Understanding EU's Regional Potential in Low-Carbon Technologies

Enrico Bergamini* and Georg Zachmann†

September 2019

Draft Paper

Abstract

This research builds on the regional innovation literature, and aims at better understanding of the potential and development of low-carbon technologies in the European Union. Exploiting OECD's REGPAT for regionalised patent data, we estimate the potential advantage in 14 green techonologies for European NUTS2 regions. We use network proximity between both technologies and the regions to understand clusters of revealed technological advantage, and build the regressors for estimating potential advantage via Zero-inflated Beta Regressions. Subsequently, we construct a dataset of lagged potentials and labour market, economic and demographic variables and we perform an Elastic net regularization to understand the association with current revealed advantage. In addition, we explore the region-technology networks, finding two gravity centers for green innovations in France and Germany's industrial and high-tech hubs (Île de France, Stuttgart and Oberbyern). Our approach indicates an association of technological advantage in green technologies with the (lags of) participation rates to labour market, sectoral employment in STEM, general higher education, duration of employment, percentage of GDP spend in R&D (public and private) and intramural expenditure in R&D.

1 Introduction

Keeping global temperature increase below 2°C will require an almost complete decarbonisation of our energy system in the foreseeable future (early in the second half of this century). There will, therefore, be a growing market for different kinds of low-carbon technologies (vehicles, power plants, appliances, batteries...) that will replace the existing stock of high-carbon technologies. Policies to develop and deploy corresponding technologies might not only speed up the necessary learning but also enable the development of comparative advantages.

*enrico.bergamini@bruegel.org, Bruegel
†georg.zachmann@bruegel.org, Bruegel

The urge of investigating how to foster the development of low-carbon technologies lies not only in the climate argument but also in the possibility for Europe to compete in the low-carbon technologies industrial race. Advantages in specific technologies typically develop in industrial clusters and not all regions have the potential to excel in all low-carbon technologies. Hence, it might be sensible for regional and national policymakers to make educated guesses on what type of support might fall on the most fertile ground in their jurisdiction.

Our analysis relies on systematic evidence originating from the regional growth literature triggered by Hidalgo et al. (2007), and builds on the analytical approaches in patent analysis for low-carbon technologies based on similar regions' current advantage. If policy makers want to create or strengthen comparative advantages, they need to understand the current and potential regional specializations. Our paper firstly explores this dimension, in order to estimate the potential for low-carbon technologies inside European countries. Hence, our first research question is:

- What are the current and potential advantages of EU regions for sectoral specialisation in low-carbon technologies?

Secondly, it investigates the policy, labour market, and institutional aspects that might lead regions to create, realise and exploit this potential, leading to our second research question:

- Which labour market, economic and demographic conditions are associated, together with potential specialisation, to a stronger relative technological advantage?

The novelty that we present with the respect to previous literature is a more granular specification that allows us to understand the evolution of potential and advantage of regional hubs and innovation clusters beyond the national level. Our analysis is based on a two-stage approach in which we first estimate potential and revealed green specialisation and subsequently select labour market, demographic and economic variables that are associated with it.
The paper proceeds as follows: Section 2 explains the data sources used, Section 3 contains the empirical strategy that we apply for calculating potential regional advantage, Section 4 explains our methodology for the second stage regressions in which we perform a data-driven selection of variables, and in Section 5 the results are discussed. Section 6 concludes and discusses possible policy implications.

2 Data

Innovation activity is approximated by the number of patents filed in a specific patent category in a region. Patent data stems from the OECD's REGPAT database, that allows us to have more patent data geolocated to NUTS2 and NUTS3 level, compared to EPO's PATSTAT. REGPAT contains patents listed under the Patent Cooperation Treaty (PCT) and the European Patent Office (EPO). We combine the patents from both sources, preferring EPO over PCT by keeping the PCT entries only where the patent is not filed under both.

The analysis is based on patent technology codes classified by the Cooperative Patent Classification (CPC) scheme and the International Patent Classification (IPC). We use IPC definition for all the technologies and we use only the Y class CPC codes to identify low-carbon techonologies. The number of patents attributed to a region is based on the location of patent inventors that applied at the EPO or international patents under the Patent Cooperation Treaty (PCT). The earliest application of individual patent families is used and attributed in fractions to all inventor countries and technology codes. We select the definition of low-carbon technologies based on the Joint Reseearch Center's definition, as in Fiorini et al (2017). CPC codes are grouped for 14 technologies, namely solar panels, hydrogen-related technologies, solar and thermal energy, wind energy, hydro energy, energy management, efficient lighting, efficient heating and cooling, combustion, residential insulation, bio-fuels, batteries, electric cars, efficient rail transport and nuclear energy. The list of relevant CPC-Y codes is in table A.1 of the Appendix.

In the following sections, we will give detail on the estimations of revealed technological advantage and of potential technological advantage, which will be included respectively in the left and right-hand side of our final dataset. The economic, demographic, labour market variables that we include in our dataset are instead based on the full Eurostat and Urban Data Platform[1] databases, and cover a very wind range of fields.

3 Empirical strategy

3.1 Estimation of potential advantage

Revealed Technological Advantage (RTA) is measured à-la-Balassa, calculating the relative specialisation in patents of a region, in the same fashion exports are used to build Revealed Comparative Advantage (RCA). In this section, our methodologies relies and builds on Zachmann and Roth (2018). The revealed advantage in a technology of a country is defined by a fraction of two shares:

$$RTA = \frac{\frac{x_{il}}{\sum_i x_{il}}}{\frac{\sum_l x_{il}}{\sum_{il} x_{il}}} \tag{1}$$

where:

x_{il} is the number of patents of technology i in region l

$\sum_l x_{il}$ is the sum of patents of technology i across all regions

$\sum_{il} x_{il}$ is the sum of all patents across all regions

All the RTAs are generated while excluding green technologies from the sample, as in Hausmann (2007). RTAs are subsequently standardised in the following fashion:

[1] The Urban Data Platform was created with the joint efforts of the European's Commission Joint Research Center and the Directorate General for Regional and Urban Policy (DG REGIO). Available at: https://urban.jrc.ec.europa.eu

$$sRTA = \frac{\frac{RTA-1}{RTA+1} + 1}{2} \tag{2}$$

where RTA is derived as in equation (1).

In order to estimate the potential technological advantage of regions (pRTA), we apply the methodology inspired by Hausmann et al (2014) and thus alike to calculating pRTA, bringing the work of Zachmann and Roth (2018) to a sub-national level. This methodology assumes a relationship between the comparative advantage of different products or technologies. For instance, a region's comparative strength in one product can imply a potential strength of another product, given that there is a link, a similarity, either between the products or regions. Our measures are based on four different levels of measurement (country, NUTS region, inventor, application) and different methods to calculate the links between them.

The intuition behind potential advantage, is the attempt to estimate correlations between regions and technologies that are based on latent factors that are unknown a priori. For example, the latent factors that make regions similar could be factor costs, infrastructures, geography, domestic market sizes. These correlations could also be based on technological links (e.g. similar value chains, technological spill-overs, degree of complexity).

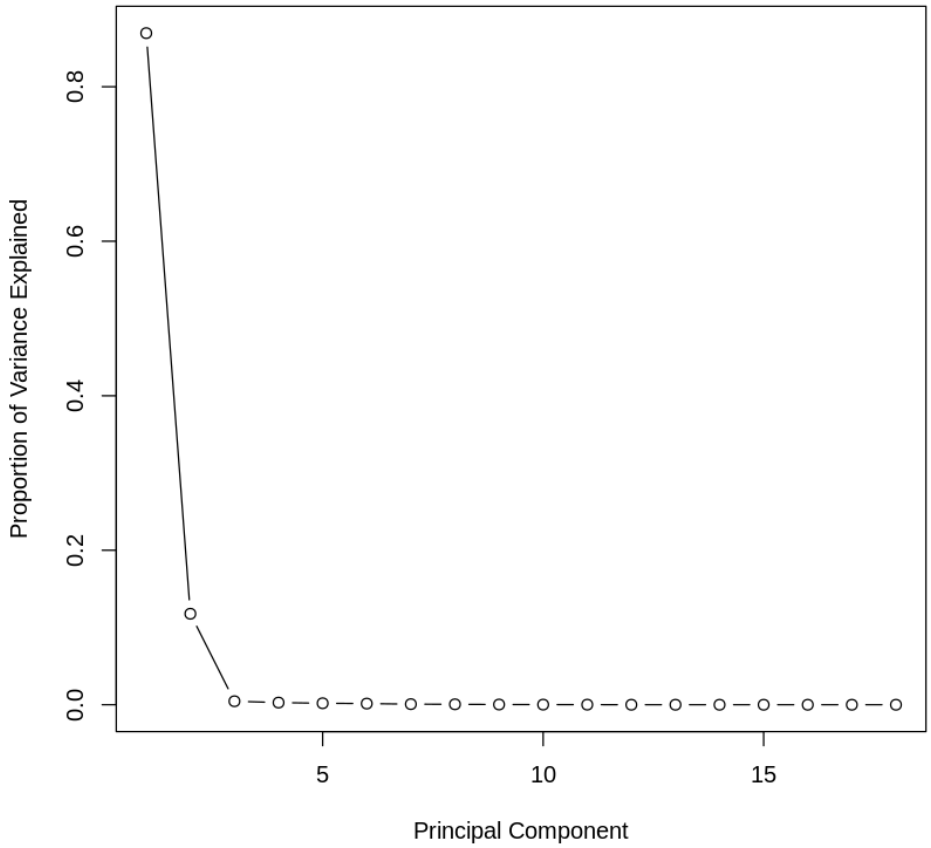
The dataset for RTAs is structured, by choice, on three-year non-overlapping sums of patent counts, from 1990 to 2016, in order to smooth out the volatility in patent activity. Once RTAs are obtained from patent-technology cross tables, we construct 18 different region-technology proximity networks.

We borrow the definitions of the technological networks from two papers (Yan and Luo, 2017; Stellner, 2014). The methods applied include simple correlations, minimum pairwise conditional probabilities, class-to-class cosine similarity, class-to-patent cosine similarity, co-classification, co-occurrence to generate the networks on the four different levels, geographic (regions and countries) and personal (inventors and applicants).

The methodology is inspired by Hausmann et al (2014), however, there are a few differences. Instead of calculating a related product and a regional density to explain future RTA, we only use relatedness in technologies and do not factor in the region component. However, instead of relying only on one measure of technological relatedness (such as RTA correlation for pRTA), we use different methodologies to calculate 18 measures for proximity. In a final step, all the matrices that contain the weighted product and region densities are stacked vertically, so we obtain two column vectors with each of length of a product between IPC class-regions pairs. Being the 18 networks very collinear we apply Principal Component Analysis, and obtain the regressors for the pRTA estimations.

Figure 1 illustrates the portion of variance explained by the Principal Components based on the 18 technology-region networks. The correlation between the networks and the first principal component is on average higher than 0.93. Using a Zero-inflated Beta regression, we obtain the parameters of the model by regressing the principal components in t_1 on t_2 RTA values, and subsequently fitting the model on t_2 matrices in order to obtain the predicted values for technological advantage for t_3, as in Zachmann and Roth (2018).

Figure 1: Principal components analysis



For the Zero-inflated Beta model, we rely on the R package *GAMLSS* and its function *BEZI()*. These networks-based vectors are also used as regressors in a linear regression that serves as a baseline to evaluate the PCA model.

To obtain the value of the potential revealed comparative advantage (pRTA) of every region-technology combination, we fit a zero-inflated beta regression. The beta distribution can only take values in the range between zero and one. By using a zero-inflated beta distribution, zeros can also be modelled (Ospina and Ferrari, 2012). The zero-inflated beta regression takes the following functional form.

$$\begin{aligned} &\text{if y} = 0: && f(y) = \upsilon \\ &\text{if y} = (0,\ 1): && f(y|\mu,\sigma) = (1-\upsilon)\,\frac{\Gamma(\sigma)}{\Gamma(\mu\sigma)\Gamma((1-\mu)\sigma)} y^{\mu\sigma}(1-y)^{((1-\mu)\sigma)-1} \end{aligned} \tag{3}$$

The resulting predicted data will be a measure of potential technological advantage at *t+1*. The zero inflated beta regressions show a mean squared error of 0.05 compared to the baseline, and an average R^2, for all the stacks, not higher than 0.35. These statistics, if compared to a similar previous paper by Roth and Zachmann (2018) that performed a country-level analysis, exhibit a poorer performance of our regional models. We will discuss the implications and the possible ways to improve the models in Section 5.1.

In the following Section, we present examples of the pRTA estimations and apply network analysis techniques to understand collaboration in green technologies.

3.2 The geographical dimension of potential advantage

After estimating revealed and potential technological advantage, the first thing that we observe is that certain low-carbon products show a pattern of strong concentration in few regions, such as Rhône-Alpes in France, Dresden and Stuttgart in Germany, Lombardia in Italy: well-known industrial districts and technological hubs. This phenomenon is even more evident when looking at patent counts itself.

Over time, we observe a general increase of low-carbon technological advantage across Europe, although our measure of RTA seems to be quite volatile, despite the three-year smoothing already mentioned.

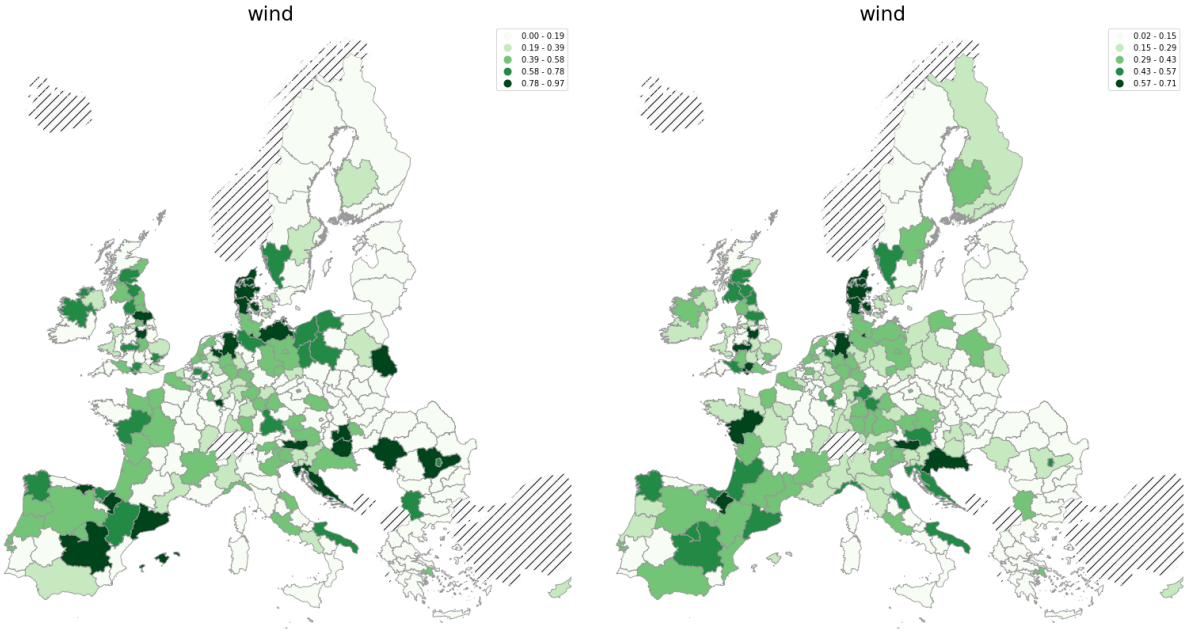
In terms of innovation specialisation, certain technologies, such as nuclear, remain exclusive for a smaller number of regions that are already strong in innovating nuclear technology, as shown in Figure 3. Other technologies, such as wind and hydro power appear to be promising for many regions.

In Figure 2 it is possible to observe how the wind-related technologies have similar geographical distribution looking at RTA in 2012 besides the potential RTA for the successive period, 2015. Many countries have at least one region with some degree of specialisation that results in an advantage for 2012 in Denmark, Germany and Spain. Two regions that exhibit strong potential advantage are Continental Croatia and Pays de la Loire in the north-west of France.

This, most likely, has to do with the technological complexity involved in producing these products. While the production of products for nuclear power plants involves itself a lot of sophisticated technologies, thus the entry barrier for companies is high, other low-carbon technologies allow an easier access for newcomers and thus a wider spread over several countries. Industrial hubs reveal to have an advantage or a potential across many fields as they present a very complex product mix.

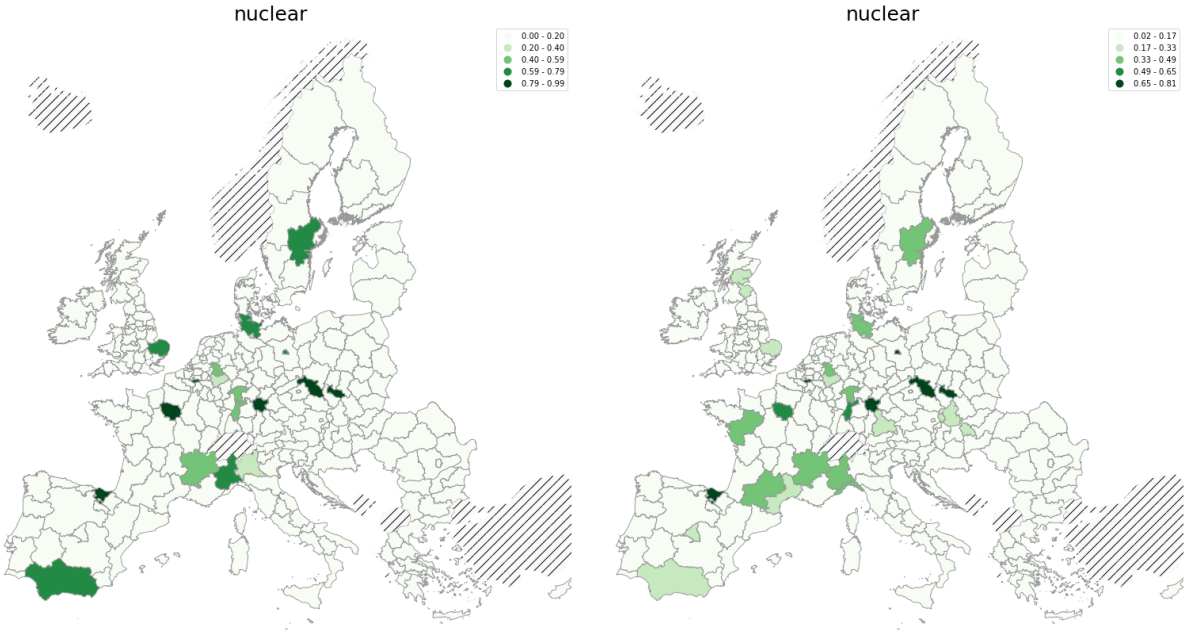
3.2.1 European networks of green innovation

A growing body of literature applies network theory to patent analysis. Visualising the technological space can lead to an understanding of technological proximity and possible developments of a technology. Mariani et al. (2019), focus on citations of patents and use network centrality for technological forecasting. Wu and Yao (2012), create and test on a specific technical field, an artificial intelligence based method for network analysis, combining text-mining techniques.



(a) RTA 2012 (b) pRTA 2015

Figure 2: *Revealed and potential technological advantage in wind technologies*



(a) RTA 2012 (b) pRTA 2015

Figure 3: *Revealed and potential technological advantage in nuclear technologies*

Song et al. (2016) apply overlay patent networks to analyse the design space evolution by looking at co-references of patents, in order to understand the possible directions of the most likely expansion paths.

In this section, we explore with network analysis the relatedness of European regions in low-carbon technologies. We start by looking at the technological space with a focus on our selected fields. The European technological space is plotted in Figure 4. This graph is built by constructing a technology-technology matrix between the IPC classes and the low-carbon technologies. Each node is a 4-digit IPC class or a green technology, and the weight of the nodes is given by the correlation between the RTAs for the 2013 stack.

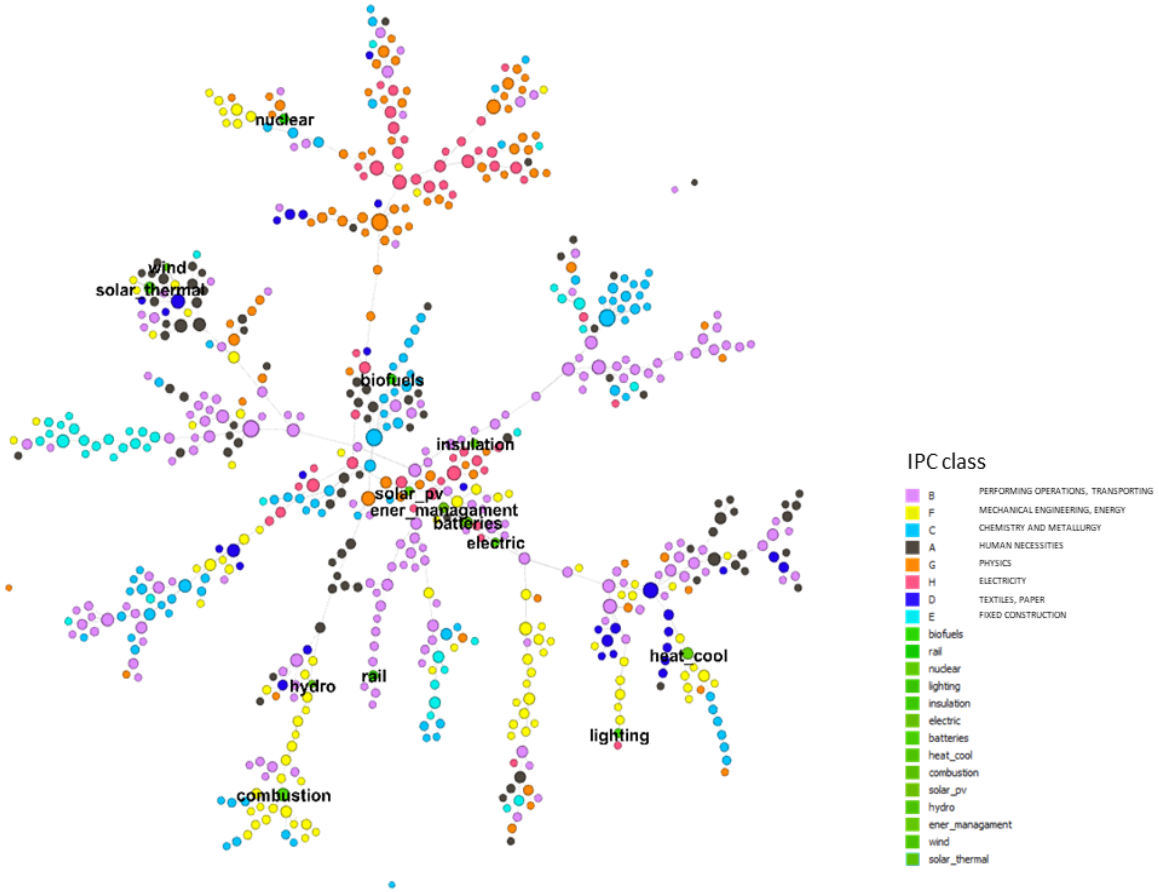


Figure 4: *European product space based on patents registered in REGPAT 2018. The network is built on the RTA correlations between IPC 1-digit technological classes and low-carbon technologies based on the relevant CPC Y-codes.*

Being the network extremely dense, we present here a visualization of the maximum spanning tree. We can observe how solar panels, energy management, batteries and electric cars seem to have good proximity in the network. Nuclear is more distributed and closer to G and F classes (physics and mechanical and energy engineering). Rail is instead quite well connected with the IPC codes that fall under IPC-class B (Performing operations, transportation).

In Figure 5 we recreate the same network with a transposed matrix, in order to understand the proximity of European regions based on the correlations between the technological structures of their economies. The size of the nodes represents their in-degrees. A general, first, observation from this graph is that European regions have product mixes which go beyond their country of belonging.

Looking closer to the division, we notice the wedge between productive regions (bottom left clusters) and less productive regions (mid right branch). The regions that cluster on the left around Veneto (ITH3) and Upper Franconia (DE24) appear to be the highly productive cluster, listing also Rhône-Alpes for France. Here the regions of northern Italy and most of the highly industrialized regions of the Germany branch out. Another interesting cluster is that at the bottom-right: we observe that the dutch region of Noord-Brabant (NL41), a renowned high-tech region, is close to East Anglia (UKH1), the danish capital region of Hovedstaden (DK01) and Île de France (FR10).

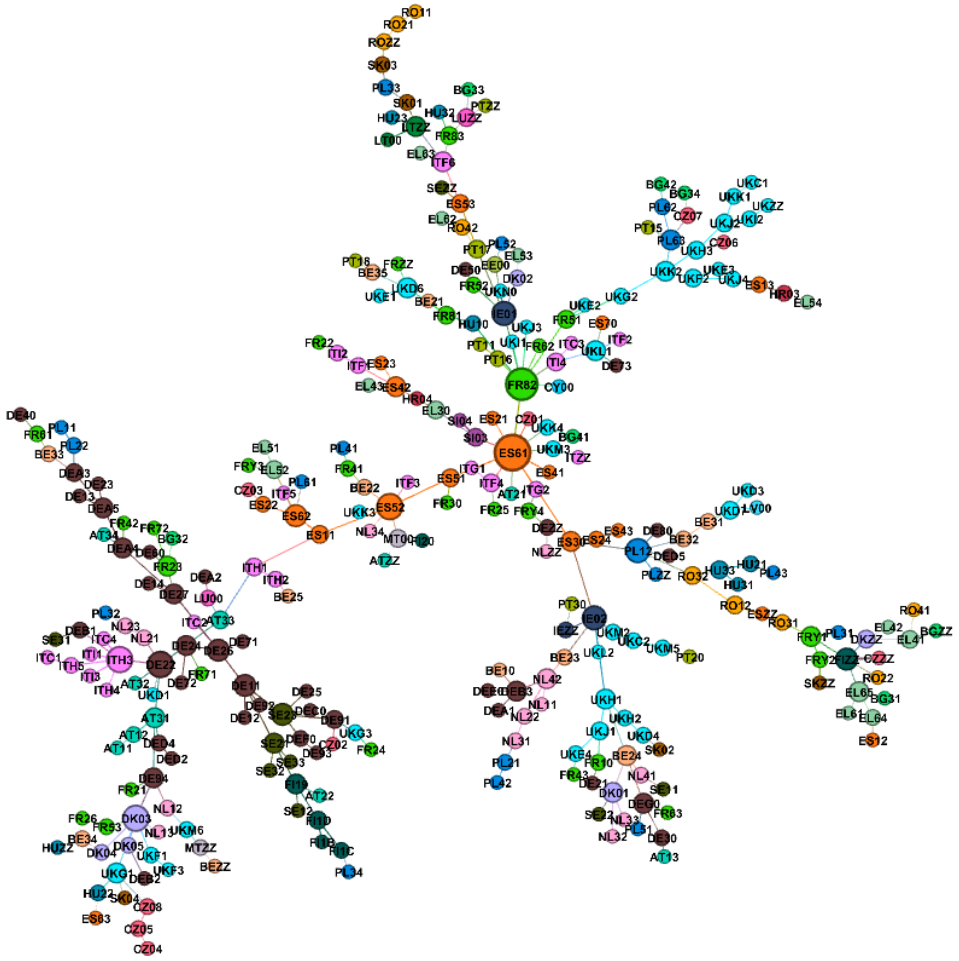


Figure 5: *Proximity of NUTS2 regions based on revealed comparative advantage in 2012.*

This network is based on the relative technological advantage of all the 637 4-digit IPC

codes. With the intention of observing how European regions collaborate in patenting low-carbon technologies, in order to highlight green industrial clusters, we build other networks based on simple co-patenting figures.

In Figure 6, we plot a graph only based on low-carbon technologies, in which the size of the nodes is relative to the number of patent applications, and the weight of the edges represents the number of co-patents between two regions. The network is then clustered based on its modularity, a structural measure that tells how well the graph can be divided in different modules. Specifically the OpenOrd algorithm (Martin et al, 2011) is applied.

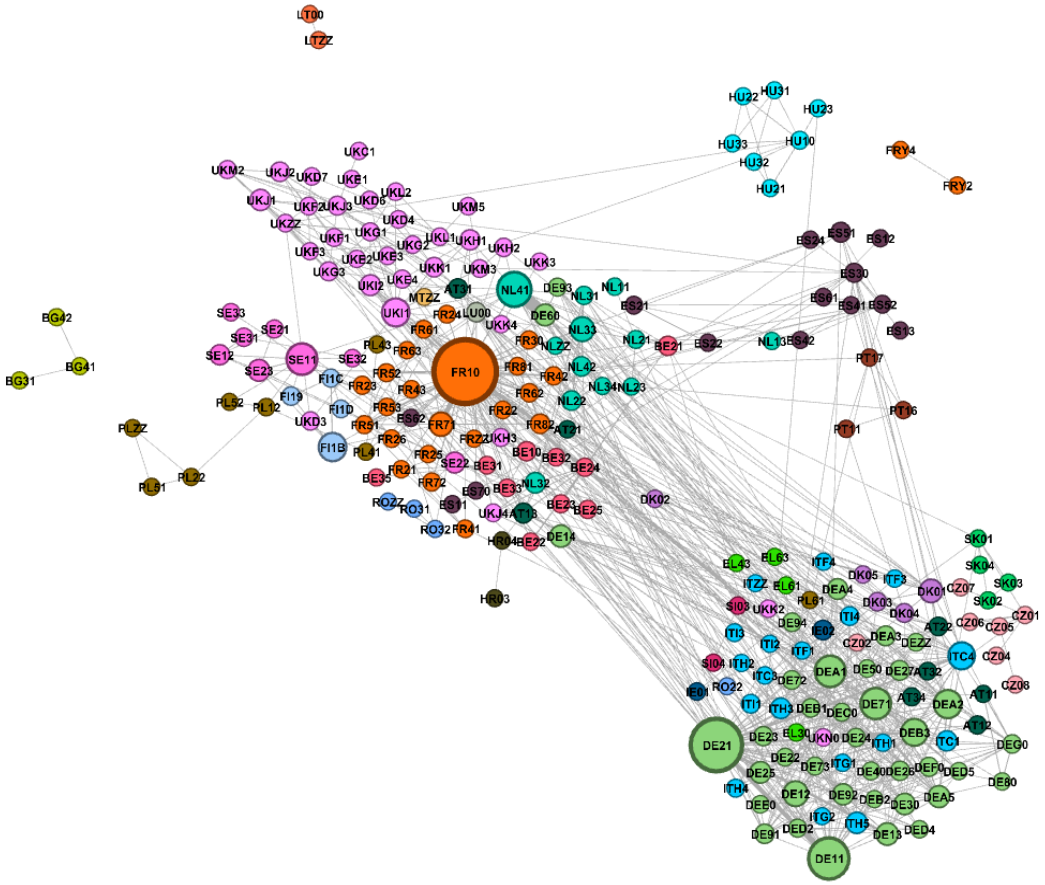
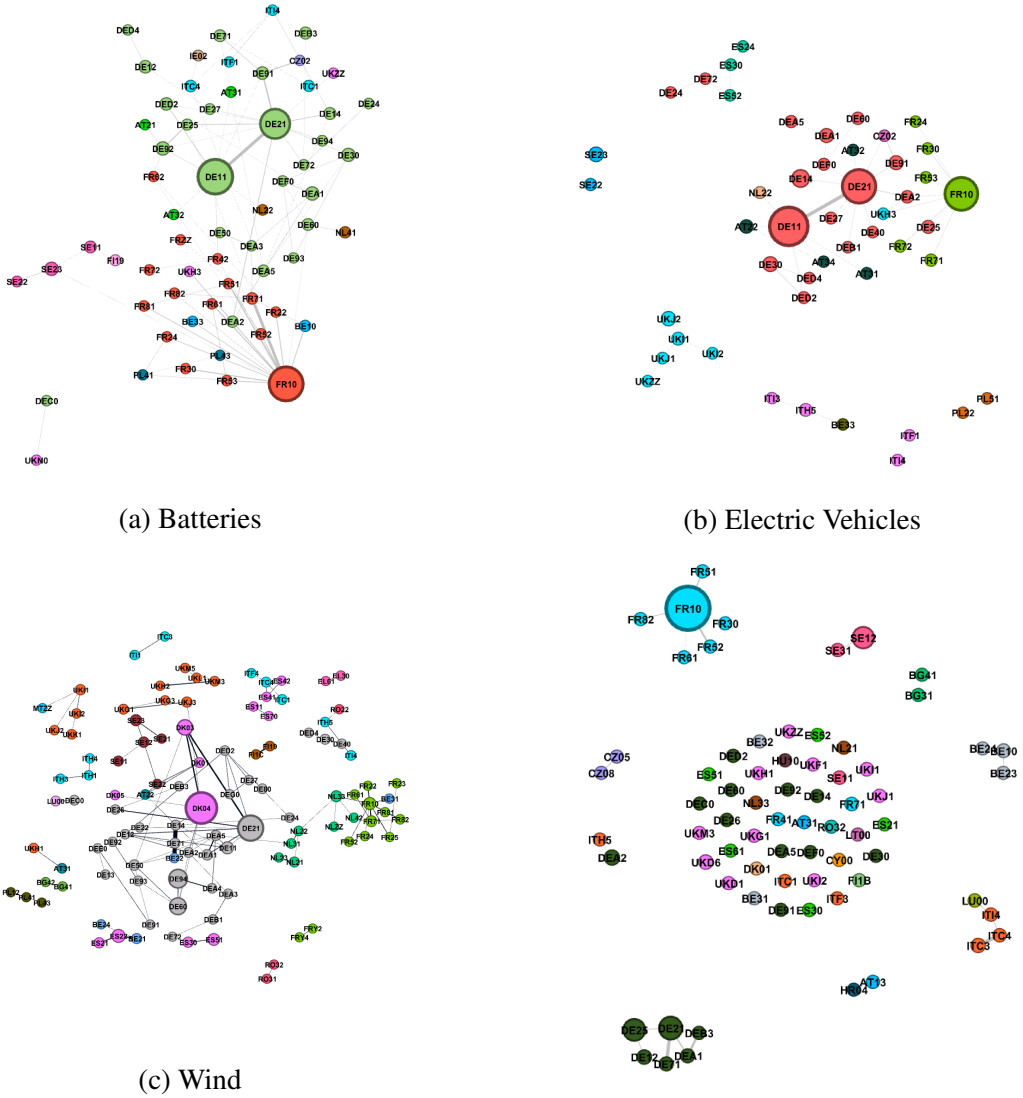


Figure 6: *Co-patenting of European regions in low-carbon technologies.*

Two clusters emerge, although quite tightly connected. One is dominated by Île de France (FR10), region of the capital of France. The other by Germany, with Oberbayern (DE21) and Stuttgart (DE11). The United Kingdom, the Netherlands, Belgium and Sweden, among other, seem to be clustered more tightly with France, whereas on the other side we see Italy, Germany, Slovakia and Austria.

In addition, we compute the same co-patenting networks considering one low-carbon technology at the time. In the four panels of Figure 7 we show the examples of batteries, electric vehicles, wind and nuclear technologies. What are the clusters of innovations? In the case of

batteries and electric vehicles we can see the clusters in France and Germany, whereas for wind Denmark and Germany are the most central. In these three panels, we filter out the nodes that have no co-patenting and a small number of patents, contrary to the case of the fourth panel. Interestingly, in the case of a nuclear technology we can observe how the clusters almost reflect the national boundaries. At the center of the network are represented nodes of low degree without edges.



(a) Batteries

(b) Electric Vehicles

(c) Wind

(d) Nuclear

Figure 7: *Co-patenting for European regions based on Applicant's location for the period 1990-2016.*

In the following sections we move forward with the second part of the study that makes use of the full dataset and the dimensionality reduction algorithm in order to explore the association

between the measures for potential and expressed technological specialisation and the variables of the Eurostat and Commission's databases.

4 Multi-dimensionality reduction: a data-driven selection of variables

Our approach is exploratory, and moves on from the question: what can affect the presence of an advantage in a low-carbon technology? We aim at understanding the association between our measure for revealed advantage in a technology for region *r* at time *t*, the potential in that same technology at *t-1* and the large number of regressors on the right-hand side. We begin with an agnostic approach about what could be associated with RTA: we first build a wide dataset at the NUTS2-year dimension with all the variables present in the Eurostat database and in the Urban Data Platform. We subsequently apply an Elastic Net regularization to the wide dataset in order to isolate the relevant variables.

In the following paragraph we explain the imputation methodology that we apply to the panel dataset at the NUTS2 level, in order to fill in the gaps across regions and time. Subsequently, in Section 4.2, we present our Elastic net regularization for a data-driven variables selection.

4.1 Imputation methodology

Data collection of regional statistics at the NUTS2 level, though improving, is highly inconsistent. As a result, many of the data sheets used to generate our dataset are incomplete on the bases of time and location. Proper utilization of NUTS2 regional statistics is therefore a challenge given such complications. Limitations will be discussed further in section 5.1.

As a consequence of the indiscriminate scraping techniques used to generate it, the dataset is incomplete and highly multicollinear, containing several repeated and aggregated indicators. Regularized regression techniques such as LASSO are often used to reduce dimensionality through variable selection; however, these techniques fail in the presence of missing data. Therefore, we choose to impute missing values before using any data-driven variable selection technique.

Multiple Imputation by Chained Equations (MICE) is a data imputation algorithm which can attempt to capture the uncertainty associated with missing data values by "randomly drawing multiple imputations from a distribution of imputations and also by introducing additional error variance to each imputation". MICE makes the assumption that data is missing at random (MAR), meaning that the presence of an underlying relationship between the propensity of a region to be missing data and the value of the missing data causes any results taken from imputation to be problematic (i.e. due to estimator bias). Additionally, MICE fails when imputing on non-invertible matrices; therefore, it is important to reduce collinearity as much as possible prior to imputation.

Given the poor performance of advanced imputation methods on high-dimensional low-rank matrices, we first hand-select indicators using domain knowledge to reduce the total number of indicators from 476 to 245. Additionally, we remove regions which are defined as extra NUTS2

regions (encoded with a ZZ) which are not useful for the purposes of this study. We also remove overseas territories (e.g. PT20 or FRY1) or regions with NUTS2 codes that have been replaced (e.g. UKI1 or IE01) but still persist in Eurostat or Urban Data Platform databases. Further, these methods often utilize linear regressions to estimate missing values and lose a great deal of stability above certain thresholds of missingness. A 2017 thesis out of East Tennessee State University showed that MICE imputation using non-Bayesian linear regression exhibits stability for datasets missing up to 50% of observations. For our purposes, we place this threshold at 30% to get a dataset of 110 indicators for 258 NUTS2 regions.

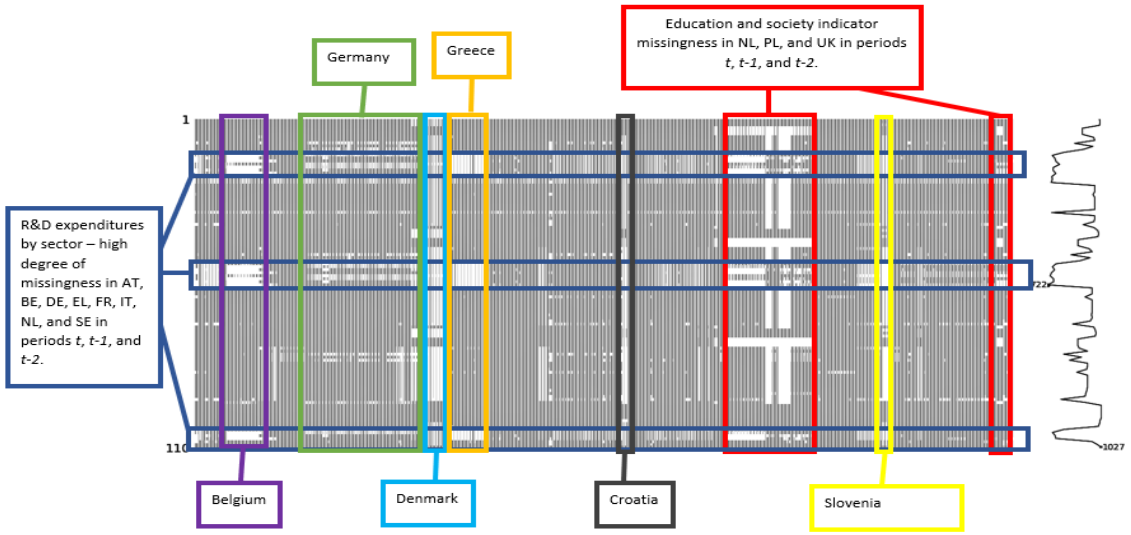
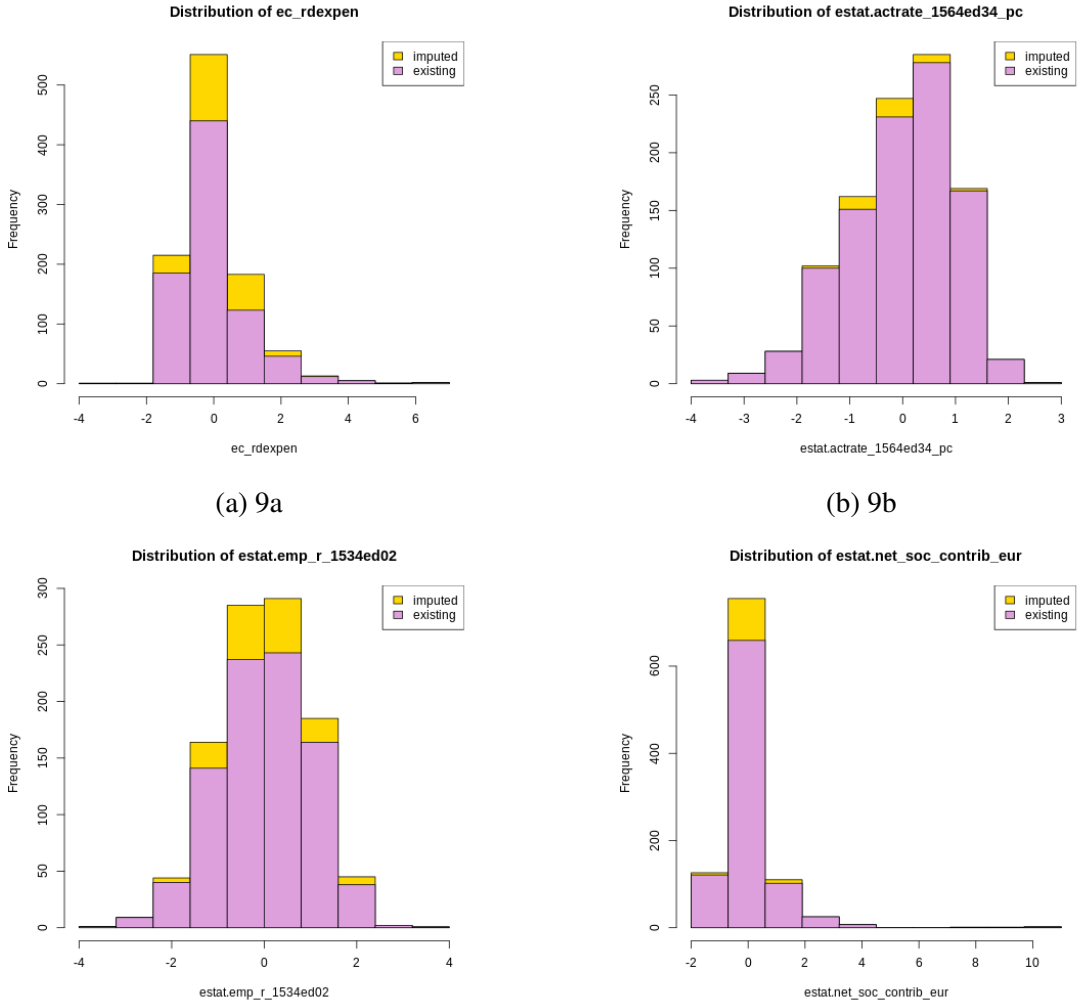


Figure 8: *Visualization of dataset with missing data is shown in white. Horizontal patterns of missingness indicate data missing across regions while vertical patterns indicate missingness within a region.*

Using the Python missing data visualization package missingno, we examine our dataset for patterns which might invalidate the MAR assumption. Investigation shows that several areas of missingness can be attributed to lack of availability (i.e. Denmark in 2005). Given these patterns and the completeness of the remaining data, we proceed under MAR assumptions and impute with MICE.

In line with MICE common practice, we first allow the algorithm to identify constant or collinear variables which could present problems during the imputation step. Three covariates are identified as being collinear and are removed. We then impute using MICE using non-Bayesian linear regression. See Figure 9 for histograms showing the results of imputation on selected covariates.



(a) 9a

(b) 9b

(c) 9c

(d) 9d

Figure 9: *Distributions of selected variables before and after imputation*

4.2 Dimensionality reduction

After using MICE imputation to fill in the gaps without biasing the distributions of the variables, we apply Elastic net regularization to the dataset. As mentioned, this approach is experimental and does not seek to make causal inference but to look at mere associations.

Our dataset is based on 4 time stacks based on RTAs and pRTAs estimated for non-overlapping periods of three years (2006, 2009, 2012, 2015). Considering both Eurostat and Ubran data platorm's, it contains 1032 observations and 110 variables, after the first selection done on the Eurostat database to reduce multicollinearity, and the threshold filtering, as explained in the previous paragraph.

We define RTA, an observed advantage in region r at time t, as a function of all the other variables and pRTA, as in equation (4).

$$RTA_{\mathrm{r,t,i}} = \beta_0 + \sum_{k=1}^{K} \omega_{\mathrm{k}} x_{\mathrm{r,\,t,\,i,\,k}} + \sum_{k=1}^{K} \gamma_{\mathrm{k}} x_{\mathrm{r,\,t\text{-}1,\,i,\,k}} + \sum_{k=1}^{K} \theta_{\mathrm{k}} x_{\mathrm{r,\,t\text{-}2,\,i,\,k}} + \theta pRTA_{\mathrm{r,\,t\text{-}1,\,i}} \tag{4}$$

where:

$r = 258$ *(NUTS2 regions)*

i = Low-carbon technology

$t = 3$ years non overlapping time stack

$K = 110$

The regularization procedure which γ coefficients are significantly associated with RTA. The two most common applications of regularization regression are LASSO (Least Absolute Shrinkage and Selection Operator) and Ridge regressions. The intuition behind these techniques is to apply a penalty score on the magnitude of the coefficients of an OLS regression.

In the data science literature, they are typically known as the L1 and L2 types of regularization. L1 (Ridge regression) and L2 (Lasso regression), are different in the way the penalty is applied to the cost function. LASSO regularization, with the respect to Ridge, can shrink coefficients to zero, performing a real variable selection, although in presence of a large number of multicollinear variables, the selection of the feature is random. On the other hand, Ridge can handle multicollinearity non at random, but performs poorly with a high number of dimensions.

Elastic net regularization combines L1 and L2 penalty scores, as in equation 5, overcoming the respective limitations. This method is more flexible to our purpose, being able to better handle the high multi-dimensionality and collinearity of our dataset. In addition, the flexibility stems from the ability to dynamically balance the model parameters depending on the best score.

$$\hat{\beta} = argmin \parallel y - X\beta \parallel^2 + \lambda_2 \parallel \beta \parallel^2 + \lambda_1 \parallel \beta \parallel^2 \tag{5}$$

In order to estimate the coefficients for the right-hand side, we first subset the panel to four cross-sectional datasets based on the 2006, 2009, 2012 and 2015 stacks. The regularization method will highlight the cross-regional effects. The dataset is divided in training and testing samples with a 10 fold cross-validation strategy, and standardized on both the right and left hand sides.

We make use of the `ElasticNetCV` module of the python library `sklearn` to perform the regularization. The module allows an automatic choice of the L1-L2 ratio, influencing the weight given to each penalty factor, by feeding in an array of possible values. The L1-L2 ratio ranges from 0 to 1, and tells how skewed the model should be towards LASSO or Ridge. The array chosen is skewed towards the LASSO-type regression, including more values above 0.5 than below. In the same fashion, the alpha level (overall magnitude of the penalty score) is also automatically selected.

Figure 10 shows the results of the regularization for electric vehicles, for the 2012 and 2015. On the horizontal axis we plot the 110 coefficients estimated, and on the vertical one their magnitude. The last coefficient on the right-hand side is pRTA.

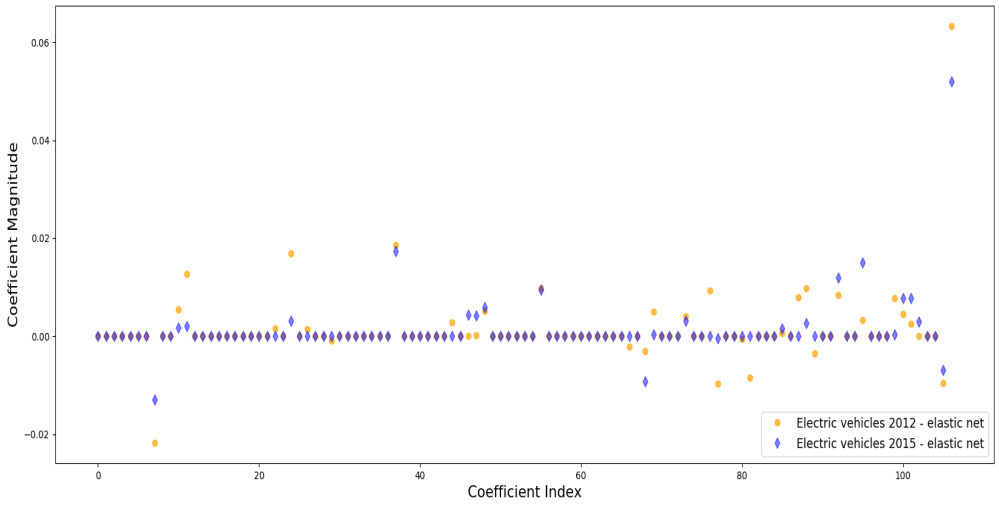


Figure 10: *Results of the Elastic net regularization where for electric vehicles in the 2012 and 2015 stacks, and the other variables and interaction factors are estimated at t-1*

After applying the regularization to each low-carbon technology all the resulting non-zero coefficients, despite their direction, are combined for all technologies. Tables 3 and 4 of the appendix present the results respectively for the 2012 and 2015 cross sections. In the columns, we can count how many times, that variable survives the regularization in *t*, as a first and second lag, and in total. pRTAs at *t-1* are found to be always the highest ranking coefficients, and are omitted in the tables.

In addition, we make an attempt to include the interaction factors between pRTA and the right hand side variables, in order to understand which variables are the most relevant combined with pRTA. The in terms of interaction terms at *t-1* are not too dissimilar from what we estimate in the previous specification.

However, the Elastic net model does not allow us to properly assess the interactions. This results in R^2 scores lower than 0.1. This issue could be overcome, possibly in a further second stage with less collinear variables, in a group LASSO model that can hierarchically explore interactions, as explained by Lim and Hastie (2015).

In order to observe the time-varying effects of these variables, the panel dataset is demeaned, following a two-way fixed effects approach as in Imai and Kim (2018). However, differently from the two-way methodology, the panel is only demeaned for region-fixed effects. For each region, in each time period, the average of that region over the years is subtracted. The subsequent regularization is therefore performed under the assumption of fixed effects estimation. The results of the fixed effects regularization are presented in Table 4 of the appendix. The R^2 values yielding from this model are all lower than 0.5%, suggesting that the variation in RTA comes more from cross-sectional differences rather than time-varying effects.

As for the cross-sectional models, in order to produce genuine estimators for the interaction

terms in fixed effects, as explained by Giesselmann and Schmidt-Catran (2018), the simple demeaning of the interaction terms is not sufficient. This issue will be explored in future research.

5 Results

Tables 2, and 3 of the Appendix summarize the results of the regularization via Elastic net, by showing the number of times a coefficient is found to be non-zero, at time *t*, *t-1*, *t-2* and in total. All our results are not causally identified, and only represent an association. The variable that survives the greatest number of times is the activity rate (labour participation) of people that have an ISCED level of education[2], higher than the upper secondary level.

A longer duration of employment is also associated positively in many of the specifications with the advantage in low-carbon technologies. Logically, the number of engineers and scientists is also significant if included at any time period. In general, higher educational attainments of the population are associated with more innovation in green technologies. We also find some evidence on a positive association with tertiary educational attainment and (lower) unemployment rates of women.

In addition, different measures for R&D expenditure are resulting. Both private and public sector spending are selected, but seem to appear more often at time *t* rather than as a lag. The measure that scores higher in our framework, for the fixed effects model, is the percentage of GDP spent by the government on R&D. Intramural expenditure, defined by the OECD as the: *"amount of money spent on R&D that is performed within a reporting unit"* (Frascati Manual, 2015). In a causal model, future research shall try to establish the precise relationship between R&D expenditure in the private sector, in higher education institutions, and business enterprises.

Tables 5 and 6, instead, present the count of coefficients indicated by the regularization procedure, using the *glinternet* R-package and the one-way demeaned panel approach discussed in the previous section. Table 4 summarizes the main effects, while table 5 summarizes the interaction factors that we force with pRTA at *t-1*. Overall, our fixed-effects regularization approach does not yield particularly different results compared to the cross-sectional specifications.

5.1 Limitations and further research

The first stage of these paper used zero inflated beta regressions to predict pRTA values. Looking at the R^2 statistics, not higher than 0.35, suggests that the modelling could be improved. pRTA values correlate at 0.4, on average, with RTA. The definition of the stacks (three years versus a larger period), appears to be less problematic than the modelling, as does not seem to have too much effect on the volatility of the RTAs, although making the RTA measures probably more precise. In further research, a model with less volatile RTAs could yield more consistent predictions of pRTA.

One of the most significant limitations of this study is data availability at the NUTS2 level. Prior to dimensionality reduction, over 70% of indicators are missing more than 10% of observations. Of these, around 60% are time-lagged and 40% are non-lagged. As a consequence of

[2]International Standard Classification of Education, classified by the UNESCO

the missingness, the 30% threshold used for keeping covariates is higher than preferred. Though some in the literature suggest decent performance of multiple imputation on high dimensional matrices with significant amounts of missing data, stability and quality of imputed values are obviously improved given a more complete starting dataset.

Moreover, because of the need to reduce dimensionality and missingness when using multiple imputation and regularized regression, we may be eliminating covariates or interactions among covariates and pRTA/RTA which have significant predictive power. For some covariates, though, the missingness is so pervasive that the ability to have meaningful predictive power is precluded. Better and more consistent data collection at the NUTS-2 level will help solve these dilemmas.

Related to the data availability problem, MICE imputation assumes that the missingness mechanism of the underlying data is MAR. This would imply that, conditional on our observed values, the values of missing data have no relation with the missing data. After reviewing the patterns of missing data by indicator, region, and time, it could be argued that missingness is heterogenous in its mechanism, with some being MAR and others being missing not at random (MNAR). Given a high proportion of missing data being consolidated across similar indicators and time periods, we felt comfortable making a blanket MAR assumption; however, a much closer evaluation of missingness should be performed to confirm this assumption.

Furthermore, the regularization methodology should move forward in order to take into account time-varying effects and the interaction factors between potential and right-hand side variables. As discussed, this is in our view one of the most relevant potential avenues.

Because of the flexibility of this pipeline, the cost of adding more right-hand side variables in the model is very small, and adding different NUTS2 level datasets and types of variables could lead to different results. In particular, we believe that the use of datasets with diverse scope and extensions could be particularly relevant, notably focusing on the infrastructural dimension, market structure, competitiveness and institutions.

Within the current framework, having recognized an association between our right hand-side variables, further research should try to establish with a causal approach the relationship between potential advantage in a region and its education, labour markets and R&D public and private spending.

Although RTA represents the regional specialisation in a technology, the same approach could make use of different models for patent counts in order to conduct a different but quite related exercise. In this sense and given that our results point towards what is generally recognized as good policy for innovation, further research should also try to model a better counterfactual on what distinguishes green innovation from general technological innovation, leading to more precise policy recommendations.

6 Conclusions and possible implication for policy

In this paper we explored, at European regional level, the innovation in low-carbon technologies. The motivation for our study stems from the necessity that European states will face in the years to come to foster innovation in low-carbon technologies. This necessity does not only find its roots in the need to contain global warming below 2°C. The decarbonisation process will also

push for a strong change in our production systems, and consequently on the labour markets and industrial sectors.

Hence, we consider this study an exploratory attempt to contribute to the debate around a horizontal green industrial policy for regions, with a data-driven approach.

Our results are, generally, in line with the literature on horizontal industrial policy. We find an association between a regional advantage in low-carbon products with higher activity rate for people which have a level of education above upper secondary, and a higher presence of science and technology knowledge intensive workers. In addition, in terms of labour markets, we have evidence of this positive association where the total duration of employment is longer.

In terms of R&D spending the correlations seem to be important for both the public general spending in R&D, as well as intramural expenditure both in the private sector and in higher education institutions. Aside from overall and sectorial government spending, our study points towards the relevance of policy measures that aim at increasing the incentives for private firms to invest in R&D, in order for regions to increase their relative advantage.

Furthermore, using network analysis techniques, we find that European regions have a good degree of in-country technological diversification, and that European policies could address smart specialisation in green technologies based on a clustering that is country-independent, and is linked to the local economic characteristics. Even though, as expected, a small number of leading regions (notably in France and Germany) are pushing the frontier of green patenting, we find a large number of others that have potential to develop a technological specialization in the low-carbon technologies in the future.

Although other factors play a role in determining competitive advantages, technological specialisation can promote competitive industries, thereby shaping long-run growth dynamics. Policy can leverage strength in similar technologies by shaping innovation paths, strengthening learning capabilities, targeting sector-specific innovation regimes, and coordinating sectoral, national and regional policies.

Although our results are in line with those of the literature, our empirical approach has several important limitations, and should be further refined in order to bring the data-driven selection of variables to a more rigorous third stage approach aiming to establish robust causal linkages, from which we could infer more informed policy recommendations.

References

[1] Balland, Pierre-Alexandre, Boschma, Ron, Crespo and Rigby L. David (2019), Smart specialization policy in the European Union: relatedness, knowledge complexity and regional diversification, Regional Studies, 53:9,1252-1268

[2] Dechezleprêtre et al. (2015). Climate Change Mitigation Technologies in Europe – Evidence from Patent and Economic Data. UNEP and EPO

[3] Fiorini, Al, et al., (2017) Monitoring R&I in Low-Carbon Energy Technologies.

[4] Frascati Manual (2015). Guidelines for collecting and reporting data on Research and Experimental Development, OECD.

[5] Giesselmann, M. and Schmidt-Catran, A. (2018). Interactions in Fixed Effects Regression Models.

[6] Hausmann, Ricardo, et al. "Implied comparative advantage." (2014).

[7] Hidalgo, C. A., Klinger, B., Barabási, A. L., Hausmann, R. (2007). The product space conditions the development of nations. Science, 317(5837), 482-487.

[8] Imai, K. and Kim, I. S. (2018). On the use of two-way fixed effects regression models for causal inference with panel data.

[9] Lim, M. and Hastie, T., 2015. Learning interactions via hierarchical group-lasso regularization. Journal of Computational and Graphical Statistics, 24(3), pp.627-654.

[10] Lodder, Paul. "To impute or not impute: That's the question". Advising on research methods: Selected topics, Johannes van Kessel Publishing, Huizen (2013).

[11] Mariani, M. S., Medo, M., Lafond, F. (2019). Early identification of important patents: Design and validation of citation network metrics. Technological forecasting and social change, 146, 644-654.

[12] Martin, Shawn, et al. OpenOrd: an open-source toolbox for large graph layout. Visualization and Data Analysis 2011. Vol. 7868. International Society for Optics and Photonics, 2011.

[13] Oketch, Tobias. Performance of Imputation Algorithms on Artificially Produced Missing at Random Data. Electronic Theses and Dissertations (2017).

[14] Ospina, R. and Ferrari, S.L., 2012. A general class of zero-or-one inflated beta regression models. Computational Statistics Data Analysis, 56(6), pp.1609-1623.

[15] Stellner, Florian. Technological distance measures: theoretical foundation and empirics. DRUID Society Conference. 2014.

[16] Song, B., Triulzi, G., Alstott, J., Yan, B., Luo, J. (2016). Overlay patent network to analyze the design space of a technology domain: the case of hybrid electrical vehicles. In DS 84: Proceedings of the DESIGN 2016 14th International Design Conference (pp. 1145-1154).

[17] Yan, Bowen, and Jianxi Luo. Measuring technological distance for patent mapping. Journal of the Association for Information Science and Technology 68.2 (2017): 423-437.

[18] Wu, C. C., Yao, C. B. (2012). Constructing an intelligent patent network analysis method. Data Science Journal, 011-003.

[19] Zachmann, Georg. An approach to identify the sources of low-carbon growth for Europe. No. 2016/16. Bruegel Policy Contribution, 2016.

[20] Zachmann, Georg, and Alexander Roth. Learning for Decarbonisation: start early, concentrate on promising technologies, exploit regional strength and work with your national system. Bruegel Policy Brief October 2018." (2018).

[21] Zachmann, Georg, and Robert Kalcik. (2018) "Export and patent specialization in low-carbon technologies." Global Innovation index (2018): 107.

[22] Zou, H., Hastie, T. (2005). Regularization and variable selection via the elastic net. Journal of the royal statistical society: series B (statistical methodology), 67(2), 301-320.

A Appendix

A.1 List of CPC-Y codes for low-carbon technologies definitions

Technology	CPC-Y codes (patents)
Solar PV	Y02E1050, Y02E1052, Y02E1054, Y02E10541, Y02E10542, Y02E10543, Y02E10544, Y02E10545, Y02E10546, Y02E10547, Y02E10548, Y02E10549, Y02E1056, Y02E10563, Y02E10566, Y02E1058
Solar Thermal	Y02E1040, Y02E1041, Y02E1042, Y02E1043, Y02E1044, Y02E1045, Y02E1046, Y02E10465, Y02E1047
Wind	Y02E1070, Y02E1072, Y02E10721, Y02E10722, Y02E10723, Y02E10725, Y02E10726, Y02E10727, Y02E10728, Y02E1074, Y02E1076, Y02E10763, Y02E10766
Hydro	Y02E1020, Y02E1022, Y02E10223, Y02E10226, Y02E1028
Energy management	Y02B7030, Y02B7032, Y02B703208, Y02B703216, Y02B703225, Y02B703233, Y02B703241, Y02B70325, Y02B703258, Y02B703266, Y02B703275, Y02B703283, Y02B703291, Y02B7034, Y02B70343, Y02B70346
Lighting	Y02B2010, Y02B2012, Y02B20125, Y02B2014, Y02B20142, Y02B20144, Y02B20146, Y02B20148, Y02B2016, Y02B2018, Y02B20181, Y02B20183, Y02B20185, Y02B20186, Y02B20188, Y02B2019, Y02B2020, Y02B20202, Y02B20204, Y02B20206, Y02B20208, Y02B2022, Y02B2030, Y02B2032, Y02B20325, Y02B2034, Y02B20341, Y02B20342, Y02B20343, Y02B20345, Y02B20346, Y02B20347, Y02B20348, Y02B2036, Y02B2038, Y02B20383, Y02B20386, Y02B2040, Y02B2042, Y02B2044, Y02B20445, Y02B2046, Y02B2048, Y02B2070, Y02B2072
Heating and cooling	Y02B3008, Y02B3010, Y02B30102, Y02B30104, Y02B30106, Y02B30108, Y02B3012, Y02B30123, Y02B30126, Y02B3014, Y02B3016, Y02B3018, Y02B3020, Y02B3022, Y02B3024, Y02B3026, Y02B3028, Y02B3050, Y02B3052, Y02B3054, Y02B30542, Y02B30545, Y02B30547, Y02B3056, Y02B30563, Y02B30566, Y02B3060, Y02B3062, Y02B30625, Y02B3064, Y02B3066, Y02B3070, Y02B3072, Y02B3074, Y02B30741, Y02B30743, Y02B30745, Y02B30746, Y02B30748, Y02B3076, Y02B30762, Y02B30765, Y02B30767, Y02B3078, Y02B3080, Y02B3090, Y02B3092, Y02B3094
Combustion	Y02B8010, Y02B8012, Y02B8014, Y02B8020, Y02B8022, Y02B8024, Y02B8026, Y02B8028, Y02B8030, Y02B8032, Y02B8034, Y02B8040, Y02B8050
Residential insulation	Y02E2010, Y02E2012, Y02E2014, Y02E2016, Y02E2018, Y02E2030, Y02E2032, Y02E20322, Y02E20324, Y02E20326, Y02E20328, Y02E2034, Y02E20342, Y02E20346, Y02E20348, Y02E2036, Y02E20363, Y02E20366, Y02E20185, Y02E20344
Biofuels	Y02E5010, Y02E5011, Y02E5012, Y02E5013, Y02E5014, Y02E5015, Y02E5016, Y02E5017, Y02E5018, Y02E5030, Y02E5032, Y02E5034, Y02E50343, Y02E50346

Batteries	Y02E6012, Y02E60122, Y02E60124, Y02E60126, Y02E60128, Y02T1070, Y02T107005, Y02T107011, Y02T107016, Y02T107022, Y02T107027, Y02T107033, Y02T107038, Y02T107044, Y02T10705, Y02T107055, Y02T107061, Y02T107066, Y02T107072, Y02T107077, Y02T107083, Y02T107088, Y02T107094, Y02T1072, Y02T107208, Y02T107216, Y02T107225, Y02T107233, Y02T107241, Y02T10725, Y02T107258, Y02T107266, Y02T107275, Y02T107283, Y02T107291
Electric cars	Y02T1064, Y02T10641, Y02T10642, Y02T10643, Y02T10644, Y02T10645, Y02T10646, Y02T10647, Y02T10648, Y02T10649, Y02T1062, Y02T106204, Y02T106208, Y02T106213, Y02T106217, Y02T106221, Y02T106226, Y02T10623, Y02T106234, Y02T106239, Y02T106243, Y02T106247, Y02T106252, Y02T106256, Y02T10626, Y02T106265, Y02T106269, Y02T106273, Y02T106278, Y02T106282, Y02T106286, Y02T106291, Y02T106295
Rail transport	Y02T3000, Y02T3010, Y02T3012, Y02T3014, Y02T3016, Y02T3018, Y02T3030, Y02T3032, Y02T3034, Y02T3036, Y02T3038, Y02T3040, Y02T3042
Nuclear	Y02E3030, Y02E3031, Y02E3032, Y02E3033, Y02E3034, Y02E3035, Y02E3037, Y02E3038, Y02E3039, Y02E3040

Table 2: Cross-sectional estimation 2012

	t	l1	l2	tot
Activity rates ISCED ≥ 3	15	16	21	52
Total duration	13	21	14	48
Scientists and engineers	4	7	7	18
HH Paid current taxes on income wealth etc. mil EUR	5	6	5	16
HH Social benefits other than social transfers in kind recived mil EUR	5	8	3	16
Persons employed in science and technology	8	2	4	14
Long term uneployment (12 months or longer) in thousands	6	3	5	14
Unemployment rate by age	4	5	5	14
Average number of usual weekly hours in main job by age in hours	6	1	5	12
Participation rate in education and training (last 4 weeks) total age 25-64	0	6	5	11
Gross domestic expenditure on R&D million EUR government	5	6	0	11
Self-employed persons	2	4	5	11
HH Net social contributions mil EUR	3	4	4	11
Gross domestic expenditure on R&D million EUR bussines enterprise sector	6	4	0	10
Age dependency ratio (0-19 and over 60 to pop. aged 20-59)	3	3	3	9
Total R&D personnel higher education sector full time equivalent (FTE)	5	4	0	9
Population aged 25-64 education levels 5-8 (ISCED 2011) total %	4	3	2	9
Persons with tertiary education (ISCED) and/or employed in science and technology % of active population	3	2	4	9
Adult Participation in Learning	9	0	0	9
Intramural R&D expenditure - Government	8	0	0	8
Intramural R&D expenditure - Higher Education	8	0	0	8
Proportion of population aged 20-39	5	0	2	7
Early Leavers from Education and Training	7	0	0	7
Compensation of Employees - Industry (1980 - 2015)	7	0	0	7
Hours Worked - Industry (1980 - 2015)	6	0	0	6
Total R&D personnelgovernment sector full time equivalent (FTE)	2	4	0	6
Professional scientific and technical activities	5	0	0	5
Economic activity rates age over 15 %	1	3	1	5
Intramural R&D expenditure - Enterprise	5	0	0	5

Table 3: Cross-sectional estimation 2015

	t	l1	l2	tot
Activity rates ISCED ≥ 3	12	14	13	39
Total duration	11	12	10	33
Scientists and engineers	6	7	5	18
Average number of usual weekly hours in main job by age in hours	7	4	5	16
Persons employed in science and technology	8	4	4	16
HH Paid current taxes on income wealth etc. mil EUR	4	6	4	14
Long term uneployment (12 months or longer) in thousands	7	2	4	13
Self-employed persons	3	2	6	11
Proportion of population aged 20-39	4	2	5	11
Participation rate in education and training (last 4 weeks) total age 25-64	0	6	5	11
HH Social benefits other than social transfers in kind recived mil EUR	4	7	0	11
Gross domestic expenditure on R&D million EUR government	8	3	0	11
Persons with tertiary education (ISCED) and/or employed in science and technology % of active population	2	4	3	9
Unemployment rate by age	2	5	2	9
Intramural R&D expenditure - Higher Education	8	0	0	8
Age dependency ratio (0-19 and over 60 to pop. aged 20-59)	1	3	4	8
Gross domestic expenditure on R&D million EUR bussines enterprise sector	3	5	0	8
Compensation of Employees - Industry (1980 - 2015)	8	0	0	8
Gross domestic expenditure on R&D million EUR all sectors	3	5	0	8
Persons with tertiary education (ISCED) and employed in science and technology	4	2	1	7
Total R&D personnel higher education sector full time equivalent (FTE)	4	3	0	7
HH Net social contributions mil EUR	1	3	3	7
Hours Worked - Industry (1980 - 2015)	6	0	0	6
Total R&D personnelgovernment sector full time equivalent (FTE)	2	4	0	6
Intramural R&D expenditure - Government	6	0	0	6
Early Leavers from Education and Training	6	0	0	6
Economic activity rates age over 15 %	0	3	2	5
Population aged 25-64 education levels 5-8 (ISCED 2011) total %	2	2	1	5
GFCF - Industry 1980-2015	5	0	0	5
Intramural R&D expenditure - Enterprise	5	0	0	5

Table 4: Fixed effects estimator - Main effects

	t	lag1	lag2	all
Total duration	12	12	17	41
Activity rates ISCED>3	11	12	13	36
Age dependency ratio (0-19 and over 60 to pop. aged 20-59)	4	5	5	14
Scientists and engineers	6	4	4	14
Long term unemployment (12 months or longer) in thousands	6	2	6	14
Students (ISCED 5-6) at regional level - as % of total country level students (ISCED 5-6)	0	6	7	13
Persons employed in science and technology	8	3	2	13
Gross domestic expenditure on R&D million EUR bussines enterprise sector	6	4	3	13
HH Paid current taxes on income wealth etc. mil EUR	2	5	4	11
SBS Manufacturing Share of employment in manufacturing total - percentage	11	0	0	11
Motorways networks in KM	6	2	3	11
Unemployment rate by age	3	4	4	11
Average number of usual weekly hours in main job by age in hours	4	3	4	11
Gross domestic expenditure on R&D million EUR all sectors	4	4	1	9
Participation rate in education and training (last 4 weeks) total age 25-64	0	5	4	9
Gross domestic expenditure on R&D million EUR government	5	2	1	8
Total R&D personnel higher education sector full time equivalent (FTE)	4	2	2	8
Proportion of population aged 20-39	4	2	2	8
Intramural R&D expenditure - Government	8	0	0	8

Table 5: Fixed effects estimator - Interactions

	t	l1	l2	tot
Total duration	0	2	3	5
Activity rates ISCED >3	2	1	1	4
HH Net social contributions mil EUR	0	1	1	2
Self-employed persons	0	2	0	2
Total R&D personnel bussiness enterprise sector full time equivalent (FTE)	1	1	0	2
Unemployment Rate (Females)	2	0	0	2
Adult Participation in Learning	1	0	0	1
Economic activity rates age over 15 %	1	0	0	1
Gross domestic expenditure on R&D million EUR government	0	1	0	1
Other knowledge-intensive services	1	0	0	1
Participation rate in education and training (last 4 weeks) total age 25-64	0	1	0	1
Tertiary Educational Attainment (females)	1	0	0	1