paper.

The estimates are based on a novel dataset built for the specific purpose of the paper by gathering information from multiple sources. Data on universities is available in the European Tertiary Education Register (ETER) that collects information on 3,000 HEIs in Europe. ETER provides data on the characteristics of higher education institutions and their geographical position, as well as on their educational activities, staff, finances and research activities. The relevance of this database is proved by the large number of papers employing ETER data for empirical modelling of HEIs characteristics and performance (see, for example, Vieira and Lepori, 2016 and Santoalha et al., 2018). Data on the productivity and the quality of the academic publications are collected instead from the InCites database which gathers information on all documents published in Web of Science. Lastly, for indicators on regional economic performance and regional characteristics, we use the data available in Eurostat and OECD Regions and Cities databases.

Table 2 reports the description of the variables included in the empirical models, together with the respective time-span, geographical level and data source.

	Variable name	Description	Time-span	Geographical level	Databases
Dependent Variable	GDP per capita	Regional gross domestic product expressed as PPS per inhabitant by NUTS-3 regions.	2014-2016	NUTS 3	Eurostat
HES Variables	Number of universities	Number of institutions in the region	2011-2013	NUTS 3	ETER
	Students enrolled	Number of students enrolled in ISCED 5 - ISCED 7 programmes	2011-2013	NUTS 3	ETER
	Share of students in private universities	Share of students (ISCED levels from 5 to 7) in private universities over the total number of students	2011-2013	NUTS 3	ETER
	Medical Universities	Share of medical universities, over the total number of universities in the region	2011-2013	NUTS 3	ETER
	Applied science universities	Share of applied science universities in the region	2011-2013	NUTS 3	ETER
	Student/teacher ratio	Number of academic staff (expressed as head count) over the total number of students (ISCED levels from 5 to 7).	2011-2013	NUTS 3	ETER
	Graduation rate	Share of graduated students over the total number of students enrolled (ISCED levels from 5 to 7)	2011-2013	NUTS 3	ETER
	% STEM	Share of enrolled students (ISCED levels from 5 to 7) in STEM disciplines. STEM subjects include natural sciences, mathematics and statistics, engineering, manufacturing and construction and Information and communication technologies	2011-2013	NUTS 3	ETER
	Doctoral students enrolled	Number of PhD students enrolled (ISCED level 8 programmes)	2011-2013	NUTS 3	ETER
	Erasmus students incoming (%)	Number of incoming Erasmus students over the total number of students (ISCED levels from 5 to 7)	2011-2013	NUTS 3	ETER
	Erasmus students outgoing (%)	Number of outgoing Erasmus students over the total number of students (ISCED levels from 5 to 7)	2011-2013	NUTS 3	ETER
	Publications per researcher	Number of documents in web of science published by the universities in the region divided the number of academic staff (sum of the staff in head count of the universities in the region)	2011-2013	NUTS 3	InCites
	Top 10% documents	The number of documents in Web of Science in the top 10% based on citations by category, year, and document type (mean based on the number of documents per each university)	2011-2013	NUTS 3	InCites
	Industry collaboration	Percentage of publications that have co-authors from industry (mean based on the number of documents per each university)	2011-2013	NUTS 3	InCites
	International collaboration	Percentage of publications that have international co-authors (mean based on the number of documents per each university)	2011-2013	NUTS 3	InCites
Controls	Fixed capital	Gross fixed capital formation of the region, expressed in million euros. It is calculated as resident producers' acquisitions, less disposals, of fixed assets, plus certain additions to the value of non-produced assets realised by the productive activity of producer or institutional units.	2011-2013	NUTS 2	Eurostat
	Employment rate	Number of people employed in the region (all NACE considered) divided the population of the region	2011-2013	NUTS 3	Eurostat
	Human Capital	Share of population (25-64 years old) in the region with higher education	2011-2014	NUTS 2	Eurostat
	Net migration rate	Inter-regional net flows mobility rate, calculated as a percentage of net flows over population	2011-2014	NUTS 3	OECD Regional Database
	Population density	Population density by NUTS-2 regions, expressed as inhabitants per km^2.	2011-2013	NUTS 3	Eurostat

Table 2. Descriptions of the variables employed in the empirical analyses./ *Note*: Produced by the authors.

ETER and InCites data, which is available at institutional level, are aggregated at regional NUTS-3 level following a systematic process (see Section A2 in the Annex for details). Among the institutions reported in ETER, we consider only the ones defined as "university" (category 1) and "university of applied science" (category 2). The "other" institutes, representing mainly art institutes, research institutes, military academies and conservatories, are excluded due to their high heterogeneity. We also exclude online universities, since they do not operate in a physical place and their students and staff can be largely spread over national and international territory.

The descriptive statistics of all variables employed in the analyses are reported in Table 3. The maps in Figure 1 provide a spatial representation of regional differences in terms of GDP per capita and number of universities per region.

Variable	Obs.	Mean	Std. Dev.	Min	Max
GDP per capita	2,410	30678.76	22344.32	6200	395300
Number of universities	2,586	2.45	3.52	1	42
Share of private students	2,477	27.34	41.77	0	100
Medical Universities (%)	2,586	19.32	32.76	0	100
Applied science universities (%)	2,554	35.21	42.33	0	100
Students enrolled	2,481	24139.68	34594.81	0	478933
Graduation rate	2,447	24.74	10.47	0	103.14
STEM (%)	1,877	26.07	16.83	0	100
Doctoral students enrolled	2,349	953.8	1656.29	0	14467.7
Erasmus students incoming (%)	2,477	1.15	1.05	0	12.35
Erasmus students outgoing (%)	2,477	1.44	1.25	0	12.9
Student/teacher ratio	2,066	13.91	8.73	0	126.92
Publications per researcher	2,022	0.49	0.51	0	3.86
International publications (%)	2,510	30.78	19.52	0	100
Industry collaboration (%)	2,510	1.72	2.17	0	50
Top 10% documents	2,510	9.83	7.4	0	100
Employment rate	2,139	47.8	13.22	23.07	99.76
Human Capital (%)	2,535	28.56	9.4	9.9	69.9
Net migration rate	2,000	-0.02	0.52	-8.43	3.74
Population density	2,547	835.39	1854.81	1.6	21490
Fixed capital	2,366	11755.21	10838.38	243.54	70011.6

Table 3. Descriptive Statistics./ *Source*: Produced by the authors using Stata 14.

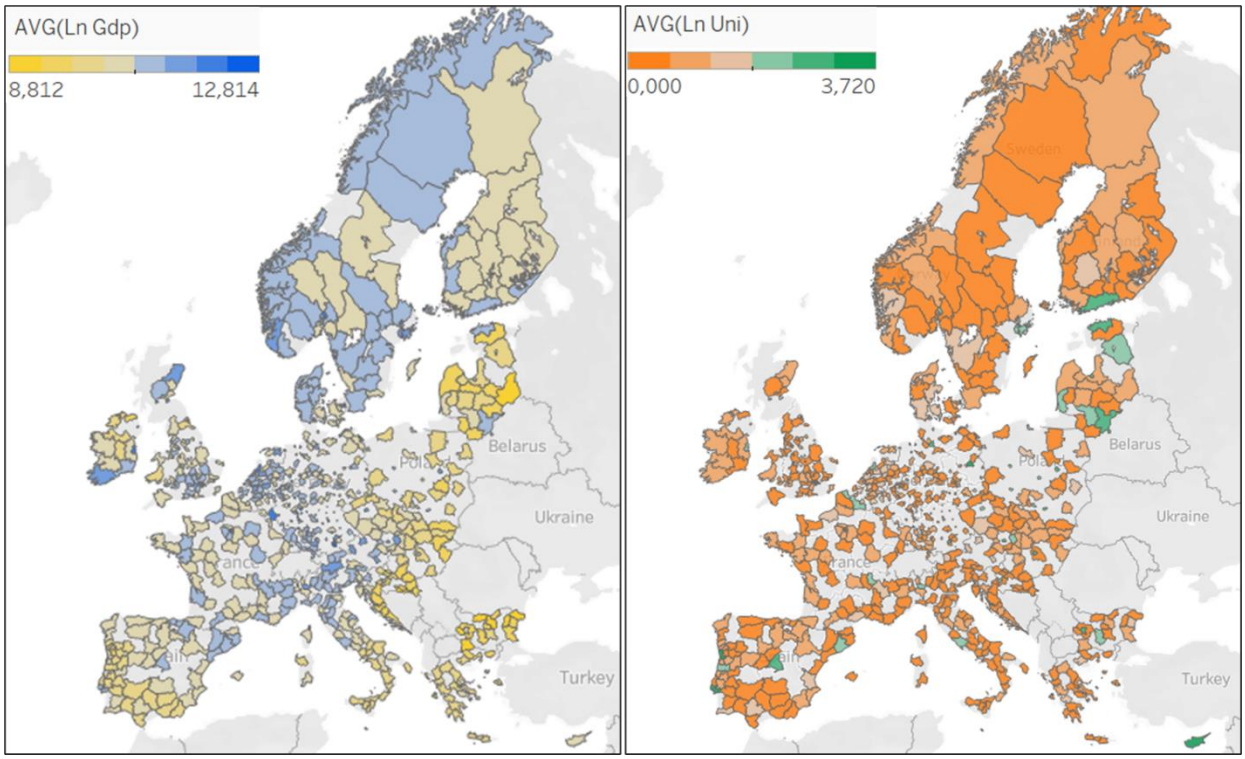

Figure 1. GDP per capita and number of universities per region (NUTS 3)./ *Note:* the graph on the left side represents the mean over years of the natural logarithm of GDP per capita. On the right side, the map shows the mean over years of the natural logarithm of the number of universities. Both maps report data for the 649 NUTS-3 regions containing at least one university (i.e. the regions analysed in this study). *Source:* Produced by the authors using Tableau 2020.2.

In terms of GDP per capita (see the first map in Figure 1), the richest regions are mainly the capital cities (such as London, Paris, Oslo, Brussels, Dublin, Vienna and Prague; but not Rome, Madrid and Lisbon) or business cities (e.g. Munich, Frankfurt and Milan). The area of London, i.e. Camden and City of London and Westminster, reports the highest values, with 338,454 and 251,508 PPS[6] per capita respectively. The top 20 regions in terms of GDP per capita are all in the UK and Germany, with only the two exceptions of Paris and Luxembourg. A similar picture is depicted by the data on employment rates. German regions report the highest shares of regional employment, even if remarkable rates are also shown by some regions in the UK and Norway.

The map on the right side of Figure 1 represents the number of universities operating in each region. The highest densities of universities are in the area of Lisbon (PT170), with 42 universities, and in the region of Berlin (DE300), with 30 universities. A high number of universities is also reported in London and Paris, where 27 and 23 institutions operate

[6] PPS (Purchasing Power Standard) is an artificial currency created by Eurostat that is used to eliminate disparities in price levels between European countries.

respectively[7]. However, excluding the most important urban areas, the majority of the regions have a limited density of universities with an average value of 2.5 universities per region. A very different ranking is obtained instead when considering the total number of students of regional HESs. In this case, the largest value is reported by the region of Madrid (with an average 470,424 students).

Significant differences among countries emerge when we analyse the data on the public-private structure of HE systems and HEIs typology. The statistics show that the UK and Netherlands are countries with a prevalence of private institutions. Applied science universities characterise instead Germany, Ireland, Portugal, Netherlands, Belgium, Hungary, Greece, Denmark, Norway, Finland, Latvia and Estonia. These divergences reflect differences in the structure of national HESs and are controlled by employing country fixed effects in the empirical models. Unlike applied science universities, the presence of medical schools does not reveal significant differences among countries. The statistics show instead a dichotomic tendency in the values ascribable to the low number of universities generally operating in NUTS-3 regions; 11% of the HESs in the sample is composed of medical universities exclusively, while the regions without medical schools are around 65%.

Many HESs in Germany are highly specialised in STEM disciplines, even if substantial shares of students in these fields are also reported for some regions in Austria (in particular in Östliche Obersteiermark) and in the UK (e.g. Central Bedfordshire). Instead, data on the share of Erasmus students (both incoming and outgoing) does not show a clear pattern among regions but seems to reflect focused strategies and specific international agreements adopted by universities. Concerning physical resources of HESs, on average there is one member of the academic staff every 14 students, even if the values vary considerably among regions. The highest student-teacher ratio is in Boeotia (Greece), with 127 students per teacher whereas the minimum (equal to 0) refers to the Lucca region where there is only one university offering exclusively doctoral programmes[8].

In terms of teaching performance, the HESs with the best graduation rates are generally in France, even though the highest value is associated with a region in Norway (Finnmark). Moving on to research performance, the data on the number of publications per researcher

[7] It is worth mentioning that London is composed of 12 NUTS 3, with different numbers of universities. The NUTS-3 region of Camden and City of London (UKI31) has the highest number of institutions, namely 9 universities.

[8] Student-teacher ratio indicator refers to students from ISCED 5 to 7, without considering doctoral (ISCED 8) students.

shows heterogeneous values among regions. HESs with high research productivity prevail mainly in the UK (i.e. in London, Cambridgeshire and Oxfordshire), but are also reported in Greece, Germany and Italy. Instead, the share of top-cited publications and the share of publications with international co-authors do not show any relevant pattern over regions. High research performance seems, in fact, to derive from specific competitive behaviours adopted by universities. Finally, large shares of publications with industry collaboration (i.e. our measure for third-mission performance) are mainly reported by regions in Norway, Finland, Netherlands and the UK.

Overall, the descriptive statistics reported in this section are intended to demonstrate the high degree of heterogeneity across the HESs in the various European regions, under various dimensions. This diversity reinforces our argument that it is worth analysing how these various factors contribute, simultaneously and interactively, to explain different levels of economic development of the regions in which the institutions are operating.

5. Results

5.1. The results of the econometric approach

Table 4 shows the results of the regression analyses for seven different models, which gradually include all the independent variables of the model in equation (1), following a stepwise approach. From column 1 to column 4 of Table 4, we separately include the three groups of HES variables as defined in the framework shown in Table 1 (i.e. institutional characteristics and HES size, programmatic characteristics and resources, higher education systems performance). As reported in the bottom part of the table, HES variables alone can explain less than 35% of the variance of the dependent variable (R-squared of R4 = 0.3482). The goodness of fit increases significantly from column 5, when we include the control variables, explaining alone 77.5% of the variance. Instead, when we add all the HES variables to the controls (model R6), the R-squared increases only 3 percentage points, showing therefore a greater relevance of socioeconomic variables compared with the educational ones. Finally, model R7 includes country fixed effects and time dummies.

HES variables are generally associated with significant effects in the first models (from R1 to R4), but many of these coefficients lose their statistical significance when the control factors are included. Looking at the more complete models (i.e. R6 and R7 models), the results highlight in particular the importance of the number of universities in the region, the share of

students in private institutions, and the research collaboration with industry while controls are all strongly statistically significant.

VARIABLES	R1	R2	R3	R4	R5	R6	R7
	ln(GDP per capita	ln(GDP per capita	ln(GDP per capita	ln(GDP per capita	ln(GDP per capita	ln(GDP per capita	ln(GDP per capita
ln(Number of universities)	0.0134	-0.0900***	-0.0771***	-0.1212***		-0.0071	0.0190
	(0.0190)	(0.0271)	(0.0198)	(0.0249)		(0.0136)	(0.0124)
Share of private students	0.0017***	0.0015***	0.0018***	0.0012***		0.0009***	0.0010***
	(0.0002)	(0.0003)	(0.0002)	(0.0003)		(0.0002)	(0.0002)
Medical Universities (%)	-0.0002	-0.0017***	-0.0019***	-0.0032***		-0.0002	-0.0002
	(0.0003)	(0.0004)	(0.0004)	(0.0005)		(0.0002)	(0.0002)
Applied science universities (%)	0.0028***	0.0025***	0.0049***	0.0031***		0.0004	0.0000
	(0.0003)	(0.0006)	(0.0004)	(0.0005)		(0.0003)	(0.0003)
ln(Students enrolled)	0.1407***	0.1973***	0.1515***	0.2252***		0.0599***	0.0147
	(0.0130)	(0.0179)	(0.0133)	(0.0187)		(0.0113)	(0.0091)
Graduation rate			-0.0011	0.0008		-0.0016**	0.0013*
			(0.0011)	(0.0012)		(0.0008)	(0.0008)
STEM (%)		-0.0004		-0.0013*		-0.0007*	0.0007**
		(0.0007)		(0.0008)		(0.0004)	(0.0003)
ln(Doctoral students enrolled)		0.0030		-0.0256**		-0.0344***	-0.0094*
		(0.0095)		(0.0107)		(0.0069)	(0.0054)
Erasmus students incoming (%)		-0.0376***		-0.0499***		0.0202**	-0.0053
		(0.0101)		(0.0111)		(0.0080)	(0.0046)
Erasmus students outgoing (%)		0.0256**		0.0336***		0.0088	0.0092*
		(0.0106)		(0.0110)		(0.0074)	(0.0048)
ln(Students/teacher ratio)		-0.3674***		-0.3486***		-0.0371**	-0.0061
		(0.0268)		(0.0309)		(0.0147)	(0.0140)
Publications per researcher			0.1794***	0.2485***		0.0270	-0.0008
			(0.0497)	(0.0542)		(0.0193)	(0.0179)
Top 10% documents			0.0049	0.0026		0.0005	-0.0006
			(0.0032)	(0.0032)		(0.0012)	(0.0012)
International publications (%)			-0.0012	-0.0026**		0.0023***	0.0005
			(0.0010)	(0.0013)		(0.0007)	(0.0005)
Industry collaboration (%)			0.0617***	0.0488***		0.0147***	0.0094***
			(0.0085)	(0.0087)		(0.0041)	(0.0036)
Employment rate					0.0210***	0.0207***	0.0204***
					(0.0004)	(0.0006)	(0.0006)
ln(Population density)					-0.0005	-0.0061	0.0173***
					(0.0031)	(0.0058)	(0.0059)
ln(Fixed capital)					0.1060***	0.1040***	0.0447***
					(0.0061)	(0.0084)	(0.0083)
Human Capital (%)					0.0063***	0.0028***	0.0095***
					(0.0005)	(0.0009)	(0.0010)
Net migration rate					0.0211***	0.0207***	0.0240***
					(0.0067)	(0.0076)	(0.0079)
Constant	8.7464***	9.2442***	8.4725***	8.9480***	8.0858***	7.8153***	8.4666***
	(0.1237)	(0.1516)	(0.1194)	(0.1551)	(0.0596)	(0.1172)	(0.1029)
Country fixed effects	no	no	no	no	no	no	yes
Time dummies	no	no	no	no	no	no	yes
Observations	2,317	1,480	1,967	1,414	1,666	1,022	1,022
R-squared	0.1429	0.2783	0.2587	0.3471	0.7751	0.8058	0.8856
RMSE	0.42697	0.40849	0.40926	0.39083	0.17822	0.17287	0.1339

Table 4. Regressions estimates./ *Note*: *** indicates significance at the 1% level, ** at the 5% level and * at the 10% level. Robust standard errors are reported in parentheses. *Source*: Produced by the authors using Stata 14.

Nevertheless, the regression analysis tends to reflect the presence of correlations between the HES variables and the dependent variable (see Section A1) without assuring the existence of causal effects. To control for the endogeneity of the HES variables, we employ a system Generalised Method of Moments. The results of the complete model (based on R7) are reported in column 5 of Table 5 (GMM5), together with the estimates of further four models (GMM1, GMM2 and GMM3), which include separately the three areas of HES variables. All six models include the controls, country fixed effects, time dummies and the variables on HES size (i.e. the number of universities and the number of students in the regional HES).

The results of the Arellano-Bond test for second-order autocorrelation (see the bottom part of Table 5) confirm the validity of the results, which are not affected by problems of autocorrelation (see Roodman, 2018). The Hansen test does not reject the null hypothesis for all the models, proving therefore the validity of the instruments employed.

According to sys-GMM results, the regional employment rate is the most important determinant for local economic development. In particular, considering model GMM5, a rise of 10% in the employment rate of the region generates an increase of 9.86% in regional GDP per capita. Instead, among the HES variables, the number of publications per researcher provides the most relevant contribution to regional economic development, with statistically significant effects found in all the models. The analysis of the elasticity of GMM5 suggests that an increase of 10% in the number of publications per researcher produces a rise of 1.01% in the local GDP per capita. The number of universities in the region also produces positive effects, even if their coefficients are statistically significant only for GMM1 and GMM3. In detail, the results of GMM3 show that a 10% rise in the number of universities in the region is associated with an increment of 4.75% in the level of regional GDP per capita. Instead, international collaboration rate and STEM specialisation seem to produce negative effects on the regional GDP per capita. However, the effects are statistically significant only for GMM3 and GMM4, respectively. The number of enrolled students loses its statistical significance when controlling for the endogeneity through the GMM. This result seems to highlight an endogenous behaviour of HES size. Indeed, large institutions could be more likely to operate in wealthy regions, and the size of universities can express the quality and the attractiveness of the institutions.

Moreover, the limited significance of the effects found for the HES variables could be due, in general, to the linearity imposed by the adoption of parametric models. Significant but nonlinear relationships risk not being captured by regression or sys-GMM estimates (see Section 3). In addition, the results of econometric analyses can be affected by the inclusion of a large number of regressors, especially if correlated with each other – in the literature, this issue is known as the "high-dimensional problem" (see, among others, Carrasco et al., 2015). The random forest analysis, reported in the next section, addresses both the problem of linearity and the high dimensionality of the covariates (see Section 3.2), providing new evidence on the relationships between the characteristics and the performance of HESs and regional economic development.

VARIABLES	GMM1 ln(GDP per capita	GMM2 ln(GDP per capita	GMM3 ln(GDP per capita	GMM4 ln(GDP per capita	GMM5 ln(GDP per capita
ln(Number of universities)	0.7956**	0.1120	0.4748***	0.0081	0.1011
	(0.3609)	(0.1652)	(0.1621)	(0.1158)	(0.1305)
ln(Students enrolled)	-0.1990	0.1574	-0.0997	0.0157	0.0713
	(0.1671)	(0.1223)	(0.0724)	(0.0677)	(0.0465)
Share of private students	-0.0018				0.0025
	(0.0033)				(0.0056)
Medical Universities (%)	-0.0099			-0.0007	-0.0037
	(0.0208)			(0.0023)	(0.0035)
Applied science universities (%)	-0.0008				-0.0033
	(0.0136)				(0.0039)
STEM (%)		-0.0031		-0.0079**	-0.0017
		(0.0042)		(0.0037)	(0.0021)
ln(Doctoral students enrolled)		-0.0752			-0.0359
		(0.0511)			(0.0402)
ln(Students/teacher ratio)		-0.1538			-0.0594
		(0.1137)			(0.0788)
Erasmus students incoming (%)		-0.0023		-0.0053	-0.0021
		(0.0285)		(0.0479)	(0.0125)
Erasmus students outgoing (%)		0.0071			0.0003
		(0.0232)			(0.0087)
Graduation rate			0.0001	0.0033	0.0023
			(0.0023)	(0.0038)	(0.0015)
Publications per researcher			0.2216**	0.0257	0.2029***
			(0.0866)	(0.1106)	(0.0563)
Top 10% documents			0.0026		0.0018
			(0.0024)		(0.0019)
International publications (%)			-0.0058***	-0.0008	-0.0030
			(0.0022)	(0.0021)	(0.0018)
Industry collaboration (%)			0.0009		-0.0023
			(0.0031)		(0.0024)
Employment rate	0.0180***	0.0183***	0.0214***	0.0197***	0.0197***
	(0.0046)	(0.0019)	(0.0016)	(0.0019)	(0.0024)
ln(Population density)	-0.0663	0.0057	-0.0597**	0.0245	0.0085
	(0.0858)	(0.0361)	(0.0287)	(0.0224)	(0.0281)
ln(Fixed capital)	0.0434	0.0143	0.0380*	0.0466***	0.0468*
	(0.0670)	(0.0324)	(0.0205)	(0.0168)	(0.0249)
Human Capital (%)	-0.0100	0.0072**	-0.0004	0.0101***	0.0051**
	(0.0139)	(0.0030)	(0.0034)	(0.0021)	(0.0026)
Net migration rate	-0.0169	0.0227	-0.0074	0.0097	0.0156
	(0.0336)	(0.0161)	(0.0149)	(0.0194)	(0.0147)
Constant	10.8905***	8.2856***	9.8845***	8.7339***	8.4816***
	(1.7482)	(0.5084)	(0.5531)	(0.5012)	(0.5954)
Country fixed effects	yes	yes	yes	yes	yes
Time dummies	yes	yes	yes	yes	yes
Observations	1,603	1,053	1,439	1,116	1,022
Number of _nuts3	455	383	416	384	374
Arellano-Bond (2), p-value	0.292	0.819	0.317	0.196	0.343
Hansen test, p-value	0.127	0.093	0.111	0.145	0.121

Table 5. Sys-GMM estimates./ *Note*: *** indicates significance at the 1% level, ** at the 5% level and * at the 10% level. Robust standard errors are reported in parentheses. *Source*: Produced by the authors using Stata 14; GMM estimations were performed using xtabond2 command (see Roodman 2018).

5.2.The results from the random forest approach

The second stage of the empirical analyses aims at discovering nonlinear effects between HES variables and local GDP per capita and, at the same time, taking into account the "high-dimensional problem". On the other hand, in this phase, we are not interested in the causality of the effects, which have been more specifically explored through GMM analyses.

The random forest employed in this stage provides, as the main result, the relative importance of the covariates in predicting the response, measured through %IncMSE (see Section 3.2). Table 6 displays the results of the seven models tested, which gradually include the HES variables, one group at a time. Among the seven models, it is worth focusing our attention on RF4, which takes into account the effects of HES variables alone, and RF7, which is the complete model that considers all the HES and the control factors. The rankings of the relative importance associated with these two models are represented by the plots in Figure 2. Confirming the results found in the first stage, in RF7 model, the most important covariate seems to be the regional employment rate, which reports an IncMSE of 54.76% (see RF7 in Table 6). The fixed capital and share of people in the region with higher education are also particularly important, with an IncMSE of 36.37% and 26.91% respectively. Among HES variables, the size of HESs, internationalisation and research productivity seem to represent the most influential dimensions. In particular, the MSE increases by 21.17% if the share of Erasmus incoming is excluded, whereas the rise is 17.18% when the number of enrolled students is omitted. The high relevance found in the number of publications per researcher (IncMSE =13.97%) is strengthened by the results of GMM analysis, which prove the causality of this effect on regional economic development. The same interpretation holds for the number of universities in the region, which reports an IncMSE of 11.36% in RF7. In contrast, the share of medical universities, university-industry collaboration and the number of doctoral students seem to be less influential.

When we analyse the effects of HES variables alone (RF4 model), the most relevant factor is the student-teacher ratio, which significantly loses its relative importance with the inclusion of the controls in the model (see RF5 and RF6 in Table 6). This behaviour is probably due to the high correlation between the physical resources employed by HESs and the regional employment rate (see Section A1). The employment rate seems, in fact, to capture a large part of the effect that is associated with the student-teacher ratio in model RF5.

As already said, random forest allows missing data to be handled by building surrogate nodes that avoid dropping all the observations with missing values. Relying on this feature, it is possible to use a larger number of observations compared to the analyses on the first stage. However, we need to split the sample into a training set, the group of observations used to fit the parameters, and a testing set, the group of observations used to test the model performance. The number of observations in the training set and in the test set is reported for each model in the bottom part of Table 6, together with the share of variance explained by the model (% Var explained), the mean square error (MSE) and the root-mean-square error (RMSE)[9].

The shares of variance explained and the RMSE coefficients, reported in the bottom part of Table 6, are compared with the performance of the regressions models in Table 4 (see models 1, 2, 3, 4, 6 and 7 in Tables 4 and 6). The statistics suggest that RF provides a better fit to the data, especially for HES variables. Despite the socioeconomic variables still being the most relevant factors in influencing the level of local GDP per capita, by employing the RF approach the HES variables have a larger capability of capturing the variance in the regional economic development. This is shown by comparison of the share of variance of the dependent variable explained by R4 in Table 4 with RF4 in Table 6. When the model includes only HES variables, the random forest approach explains 34.3% more of GDP per capital variance than the one captured by the regression approach (see RF4 in Table 6 and R4 in Table 4). The large difference in the performance of the two approaches can be due to the presence of nonlinear relationships between HES variables and the logarithm of regional GDP per capita, which are clearly shown by partial dependence plots in Figure 3. The better fit of the nonlinear model is also statistically demonstrated by the F-test reported in Tables A3 in the Annex.

[9] The share of var explained was computed using the out-of-bag observations in the training set, while the MSE and RMSE were estimated using the observations in the test set.

Variables (%IncMSE)	RF1 ln(GDP per capita)	RF 2 ln(GDP per capita)	RF 3 ln(GDP per capita)	RF 4 ln(GDP per capita)	RF 5 ln(GDP per capita)	RF 6 ln(GDP per capita)	RF 7 ln(GDP per capita)
ln(Number of universities)	19.15	14.06	17.17	11.02	15.39	9.26	11.36
Share of private students	25.88	24.12	23.96	26.33	15.44	12.28	10.82
Medical Universities (%)	22.26	15.63	10.88	13.34	10.88	10.59	8.84
Applied science universities (%)	30.27	23.51	25.68	16.31	13.87	10.01	12.58
ln(Students enrolled)	31.74	27.57	31.53	26.17	18.04	17.02	17.18
STEM (%)		30.10		17.48	22.32	14.28	12.36
Erasmus students incoming (%)		35.34		28.55	31.61	24.47	21.17
Erasmus students outgoing (%)		21.28		16.28	24.12	11.67	13.06
ln(Students/teacher ratio)		39.50		37.95	24.83	18.05	11.20
ln(Doctoral students enrolled)		23.61		18.04	13.45	10.68	9.20
Publications per researcher			32.99	27.54	17.89	14.52	13.97
Top 10% documents			27.82	17.42	14.46	10.06	9.76
International publications (%)			20.38	15.81	11.39	18.09	15.17
Industry collaboration (%)			26.01	25.33	16.01	11.78	9.01
Graduation rate			25.32	23.92	14.19	11.02	10.91
Human Capital (%)					32.38	33.64	26.91
Net migration rate					46.08	25.13	21.46
Employment rate						86.83	54.76
ln(Population density)					64.22	15.95	19.19
ln(Fixed capital)					58.45	56.78	36.37
Country							24.31
% Var explained	59.78	68.01	57.18	69.01	88.12	92.89	93.62
MSE	0.1501	0.0679	0.0964	0.0623	0.0300	0.0129	0.0127
RMSE	0.3874	0.2606	0.3105	0.2497	0.1733	0.1134	0.1127
Number of obs. (training set)	1808	1808	1808	1808	1808	1808	1808
Number of obs. (testing set)	602	602	602	602	602	602	602

Table 6. Variables importance – RF approach./ *Note:* the table reports the value of %IncMSE associated with each variable in the models. The value is scaled by dividing for the MSE of the specific variable, providing an indicator of the relative influence of the variables. *Source*: Produced by the authors using R.

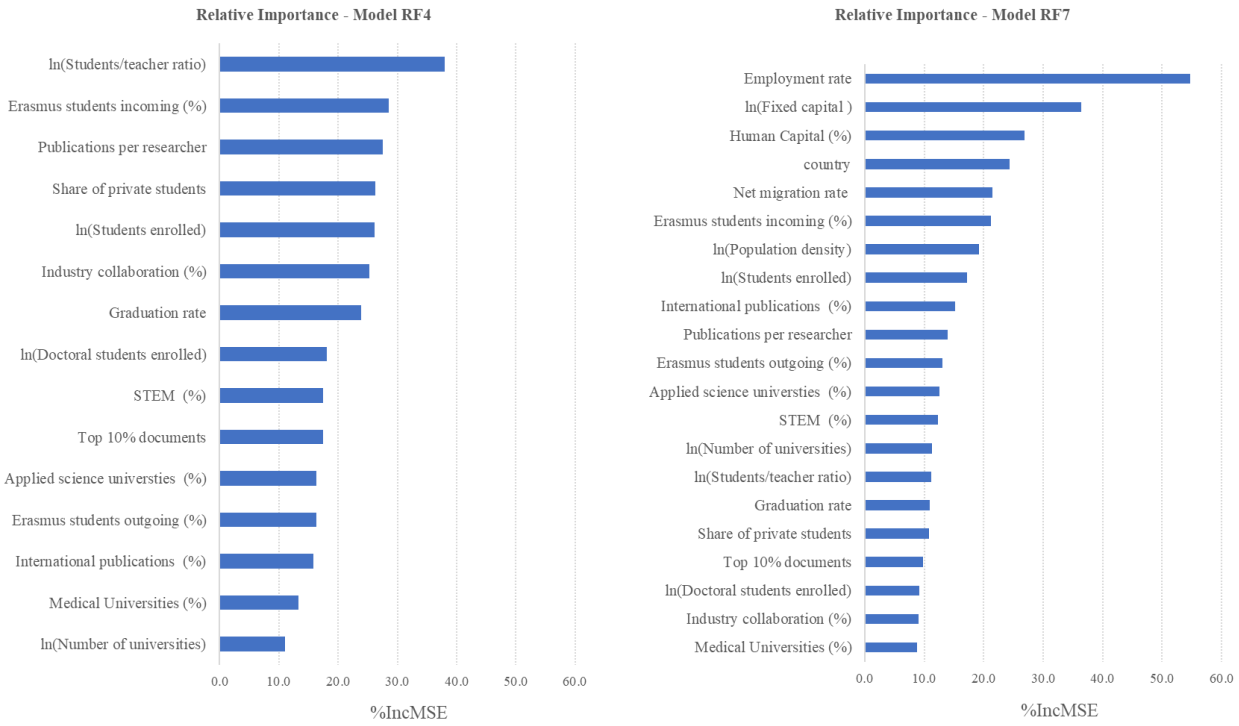

Figure 2. Plot of variables importance Model 4 and Model 7./*Note:* variables importance ranked according to %IncMSE. On the left side, we report the variables importance of model (4) that included only regional HES variables; on the right side, we report the variables importance of model (7), the complete model including control factors. *Source*: Produced by the authors based on RF results.

The plots in Figure 3 provide a graphical representation of the marginal effect of each covariate on the dependent variable, taking as a reference the complete model (i.e. RF7). In this way, we can understand how each independent variable influences the regional GDP per capita – while the relative importance of the covariates (RF7 in Table 6) suggests the magnitude of the effects. The graphs in Figure 3 show the complexity of the relationships existing between HES variables and the logarithm of regional GDP per capita – which are generally nonlinear. The nonlinearity of these relationships is also confirmed by the results of the semiparametric analysis in Section A3 – which provide information on the inference of the effects.

An increase in the number of universities in the HES is associated with positive effects on regional economic development. However, exciding a critical level between 7 and 8 universities in the region, the effect becomes negative. This evidence points to the existence of a saturation level for the HESs; after that, the establishment of a new university is no more convenient for the local economy. The result is particularly interesting if compared with the findings in the literature (see Valero and Van Reenen, 2019), which recognise a positive and

linear effect associated with the number of universities, without identifying a saturation level that cannot be captured by traditional econometric models.

The effect of the size of HESs, measured by the number of students in the system, seems to follow a different behaviour. Its marginal effect on regional GDP per capita is linear, even if no effect is registered for HESs with less than 200 students. In this case, there is no saturation level and the effect continues to grow as the number of students increases. This result, together with the high relevance found for the number of students, sheds light on the high correlation between HES size and regional economic development. However, we must note that the causality of the effect is not confirmed by the GMM analysis (see Table 5) and the mechanism likely works in both directions. The high number of graduates can contribute to the local economic output by providing highly qualified human capital and, in turn, the wellbeing of the region can significantly increase the local demand for higher education and attract students from outside the region. A linear/slightly quadratic behaviour is also reported by the partial dependent plot of the share of students in private universities (even if with a low magnitude of the effect). Similarly to the number of students, the correlation identified by the random forest could partially reflect a larger demand for private education that characterise highly developed areas (Altbach, 1999). Indeed, we should be aware of different patterns of concentration of universities between the public and private sector, with the latter operating in the regions with the highest demand of HE (Teixeira et al., 2014). This difference can be attributed to the stronger market orientation of private universities, which is in contrast with the strategy of the larger spatial coverage of the public sector (Teixeira et al., 2014).

In terms of HES resources, the partial dependent plot of student-teacher ratio shows the presence of a maximum effect for HESs that, on average, have one academic staff every 7.5 students. The result seems to benefit HESs with a significant amount of physical resources, since the maximum effect corresponds to a small student-teacher ratio, largely below the average of the sample (i.e. 13.9).

As noticed previously, the share of Erasmus students incoming positively affects regional GDP per capita, providing the largest contribution among the HES variables (see Table 6). On average, the maximum effect is associated with a share of around 3%, while a further increase does not provide any difference in the economic output. Similarly, the marginal effect of the share of Erasmus outgoing increases until 2.5%, a point at which it starts to decrease to a minimum at around 7%. However, it is worth noting that a share of 2% or 3% of Erasmus

students is significantly above the average (i.e. between 1 and 1.5%) of the sample and can be associated with HESs devoting exceptional efforts to international mobility objectives.

The number of doctoral students in the HESs can affect the regional economic output positively only above a critical level of 3,000 PhDs – there is no effect with lower values. Considering that, on average, HESs have less than 1000 students in doctoral programmes, only research-intensive institutions seem to positively contribute to the local economy. Instead, the share of students in STEM disciplines appears to be negatively correlated with regional development. This evidence seems to ascribe a greater economic contribution to HESs with generalist programmes. However, we can better interpret this result looking at the correlation analysis in Section A1. HESs focused on STEM are inversely correlated with graduation rates that, in turn, positively affect regional GDP per capita. Moreover, private universities are less likely to offer STEM programmes, which positively influence regional economic development, are more frequent in applied science universities and are generally associated with negative effects on the regional GDP per capita.

Concerning teaching performance, the partial dependent plot of graduation rates shows a dichotomous effect; rates below 30% are associated with low levels of GDP per capita, while graduation rates over 40% are associated with higher economic development. Since less than one third (27.7%) of the observations reports graduation rates over 40%, the result suggests that the only HESs with the best teaching performance can provide a significant economic contribution to the regions, even if the magnitude of the effect is relatively small (see Table 6).

Considering HES performance, research indicators show the most important effects, especially in terms of publications per researcher and share of documents with international collaboration. The indicator for research productivity shows a quasi-linear marginal effect that could explain the high statistical significance of this variable in the results of regression and sys-GMM analyses (see Table 4 and Table 5). On the other hand, the low p-values detected by the GMM analysis confirm the causality of the effect of research productivity on regional economic development. In addition, it is remarkable that the size of the effect increases for researchers publishing more than three documents per year. Instead, the contribution associated with third-mission performance, expressed as the share of publications with industry collaboration, has a limited relative importance. Nevertheless, this result should be interpreted with caution, given the limitations of this indicator presented in Section 4.1. Indeed, the economic contribution of third-mission activities could have been significantly

underestimated due to the intangible nature of this factor. Focusing on the shape of the effect, the partial dependence plot of industry collaboration shows a linear positive effect for the majority of the observations (98.3%), reporting shares of industrial collaboration that range between 0% and 7%.

Finally, Figure 3 highlights the existence of linear relationships between the controls and the dependent variable. The only exception is the marginal effect of interregional migration rate, which highlights a significant difference in the GDP per capita between regions with incoming flows (the richest regions), and the ones with outgoing flows (the poorest regions). Among the controls, it is worth mentioning the existence of a linear relationship between the level of economic development and the human capital of the region, confirming the linearity generally assumed by endogenous economic models (see Sianesi and Van Reenen 2003). The linear effects generated by these factors could explain why the difference between the goodness of fit of RF models and the fit associated with regression analysis is particularly low for the models including the controls (models 6 and 7 of Tables 4 and 6). Even if econometric models cannot properly capture most of the effects generated by HES variables, they are suitable for representing the linear effects of control factors, something which alone can well explain the variation in the GDP per capita of the region.

All the considerations discussed above hold for all the European countries we analyse, as shown in Section A4 of the Annex, which reports the analysis of country fixed effects. In this sense, our study offers a communal model to describe the relationships between HES factors and the local economic development of European countries.

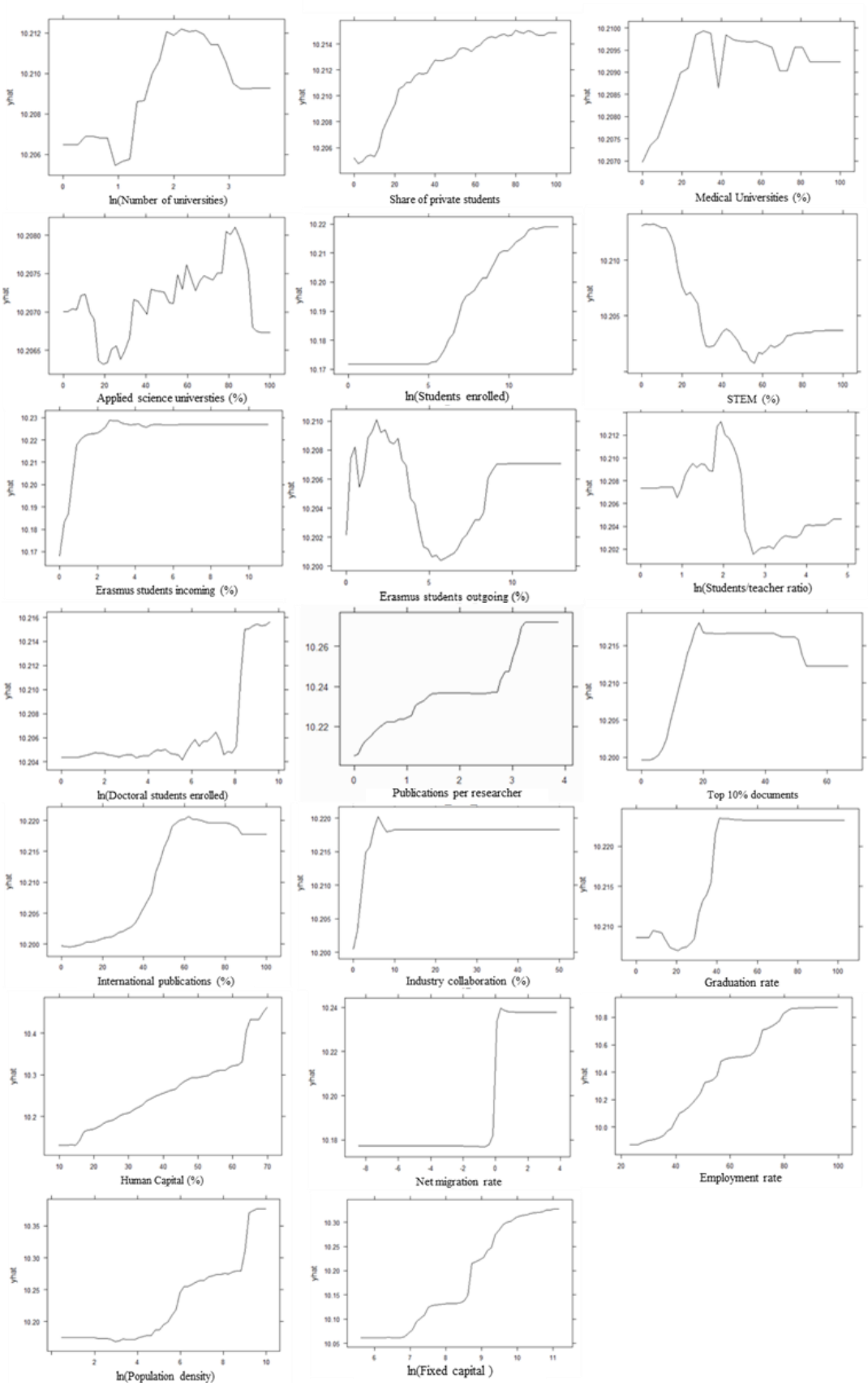

Figure 3. Partial dependence plots./ *Note:* partial dependence plots of the regressors included in model (7) in the association with the logarithm of GDP per capita of the region. *Source*: Produced by the authors using R.

Joint partial plots are employed to represent how the size of HESs and their characteristics jointly affect the logarithm of GDP per capita. In particular, we report the joint effects of the number of students in the system with the share of Erasmus students incoming (Figure 4) and the number of publications per researcher (Figure 5) – representing those of the most important HES variables (see RF7 in Table 6). The extent to which the size of HES affects the level of regional GDP per capita depends critically on how high the research productivity and the international mobility of the students are. (See Figure 4 and Figure 5).

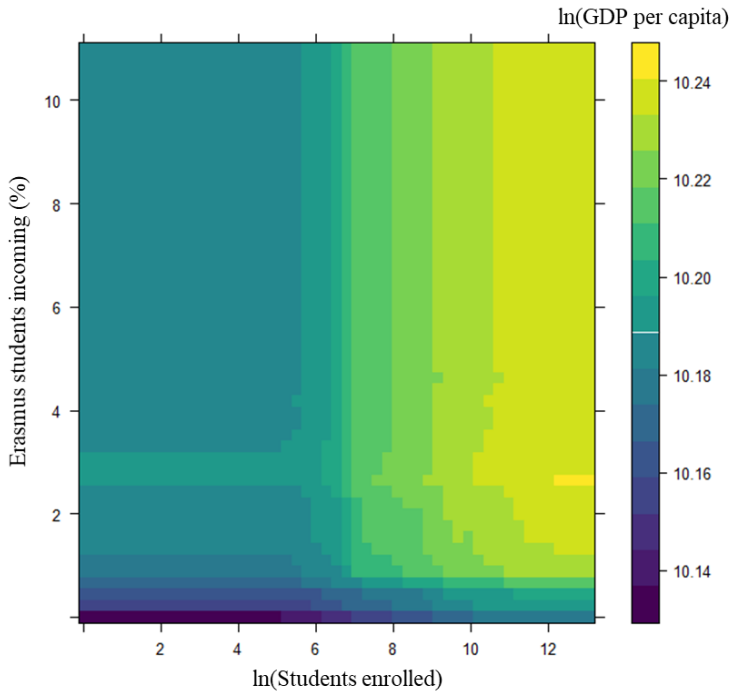

Figure 4. Joint partial plot of HES size and the share of Erasmus students incoming./ *Note:* the colour represents the scale of the values of the response. *Source*: Produced by the authors using R.

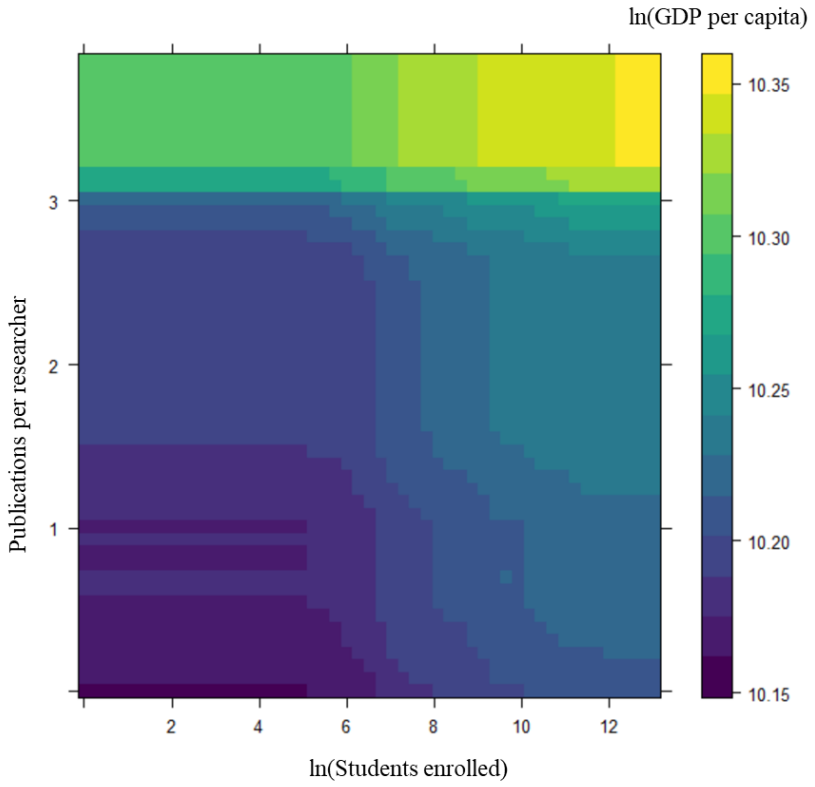

Figure 5. Joint partial plot of HES size and research productivity./ *Note:* the colour represents the scale of the values of the response. *Source*: Produced by the authors using R.

6. Discussion, policy messages and concluding remarks

The present paper provides new evidence on the relationships between HESs and regional economic development, in a context of high heterogeneity among HESs. A key contribution of this work is that we demonstrate the existence of nonlinear relationships between most of the indicators used to model HESs and the GDP per capita of European regions. In this context, traditional methodological approaches, which assume the linearity of the effects, cannot provide suitable estimations. Our results show, in fact, that machine learning techniques (in this case, random forest) can capture nonlinear effects associated with HES indicators, which are not detected by traditional econometric models (i.e. regression and GMM analysis). For this reason, the extant works in the literature could have underestimated the economic impact of HESs, detecting only the part of the total effect that is linearly associated with regional economic development. By addressing this issue, this study offers a new methodological approach that combines traditional econometric models with machine learning techniques.

Moreover, this work provides a comprehensive assessment of the economic effects associated with the characteristics and the performance of HESs. While the studies in the literature tend

to focus on specific features of universities, our analysis seeks to examine all the dimensions through which HES heterogeneity can occur. Specifically, we study a framework of 15 indicators that, based on the literature, fully represent the characteristics and the performance of HESs. The empirical findings on the relationships between regional GDP per capita and these HES indicators lead to four main observations.

Firstly, increasing the size of HESs by establishing a new university in the region is not always beneficial in terms of local economic performance (compared with Valero and Van Reenen, 2019). A better strategy for enhancing the economic contribution of HESs would be to increase the number of students in the existing universities. This result could suggest the existence of economies of scale for European universities that, in turn, may positively affect the economic contribution of HESs. In this scenario, larger universities are associated with smaller costs per unit and therefore can save physical and financial resources that may be employed for further stimulating local economic development. The existence of this specific mechanism emerges from the work of Agasisti et al. (2019), which provides empirical evidence of the link between the efficiency of universities and their local economic contribution. In addition, the size of universities may indirectly enhance local economic development through improvement in academic performance. For example, larger universities are usually associated with high research productivity (see Abramo et al., 2012). Nevertheless, we are aware of a possible problem of endogeneity of HES size, which is suggested by the lack of statistical significance found in the GMM analysis. In this sense, HES size seems to activate a virtuous cycle together with the local economic development. On the one hand, universities increase the human capital of the region by providing new graduates, on the other hand, the economic wellbeing of the region can significantly foster the demand for higher education and provide the physical and financial resources allowing an increase in the size of universities.

Secondly, the international mobility strategy implemented by HESs can significantly influence the regional GDP per capita. In particular, the share of Erasmus students incoming appears as the most important HES feature to foster the local economy. Indeed, students participating in mobility programs contribute to the local economy by generating direct expenditures (mainly related to housing, food and tuition fees), while those remaining in the regions after their studies provide longer-term benefits, for instance, by increasing local innovation and entrepreneurial activities (Owens et al., 2011). It could also be the case that Erasmus students are more willing to spend their period abroad in highly developed cities,

which offer a wide range of services and amenities or where they might find more job opportunities after their graduation. This is corroborated by the results of the GMM analysis, which does not detect a statistical significance associate with this factor. Although high p-values could be due to the inadequacy of GMM in capturing the nonlinear effects, the result highlights a risk of reverse causality. The endogeneity of this indicator should however be limited for our analysis, in which we consider a relative measure (i.e. the rate of Erasmus students over the total number of enrolled students) that allows dimensional bias to be avoided. This observation is confirmed by the descriptive statistics in Section 4, which does not detect any clear pattern over regions for this indicator. Instead, significant shares of Erasmus students are likely to be associated with focused efforts and specific strategies implemented by the universities to develop multilateral exchange and mobility programmes (Seeber et al., 2020), internationalising their curricula (Leask and Bridge 2013) and promoting the image of the university abroad (i.e. marketing strategy, see Chen, 2008). However, in the paper, we focused only on Erasmus students and the results presented in the previous section should be interpreted as a lower-bound effect of the student mobility strategy of the universities.

Thirdly, among the indicators of HES performance, research productivity provides the most significant effect on regional economic development. Its relevance is detected both by random forest and GMM estimates thus assuring the causality of the effect. The result is consistent with the works in literature, ascribing an important economic contribution to research-intensive universities (see, for instance, Lendel, 2010; Goldstein and Drucker, 2006). On the other hand, we are fully aware that the limited significance found for the indicators of teaching and third-mission performance could be due to the low tangibility of these activities. Contrary to research, teaching and third-mission performance provide less tangible outputs that could be hard to capture with empirical measures.

The major limitation of this study relates to the time dimension here investigated. By employing ETER as the main source of data, we are able to include a comprehensive set of indicators on HESs but, on the other hand, the available information covers only a restricted period of time. In detail, the analysis focuses on the GDP per capita in a three-year period and considers a two-year lag between the dependent variable and the covariates. In this sense, we should be aware that the paper does not examine the relationships between HESs and local economic development from a long-term perspective, and this is an effort to be undertaken in future research. Future works may also investigate the effects of HESs on alternative

measures of economic development (e.g. income distribution, labour force structure and quality-of-life indicators). GDP per capita allows to rely on a large availability of data and limit the issue of endogeneity[10]. However, this measure is a monetary aggregate indicator that cannot capture non-productive flows and the distribution of wellness among the people (Stiglitz et al., 2009). Accordingly, the contribution of higher education systems to the whole regional economy may be higher than the one presented in the paper. Moreover, even if our paper significantly advances the measurement of HESs in studying their economic contribution, there is still room for the improvement of some HESs variables and indicators. This is an area where further research efforts should be concentrated. Specifically, the main limitations regard the quality of teaching performance, the different sources of revenues for universities and the less tangible activities of third-mission performance, which remain represented in the model only to a limited extent. The indicators of graduation rate and student mobility could also be improved by employing more precise data on student enrolments and detailed information on students' participation in international programs. In this context, the role of international data agencies is crucial to allow the availability of these HES indicators and advance future research, also providing more granular information. The interpretation of the policy implications derived from this study must keep data quality in mind and should be corroborated by further research exploring its robustness when additional (and more precise) indicators will be made available to analysts in this field.

The findings presented in this paper provide international evidence which can be useful for policy purposes. The comprehensive vision of the contribution of HESs to the economic performance of European regions informs policymakers on the most relevant features of HESs influencing the local economy, offering them new instruments for driving and fostering the economic development of their regions. This evidence may also help to understand and possibly tackle the causes of the economic disparities existing between European regions. As the main policy advice, our findings suggest that governments should be aware that regional HESs within their national boundaries are significantly heterogeneous and this diversity influences the economic performance of regions themselves. Accordingly, the characteristics and the peculiarities of each HES should be taken into account in setting up the regulatory frameworks, allocating more powers to local governments in shaping the collaborations with the institutions operating in a given territory.

[10] With broader indicators of economic development, education could be considered both a determinant factor and an element of development itself.

References

Abramo, G., Cicero, T., & D'Angelo, C. A. (2011). A field-standardized application of DEA to national-scale research assessment of universities. *Journal of Informetrics, 5*(4), 618-628.

Agasisti, T., & Berbegal-Mirabent, J. (2021). Cross-country analysis of higher education institutions' efficiency: The role of strategic positioning. *Science and Public Policy*, *48*(1), 66-79.

Agasisti, T., & Bertoletti, A. (2019). Analysing the determinants of higher education systems' performance–a structural equation modelling approach. *Science and Public Policy*, *46*(6), 834-852.

Agasisti, T., & Bertoletti, A. (2020). Higher education and economic growth: A longitudinal study of European regions 2000–2017. *Socio-Economic Planning Sciences*, 100940. DOI: 10.1016/j.seps.2020.100940.

Agasisti, T., Barra, C., & Zotti, R. (2019). Research, knowledge transfer, and innovation: The effect of Italian universities' efficiency on local economic development 2006−2012. *Journal of Regional Science, 59*(5), 819-849.

Agrawal, A., & Henderson, R. (2002). Putting patents in context: Exploring knowledge transfer from MIT. *Management science, 48*(1), 44-60.

Albats, E., Fiegenbaum, I., & Cunningham, J. A. (2018). A micro level study of university industry collaborative lifecycle key performance indicators. *The Journal of Technology Transfer, 43*(2), 389-431.

Altbach, P. G. (Ed.). (1999). *Private Prometheus: Private higher education and development in the 21st century* (No. 77). Greenwood Publishing Group.

Amendola, A., Barra, C., & Zotti, R. (2020). Does graduate human capital production increase local economic development? An instrumental variable approach. *Journal of Regional Science*, 60(5), 959-994.

Anderson, T. R., Daim, T. U., & Lavoie, F. F. (2007). Measuring the efficiency of university technology transfer. *Technovation*, *27*(5), 306-318.

Arellano, M., & Bond, S. (1991). Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations. *The Review of Economic Studies*, *58*(2), 277-297.

Arellano, M., & Bover, O. (1995). Another look at the instrumental variable estimation of error-components models. *Journal of Econometrics*, *68*(1), 29-51.

Barra, C., Maietta, O. W., & Zotti, R. (2019). Academic excellence, local knowledge spillovers and innovation in Europe. *Regional Studies*, *53*(7), 1058-1069.

Barra, C., Maietta, O. W., & Zotti, R. (2021). The effects of university academic research on firm's propensity to innovate at local level: Evidence from Europe. *The Journal of Technology Transfer*, *46*(2), 483-530.

Barro, R. J. (2015). Convergence and modernisation. *The Economic Journal*, *125*(585), 911-942.

Becher, T., & Trowler, P. (2001). *Academic Tribes And Territories: Intellectual Enquiry and the Culture of Disciplines.* New York, NY: McGraw-Hill Education.

Birnbaum, R. (1983). *Maintaining diversity in higher education*. San Francisco, CA: Jossey-Bass, Inc.

Blundell, R., & Bond, S. (1998). Initial conditions and moment restrictions in dynamic panel data models. *Journal of Econometrics*, *87*(1), 115-143.

Bonaccorsi, A., and Daraio, C. (2007) *Universities and Strategic Knowledge Creation: Specialization and Performance in Europe.* Cheltenham, UK: Edward Elgar Publishing.

Boucher, G., Conway, C., & Van Der Meer, E. (2003). Tiers of engagement by universities in their region's development. *Regional Studies*, *37*(9), 887-897.

Bramwell, A., & Wolfe, D. A. (2008). Universities and regional economic development: The entrepreneurial University of Waterloo. *Research Policy*, *37*(8), 1175-1187.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, *24*(2), 123-140.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5-32.

Carrasco, M., Chernozhukov, V., Gonçalves, S., & Renault, E. (2015). High dimensional problems in econometrics. *Journal of Econometrics*, *186*(2), 277-279.

Casani, F., De Filippo, D., García-Zorita, C., & Sanz-Casado, E. (2014). Public versus private universities: Assessment of research performance; case study of the Spanish university system. *Research Evaluation*, *23*(1), 48-61.

Cermeño, A. L. (2019). Do universities generate spatial spillovers? Evidence from US counties between 1930 and 2010. *Journal of Economic Geography*, *19*(6), 1173-1210.

Chatterton, P., & Goddard, J. (2000). The response of higher education institutions to regional needs. *European Journal of Education*, *35*(4), 475-496.

Chen, L. H. (2008). Internationalization or international marketing? Two frameworks for understanding international students' choice of Canadian universities. *Journal of Marketing for Higher Education*, *18*(1), 1-33.

Coulombe, P.G., Leroux, C., Stevanoic, D., & Surprenant, S. (2019). How is machine learning useful for macroeconomic forecasting? *CIRANO Working Paper*, 2019s-22.

Cuaresma, J. C., Doppelhofer, G., & Feldkircher, M. (2014). The determinants of economic growth in European regions. *Regional Studies*, *48*(1), 44-67.

Daraio, C., Bonaccorsi, A., Geuna, A., Lepori, B., Bach, L., Bogetoft, P., Cardoso, M. F., Castro-Martinez, E., Crespi, G., Fernandez de Lucio, I., Fried, H., Garcia-Aracil, A., Inzelt, A., Jongbloed, B., Kempkes, G., Llerena, P., Matt, M., Olivares, M., Pohl, C., Raty, T., Rosa, M. J., Sarrico, C. S., Simar, L., Slipersaeter, S., Teixeira, P. N. & Eeckaut, P. V. (2011). The European university landscape: A micro characterization based on evidence from the Aquameth project. *Research Policy*, *40*(1), 148-164.

De la Fuente, A., & Doménech, R. (2006). Human capital in growth regressions: How much difference does data quality make? *Journal of the European Economic Association*, *4*(1), 1-36.

Denti, D. (2010). R&D spillovers and regional growth. In R. Capello & P. Nijkamp (Eds.), *Handbook of regional growth and development theories* (pp. 211-236). Cheltenham, England: Edward Elgar Publishing.

Dias, D., & Amaral, A. (2014). Assessment of Higher Education Learning Outcomes (AHELO): An OECD Feasibility Study. In M. J. Rosa & A. Amaral (Eds.), *Quality Assurance in Higher Education* (pp. 66-87). London: Palgrave Macmillan.

Diebolt, C., & Hippe, R. (2019). The long-run impact of human capital on innovation and economic development in the regions of Europe. *Applied Economics*, *51*(5), 542-563.

Dill, D. D., & Teixeira, P. (2000). Program diversity in higher education: an economic perspective. *Higher Education Policy*, *13*(1), 99-117.

ETER. (2019). *Dual vs. unitary systems in Higher Education* (Report No. 3/2019). European Tertiary Education Register. https://www.eter-project.com/uploads/analytical-reports/ETER_AnalyticalReport_03_final.pdf

European Commission (2020). *Cohesion policy*. Retrieved from: https://ec.europa.eu/regional_policy/en/policy/what/glossary/c/cohesion-policy#:~:text=Cohesion%20policy%20is%20the%20European,its%20Member%20States%20and%20regions.&text=174)%2C%20the%20EU's%20cohesion%20policy,level%20of%20development%20between%20regions

Eurostat (2020). *NUTS - Nomenclature of territorial units for statistics*. Retrieved from: https://ec.europa.eu/eurostat/web/nuts/background

Evers, G. (2019). The impact of the establishment of a university in a peripheral region on the local labour market for graduates. *Regional Studies, Regional Science*, *6*(1), 319-330.

Fournier, J. (2016). The Positive Effect of Public Investment on Potential Growth. *OECD Economics Department Working Papers* (No. 1347). Paris, FR: OECD Publishing.

Gennaioli, N., La Porta, R., De Silanes, F. L., & Shleifer, A. (2014). Growth in regions. *Journal of Economic Growth*, *19*(3), 259-309.

Geppert, K., & Stephan, A. (2008). Regional disparities in the European Union: Convergence and agglomeration. *Papers in Regional Science*, *87*(2), 193-217.

Goldstein, H., & Drucker, J. (2006). The economic development impacts of universities on regions: do size and distance matter? *Economic Development Quarterly*, *20*(1), 22-43.

Guerrero, M., Cunningham, J. A., & Urbano, D. (2015). Economic impact of entrepreneurial universities' activities: An exploratory study of the United Kingdom. *Research Policy*, *44*(3), 748-764.

Guironnet, J. P., & Peypoch, N. (2018). The geographical efficiency of education and research: The ranking of US universities. *Socio-Economic Planning Sciences*, *62*(2018), 44-55.

Hanushek, E. A. (2016). Will more higher education improve economic growth? *Oxford Review of Economic Policy*, *32*(4), 538-552.

Hanushek, E. A., & Kimko, D. D. (2000). Schooling, labor-force quality, and the growth of nations. *American Economic Review*, *90*(5), 1184-1208.

Hanushek, E. A., & Woessmann, L. (2020). Education, knowledge capital, and economic growth. *The Economics of Education,* 171-182, New York: Academic Press.

Hanushek, E. A., & Wößmann, L. (2010). Education and Economic Growth. In P. Peterson, E. Baker & B. McGaw (Eds.), *International Encyclopedia of Education* (volume 2, pp. 245-252). Oxford, UK: Elsevier.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction*. Berlin, DE: Springer Science & Business Media.

Hegde, D. (2005). Public and private universities: unequal sources of regional innovation? *Economic Development Quarterly, 19*(4), 373-386.

Hermannsson, K., Lisenkova, K., Lecca, P., McGregor, P. G., & Swales, J. K. (2017). The external benefits of higher education. *Regional Studies, 51*(7), 1077-1088.

Holmes, C. (2013). Has the expansion of higher education led to greater economic growth? *National Institute Economic Review, 224*(1), R29-R47.

Horváth, K., & Berbegal-Mirabent, J. (2020). The role of universities on the consolidation of knowledge-based sectors: A spatial econometric analysis of KIBS formation rates in Spanish regions. *Socio-Economic Planning Sciences*, 100900. DOI: 10.1016/j.seps.2020.100900.

Hurwicz, L. (1950). Least-squares bias in time series. In T. C. Koopmans (Ed.), *Statistical inference in dynamic economic models*. New York, NY: Wiley.

Iammarino, S., Rodríguez-Pose, A., & Storper, M. (2019). Regional inequality in Europe: evidence, theory and policy implications. *Journal of Economic Geography*, *19*(2), 273-298.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning*. New York, NY: Springer.

Jamison, E. A., Jamison, D. T., & Hanushek, E. A. (2007). The effects of education quality on income growth and mortality decline. *Economics of Education Review*, *26*(6), 771-788.

Keller, K. R. (2006). Investment in primary, secondary, and higher education and the effects on economic growth. *Contemporary Economic Policy*, *24*(1), 18-34.

Kivinen, O., & Rinne, R. (1996). The Problem of Diversification in Higher Education: Countertendencies Between Divergence and Convergence. In V. Lynn Meek, L. Goedegebuure, O. Kivinen & R. Rinne (Eds.), *The Mockers and Mocked: Comparative Perspectives on Differentiation; Convergence and Diversity in Higher Education* (pp. 95-116). Guildford, UK: IAU Press and Pergamon.

Krueger, A. B., & Lindahl, M. (1998). *Education and growth: why and for whom?* Princeton, NJ: Princeton University.

Leask, B., & Bridge, C. (2013). Comparing internationalisation of the curriculum in action across disciplines: Theoretical and practical perspectives. *Compare*, *43*(1), 79-101.

Lemon, S. C., Roy, J., Clark, M. A., Friedmann, P. D., & Rakowski, W. (2003). Classification and regression tree analysis in public health: methodological review and comparison with logistic regression. *Annals of Behavioral Medicine*, *26*(3), 172-181.

Lendel, I. (2010). The impact of research universities on regional economies: The concept of university products. *Economic Development Quarterly*, *24*(3), 210-230.

Lepori, B. (2021). The heterogeneity of European Higher Education Institutions: a configurational approach. *Studies in Higher Education*, 1-17. DOI: 10.1080/03075079.2021.1968368.

Lilles, A., & Rõigas, K. (2017). How higher education institutions contribute to the growth in regions of Europe? *Studies in Higher Education*, *42*(1), 65-78.

Lucas, R. E. (1988). On the Mechanics of Economic Development. *Journal of Monetary Economics*, 22(1), 3-42.

Malva, A. D., & Carree, M. (2013). The spatial distribution of innovation: evidence on the role of academic quality for seven European countries. *Economics of Innovation and New Technology*, *22*(6), 601-618.

Mankiw, N. G., Romer, D., & Weil, D. (1992). A Contribution to the Empirics of Economic Growth. *Quarterly Journal of Economics*, *107*(2), 407-37.

Marrocu, E., & Paci, R. (2010). The effects of public capital on the productivity of the Italian regions. *Applied Economics*, *42*(8), 989-1002.

Martin, B. R. (2012). Are universities and university research under threat? Towards an evolutionary model of university speciation. *Cambridge Journal of Economics*, *36*(3), 543-565.

Molas-Gallart, J., Salter, A., Patel, P., Scott, A., & Duran, X. (2002). Measuring Third Stream Activities. *Final Report to The Russell Group of Universities*. Brighton, UK: SPRU, University of Sussex.

Mueller, P. (2006). Exploring the knowledge filter: How entrepreneurship and university–industry relationships drive economic growth. *Research Policy*, *35*(10), 1499-1508.

Mullainathan, S., & Spiess, J. (2017). Machine learning: An applied econometric approach. *Journal of Economic Perspectives*, 31(2), 87-106.

OECD (2018). *OECD Regions and Cities at a Glance 2018*. Paris, France: OECD Publishing. https://doi.org/10.1787/reg_cit_glance-2018-en

Owens, D. L., Srivastava, P., & Feerasta, A. (2011). Viewing international students as state stimulus potential: Current perceptions and future possibilities. *Journal of Marketing for Higher Education*, *21*(2), 157-179.

Pelinescu, E. (2015). The impact of human capital on economic growth. *Procedia Economics and Finance*, *22*(2015), 184-190.

Perkmann, M., King, Z., & Pavelin, S. (2011). Engaging excellence? Effects of faculty quality on university engagement with industry. *Research Policy*, *40*(4), 539-552.

Pohl, H. (2021). Internationalisation, innovation, and academic–corporate co-publications. *Scientometrics*, *126*(2), 1329-1358.

Polyakov, M., Polyakov, S. & Iftekhar, M.S. (2017). Does academic collaboration equally benefit impact of research across topics? The case of agricultural, resource, environmental and ecological economics. *Scientometrics, 113*(3), 1385-1405.

Puuska, H. M., Muhonen, R., & Leino, Y. (2014). International and domestic co-publishing and their citation impact in different disciplines. *Scientometrics*, *98*(2), 823-839.

Restaino, M., Vitale, M. P., & Primerano, I. (2020). Analysing International Student Mobility Flows in Higher Education: A Comparative Study on European Countries. *Social Indicators Research*, *149*(3), 947-965.

Romer, P. (1990). Capital, Labor, and Productivity. *Brookings Papers on Economic Activity. Microeconomics, 1990*, 337-367.

Roodman, David. (2008) xtabond2: Stata Module to Extend Xtabond Dynamic Panel Data Estimator. Washington: Center for Global Development.

Rossi, F. (2010). Massification, competition and organizational diversity in higher education: evidence from Italy. *Studies in Higher Education*, *35*(3), 277-300.

Rossi, F., & Goglio, V. (2020). Satellite university campuses and economic development in peripheral regions. *Studies in Higher Education*, *45*(1), 34-54.

Santoalha, A., Biscaia, R., & Teixeira, P. (2018). Higher education and its contribution to a diverse regional supply of human capital: does the binary/unitary divide matters? *Higher Education*, *75*(2), 209-230.

Schmoch, U., & Schubert, T. (2008). Are international co-publications an indicator for quality of scientific research? *Scientometrics*, *74*(3), 361-377.

Schubert, T., & Kroll, H. (2016). Universities' effects on regional GDP and unemployment: The case of Germany. *Papers in Regional Science*, *95*(3), 467-489.

Seeber, M., Meoli, M., & Cattaneo, M. (2020). How do European higher education institutions internationalize? *Studies in Higher Education*, *45*(1), 145-162.

Shi, T., & Horvath, S. (2006). Unsupervised learning with random forest predictors. *Journal of Computational and Graphical Statistics, 15*(1), 118-138.

Sianesi, B., & Van Reenen, J. (2003). The returns to education: Macroeconomics. *Journal of Economic Surveys*, *17*(2), 157-200.

Steinacker, A. (2005). The economic effect of urban colleges on their surrounding communities. *Urban Studies*, 42(7), 1161-1175.

Stiglitz, J. E., Sen, A., & Fitoussi, J. P. (2009). Report by the commission on the measurement of economic performance and social progress.

Teichler, U. (2008). Diversification? Trends and explanations of the shape and size of higher education. *Higher Education*, *56*(3), 349-379.

Teichler, U. (2009). Internationalisation of higher education: European experiences. *Asia Pacific Education Review*, *10*(1), 93-106.

Teichler, U. (2010). Internationalising higher education: Debates and changes in Europe. In D. Mattheou (Ed.), *Changing educational landscapes: Educational policies, schooling*

systems and higher education-a comparative perspective (pp. 263-283). Dordrecht, NLD: Springer.

Teixeira, P., Rocha, V., Biscaia, R., & Cardoso, M. F. (2014). Policy changes, marketisation trends and spatial dispersion in European higher education: comparing public and private sectors. *Cambridge Journal of Regions, Economy and Society*, *7*(2), 271-288.

Thursby, J. G., & Kemp, S. (2002). Growth and productive efficiency of university intellectual property licensing. *Research Policy*, *31*(1), 109-124.

Tijssen, R. J., Visser, M., & Van Leeuwen, T. (2002). Benchmarking international scientific excellence: Are highly cited research papers an appropriate frame of reference? *Scientometrics*, *54*(3), 381-397.

Tijssen, R. J., Yegros-Yegros, A., & Winnink, J. J. (2016). University–industry R&D linkage metrics: validity and applicability in world university rankings. *Scientometrics*, *109*(2), 677-696.

Ullah, S., Akhtar, P., & Zaefarian, G. (2018). Dealing with endogeneity bias: The generalized method of moments (GMM) for panel data. *Industrial Marketing Management*, *71*, 69-78.

Valero, A., & Van Reenen, J. (2019). The economic impact of universities: Evidence from across the globe. *Economics of Education Review*, *68*, 53-67.

Vieira, E. S., & Lepori, B. (2016). The growth process of higher education institutions and public policies. *Journal of informetrics*, *10*(1), 286-298.

Wolf, A. (2002). *Does education matter? Myths about education and economic growth*. London, UK: Penguin.

Xie, Y., Fang, M., & Kimberlee, S. (2015). STEM education. *Annual Review of Sociology*, *41*, 331-357.

Yen, S. H., Ong, W. L., & Ooi, K. P. (2015). Income and employment multiplier effects of the Malaysian higher education sector. *Margin: The Journal of Applied Economic Research*, *9*(1), 61-91.

Yusuf, S., & Nabeshima, K. (2006). *How universities promote economic growth*. Washington, US: The World Bank.

Annex

Section A1. Correlation analysis

Table A1 reports the coefficients of the pairwise correlation matrix associated with the variables included in the empirical model of equation (1).

The results of the analysis detect significant correlations among HESs variables. In particular, Table A1 highlights that HESs with the prevalence of applied science universities tend to be less research-oriented, are associated with a lower number of students and composted of relatively new universities. Instead, regions with a significant presence of students in private universities are more likely to achieve better graduation rates and to offer generalist programmes, less focused on STEM disciplines. In turn, HESs specialised in STEM are negatively correlated with graduation rates but tend to employ a greater amount of physical resources (low student-teacher ratio). Moreover, the indicators of research performance are highly correlated between each other and with the number of enrolled students, especially in doctoral courses; whereas, high shares of Erasmus students (both incoming and outgoing) seem to be associated with richer universities, with low student-teacher rates.

Looking at the relationships between HES indicators and the dependent variable, the regional GDP per capita, we can detect a significant association between richer regions and HESs that report a high amount of physical resources and better performance in terms of research collaboration with industries. However, regional GDP per capita shows higher positive correlations with control factors – in particular with employment rate (with a coefficient 0.84). The control factors are also significantly correlated between each other (except migration rate) and with the HES variables. In particular, it is worth to highlight that regions with high levels of human capital (hc64_n2) are positively associated with the share of students in private institutions and the graduation rates of HESs.

	ln(GDP per capita)	ln(Number of universities)	Share of private students	Medical Universities (%)	Applied science universities (%)	ln(Students enrolled)	Graduation rate	STEM (%)	ln(Doctoral students enrolled)	Erasmus students incoming (%)	Erasmus students outgoing (%)	ln(Students/teacher ratio)	Publications per researcher	Top 10% documents	International publications (%)	Industry collaboration (%)	Employment rate	ln(Population density)	ln(Fixed capital)	Human Capital (%)	Net migration rate
ln(GDP per capita)	1																				
ln(Number of universities)	0.2491	1																			
Share of private students	-0.0184	-0.0892	1																		
Medical Universities (%)	-0.1359	-0.0494	-0.0643	1																	
Applied science universities (%)	0.2213	0.0885	-0.112	-0.4459	1																
ln(Students enrolled)	0.144	0.4858	-0.1522	0.4055	-0.4861	1															
Graduation rate	-0.1332	-0.0588	0.5061	0.0045	-0.1913	-0.0402	1														
STEM (%)	0.1311	0.0259	-0.3335	-0.1454	0.1856	0.0184	-0.3104	1													
ln(Doctoral students enrolled)	0.015	0.4047	-0.0309	0.4557	-0.7293	0.7867	0.0954	-0.0896	1												
Erasmus students incoming (%)	0.0692	0.1198	0.0188	0.0226	-0.0348	0.0106	0.1941	-0.0102	0.1217	1											
Erasmus students outgoing (%)	0.0807	0.1214	-0.1168	0.0337	0.0815	-0.0816	-0.1035	-0.0506	-0.0247	0.5829	1										
ln(Students/teacher ratio)	-0.3971	-0.1947	-0.0301	0.0648	-0.2161	0.1624	-0.002	-0.2188	0.0244	-0.275	-0.2859	1									
Publications per researcher	-0.1072	0.1336	0.0728	0.6178	-0.6274	0.5213	0.1346	-0.0724	0.6815	0.1424	-0.0089	0.1199	1								
Top 10% documents	0.0452	0.2644	-0.0049	0.3424	-0.6414	0.5725	0.0454	-0.1178	0.7316	0.088	-0.0147	0.0179	0.6032	1							
International publications (%)	0.0693	0.3411	-0.0499	0.3509	-0.6813	0.6121	0.0659	-0.0946	0.8256	0.1283	-0.0161	-0.0657	0.6359	0.8046	1						
Industry collaboration (%)	0.2753	0.3006	-0.0643	0.2199	-0.3502	0.4351	-0.0633	0.1061	0.5338	0.0582	-0.1023	-0.2001	0.392	0.5101	0.5992	1					
Employment rate	0.8394	0.2352	-0.1157	-0.1488	0.1973	0.128	-0.1538	0.1685	0.0599	-0.0002	0.0578	-0.3933	-0.1535	0.0313	0.0596	0.2233	1				
ln(Population density)	0.4454	0.285	0.2435	0.0711	-0.0973	0.4065	-0.0642	0.122	0.3086	-0.0786	-0.0602	-0.0883	0.1482	0.212	0.193	0.29	0.451	1			
ln(Fixed capital)	0.4681	0.0918	-0.0933	-0.029	0.1939	0.0755	-0.2339	0.1488	-0.096	-0.0599	-0.0368	-0.1353	-0.042	-0.0357	-0.069	0.1806	0.279	0.3308	1		
Human Capital (%)	0.3294	0.1318	0.4104	-0.1528	0.1115	0.0843	0.4289	-0.0593	0.0538	0.0975	-0.1624	-0.137	-0.0491	-0.0227	0.0096	0.0579	0.2713	0.2359	0.1301	1	
Net migration rate	0.0644	0.0199	0.0174	-0.0136	0.0108	-0.0401	-0.0009	-0.0527	-0.0241	-0.0166	0.0088	-0.0144	-0.0085	-0.0192	-0.0101	0.0167	0.0339	-0.0455	0.0329	-0.0081	1

Table A1. Pairwise correlation matrix/ *Note:* the values highlighted with a darker colour are associated with significant positive or negative correlations. *Source:* Produced by the authors using Stata 14.

Section A2. Systematic process for aggregating ETER and InCites data

The preparation of the data for empirical analyses has required a great effort. The most critical part of this activity concerned the aggregation of the data at institutional level (i.e. ETER and InCites indicators) to the regional NUTS-3 level. In order to be transparent and provide a robust set of data, we followed a systematic approach for the aggregation.

The first step aimed at matching the selected institutions in ETER with the ones available in InCites. In the case of a missing match, we needed to understand if the missing value was due to the lack of documents published in Web of Science by the specific university. This phenomenon is particularly frequent for applied science universities which, by definition, are less focused on research activities[11] (ETER, 2019). Thus, we considered InCites indicators as missing, only for category 1 universities (i.e. not applied science universities), defined as "research active" institutions in ETER, and offering at least one doctoral course[12]. Following these criteria, we found about 60 universities (3.8% of the total sample) presenting missing values for InCites variables. The remaining universities without a match were about 600 (representing 40% of the total sample) and in 80% of the cases are applied science institutions. These universities are very likely to have not published documents in Web of Science and, therefore, their research indicators have been imputed to zero.

ETER also provides the code of NUTS-3 region of each university, which is needed to aggregate the data at regional level. In the case of multi-campus institutions (which are 25.7% of the total number of universities, located in 290 NUTS-3 regions), we considered only the location of the main campus[13]. Most of the variables aggregated from institutional to regional level are indicators expressed in absolute number and, therefore, we simply summed the values for the universities belonging to the same region. The sum was weighted based on the size of each university, expressed by the number of students, or (only for research indicators) on the respecting number of documents published in Web of Science. Instead, for indicators expressed through relative measures, the aggregated indicators are given by the mean of the observations belonging to the same region. Finally, a variable has been considered as missing for a certain region when the universities with missing data represent more than 10% of the total students in the HES. The cut-off is 5% when information on the number of students is

[11] Universities of applied science are characterised by a strong professional and vocational orientation and, for this reason, research does not represent a necessary condition. Indeed, many universities of applied science are not legally authorised to award doctoral degrees (ETER, 2019).
[12] i.e. reporting a number of PhD students greater than zero in ETER. Offering a doctoral course is a proxy for the research activity of universities.
[13] The location of the main campus is defined in ETER database.

not available for all the universities in the region (40% of HESs in the sample) and, therefore, we consider the share of universities with missing data over the total number of universities in the region.

Section A3. Semiparametric analysis

The completed model (see RF7 in Table 6, GMM4 in Table 5, and R7 in Table 4) has been estimated by employing a semiparametric technique in order to provide evidence of inference of the nonlinear interactions. Indeed, semiparametric models can provide estimates of the confidence intervals, even if they accommodate nonlinear effects. However, it is worth noticing that semiparametric approaches cannot control for the endogeneity, as fully parametric techniques do (e.g. GMM approach). Therefore, this limitation must be considered when interpreting the results of the model, where HES variables have a potential problem of endogeneity.

On the other hand, many of the advantages of random forest do not hold for semiparametric techniques (see Section 3.2). For instance, semiparametric models cannot properly handle the high number of HES variables in the model and do not provide information on the importance of the nonlinear regressors. In addition, semiparametric models are not as flexible as fully nonparametric approaches are. Semiparametric techniques require specifying *a priory* which covariates are interacting nonlinearly with the response (Ruppert et al., 2003). For the analysis presented in this section, we chose the nonparametric terms based on the random forest results in Section 5.3.

The semiparametric results, reported in Table A2, have been estimated by employing Generalized Additive Model (GAM) (Hastie and Tibshirani, 1986; 1990). GAM accommodates both linear and nonlinear relationships through an additive model, where the interactions between covariates and the dependent variable follow smooth functions. Relying on GAM flexibility, the smooth functions can be linear or nonlinear, depending on the data analysed. In detail, a general GAM structure is defined as:

$$g\big(E(y)\big) = \beta + s_1(x_1) + \cdots + s_p\big(x_p\big)$$

Where *y* is the dependent variable and $E(y)$ the expected values of the response. The function $g()$ is called link function, since it defines the relationship between the predictors (x_i) and $E(y)$. Finally, the terms $s_i(x_i)$ denote the smooth functions. Following the most common

approach, we fit the smooth functions in GAM by employing penalised splines (Ruppert et al., 2003).

The results obtained by GAM estimates confirm the existence of nonlinear relationships between the higher education variables and the regional GDP per capita. In particular, the results show that the spline functions are properly fitting the data, with significant p-values for most of the HES variables. Nevertheless, some of the nonparametric variables, such as the share of incoming/outgoing Erasmus student, are not significant. This lack of significance can be due, in part, to the large number of highly correlated covariates included in the model. In other words, the effects of these non-significant predictors could be captured by the covariates with statistically significant effects.

The estimated degrees of freedom (edf) in Table A2 confirm the high complexity of the interactions between the HES variables and the predictor, with degrees significantly higher than 1 (with edf=1 suggesting a linear effect). The existence of nonlinear effects has also been formally tested by the F-test. This test compares the GAM model here described to the full parametric model (i.e. a Generalised linear model) where all the covariates are supposed to linearly influence the response. As shown in Table A3, the small p-value demonstrates that the GAM smooth function has a better fit compared to the fully parametric model.

The shapes of the penalised spline functions describing the effects of the main nonparametric variables are reported in Figure A1. As observed in the partial dependence plots in Section 5.2, the logarithm of the GDP per capita in the region varies at different values of a specific predictor, averaging out all the other regressors in the model. Comparing these graphs with the partial plots in Figure 3, we can recognise similar paths. The existence of a saturation level for the number of universities in the region is verified by the first plot in Figure A1 – even if it is higher than the one found for RF7. Indeed, the semiparametric model identifies the maximum effect on local economic development around 16 institutions. The GAM estimates also confirm the existence of a maximum effect of around 14 students per academic staff. Also the number of enrolled students and the number of publications for the researcher follow similar paths to the ones identified by random forest. On the other hand, differences between the partial plots provided by the two approaches seem to shed a light on the higher flexibility of random forest compared to the semiparametric model. This is confirmed by the data fitting associated with the two different models. As reported in Table A2, the GAM model has an R squared of 91.5%, while random forest (RF7) explains 94.14% of the variance of the dependent variable.

Parametric coefficients:					
	Estimate	Std. Error	t-value	Pr(>|t|)	
(Intercept)	8.836265	0.060258	146.642	2.00E-16	***
Human Capital (%)	0.012178	0.00073	16.679	2.00E-16	***
Employment rate	0.021794	0.000406	53.748	2.00E-16	***
ln(Population density)	0.018251	0.003892	4.69	2.95E-06	***
ln(Fixed capital)	0.018326	0.006168	2.971	0.003012	**
Country fixed effects	yes				
Approximate significance of smooth terms:					
	edf	Ref.df	F	p-value	
ln(Number of universities)	4.874	5.73	7.235	2.79E-07	***
Share of private students	2.035	2.457	8.313	0.000111	***
Medical universities (%)	8.831	8.976	3.819	7.51E-05	***
Applied science universities (%)	7.322	8.204	1.606	0.128814	
ln(Students enrolled)	7.974	8.696	4.27	2.71E-05	***
Graduation rate	7.885	8.691	2.239	0.016112	*
STEM (%)	8.471	8.916	3.475	0.001744	**
Erasmus students incoming (%)	3.535	4.477	1.183	0.296411	
Erasmus students outgoing (%)	2.16	2.768	0.306	0.737812	
ln(Students/teacher ratio)	6.589	7.686	3.072	0.001483	**
ln(Doctoral students enrolled)	7.441	8.391	2.603	0.011811	*
Publications per researcher	7.488	8.411	49.598	2.00E-16	***
Top 10% documents	4.756	5.74	0.825	0.487159	
International publications (%)	4.264	5.347	1.656	0.140091	
Industry collaboration (%)	1	1.001	0.13	0.71874	
Net migration rate	7.315	8.331	7.702	2.09E-10	***

Table A2. GAM regression results./ *Note:* R-sq.(adj) = 0.915, Deviance explained = 92.1%, n = 1808, GACV = 0.019585, Scale est. = 0.018328. Estimated degree of freedom (edf) and Reference degrees of freedom (Ref.df) are reported for the smooth terms. Significance levels: 0 '***' 0.001; '**' 0.01; '*' 0.05; '.' 0.1; ' ' 1

Model	Resid. Df	Resid. Dev.	Df	Deviance	F	Pr(>F)	
Parametric model	1760	43.634					
Semiprametric model	1672.2	30.866	87.822	12.768	7.9323	<2.2E-16	***

Table A3. Analysis of deviance table, F-test./ *Note:* Significance levels: 0 '***' 0.001; '**' 0.01; '*' 0.05; '.' 0.1; ' ' 1. *Source*: Produced by the authors using R.

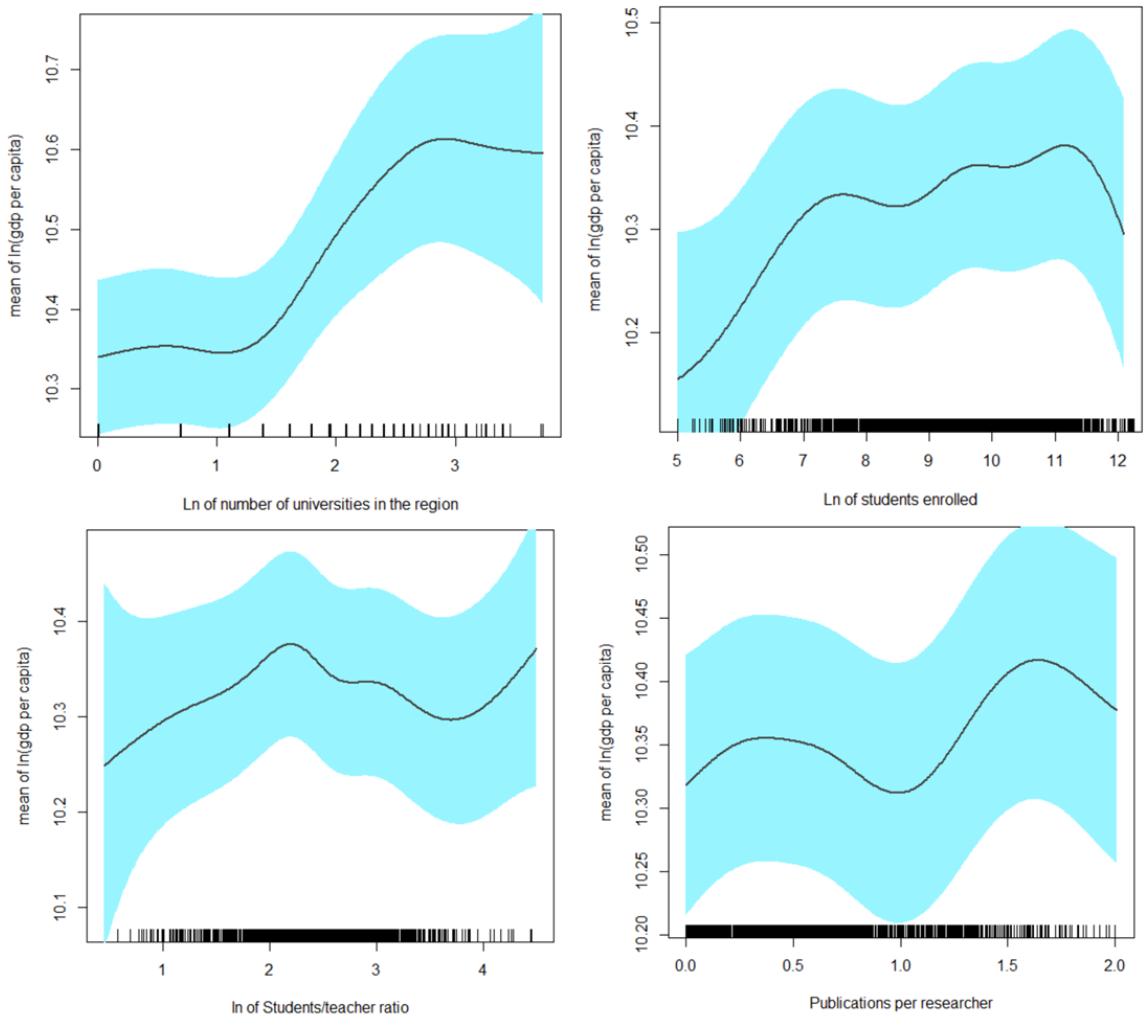

Figure A1. Estimated smooth function components./*Note:* The blue regions show the 95% pointwise confidence intervals, while the tick marks indicate the values of the predictors in the sample. *Source*: Produced by the authors using R.

Section A4. Evidence among countries

The analyses presented in Section 5 aims at offering communal evidence of the influence of HES on the local economic development of European countries. Nevertheless, the analysis of country fixed effects can provide interesting evidence on differences among national HESs. Indeed, based on model RF7, we performed the joint partial plots by considering the joint effect of country fixed effects and different HES variables on the regional GDP per capita. In general, the analysis suggests that higher educational characteristics influence local economic development in the same way, regardless of the country considered. On the contrary, country fixed effects have an impact only on the average level of the regional GDP per capita. This specific behaviour is shown in Figure A2, reporting the joint partial plot for the number of publications per researcher and country fixed effects. More rarely, the national system in

which universities are located influences the size of the effect. This is the case of the number of enrolled students (see Figure A3), where the HES size seems to have a greater economic impact in countries with low economic outputs, such as Bulgaria (BG), Latvia (LV), Greece (GR) and Slovakia (SK).

In summary, the results shed light on the possibility to recognise a communal model through which HES characteristics influence the local economic development of European countries, while differences among countries are mainly due to the heterogeneity in the levels of economic output.

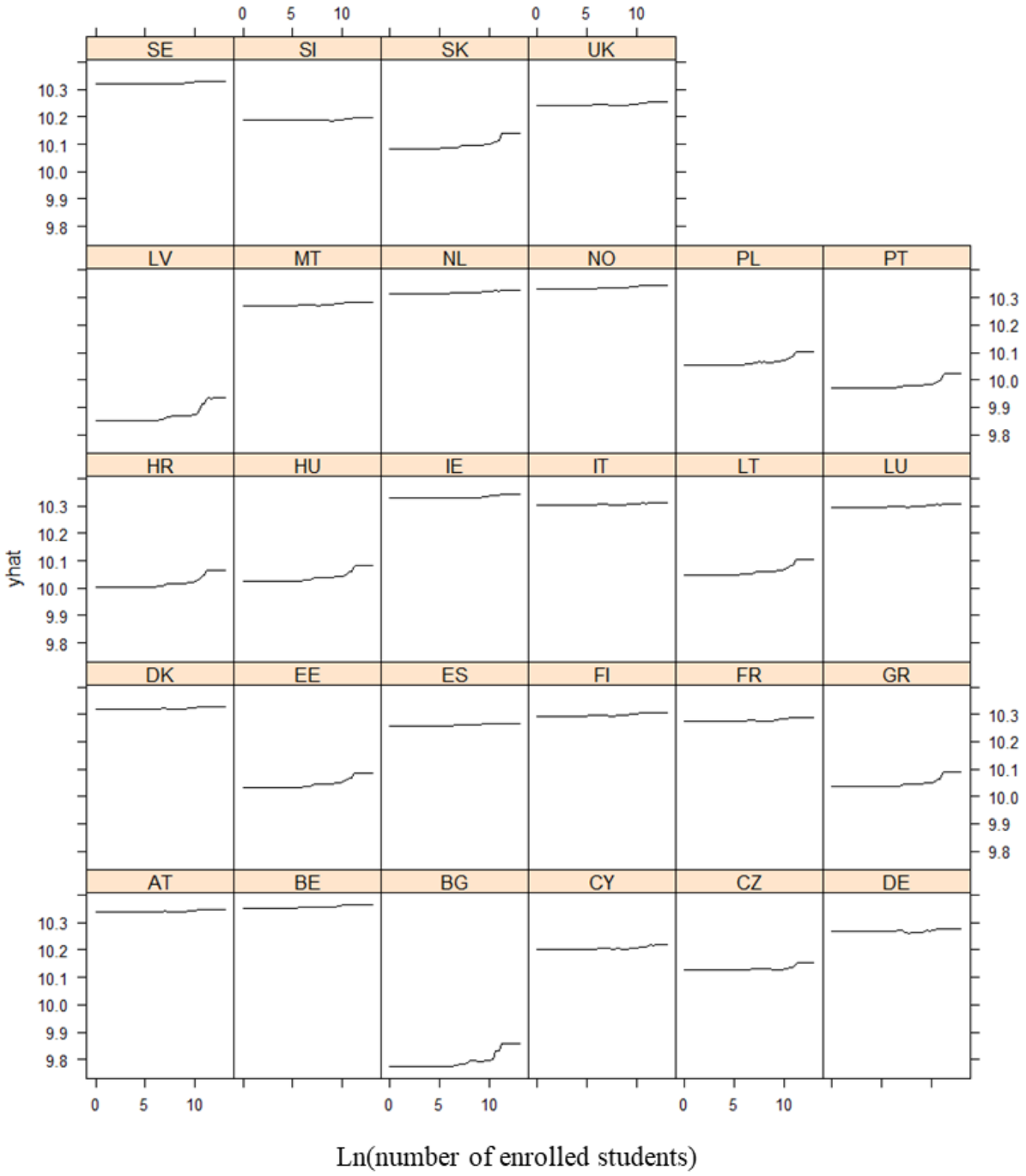

Figure A2. Joint partial plot of HES size and country fixed effects./ Note: the colour represents the scale of the values of the response. Source: Produced by the authors using R.

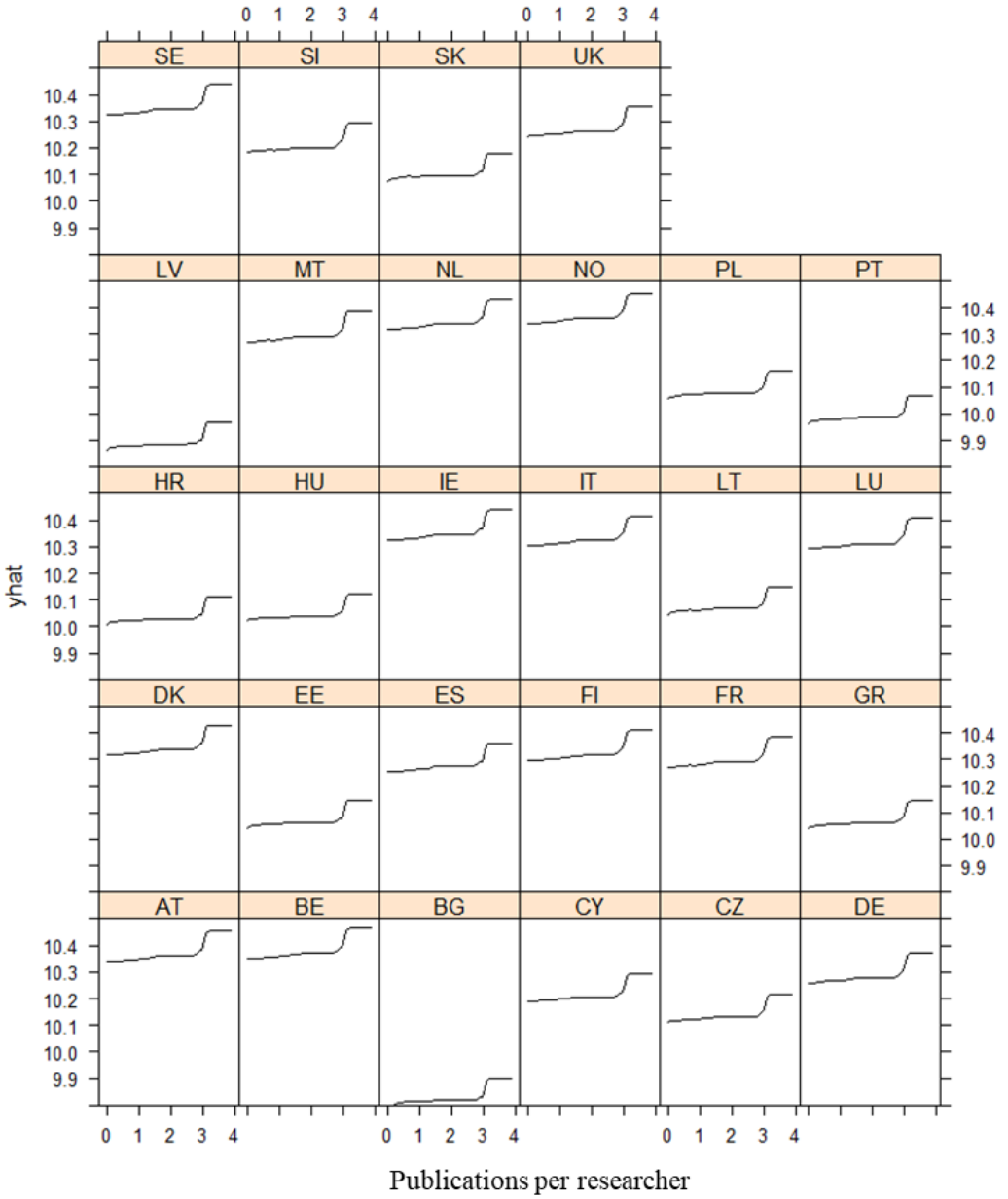

Figure A2. Joint partial plot of the research productivity and country fixed effects./ Note: the colour represents the scale of the values of the response. Source: Produced by the authors using R.

Reference of the Annex

Hastie, T., & Tibshirani, R. (1987). Generalized additive models: some applications. *Journal of the American Statistical Association*, *82*(398), 371-386.

Hastie, T., & Tibshirani, R. (1990). Exploring the nature of covariate effects in the proportional hazards model. *Biometrics*, *46*(4), 1005-1016.

Ruppert, D., Wand, M. P., & Carroll, R. J. (2003). *Semiparametric regression* (No. 12). Cambridge, UK: Cambridge university press.