Maastricht University

UNITED NATIONS UNIVERSITY

UNU-MERIT

Working Paper Series

#2022-016

Canonical correlation complexity of European regions

Önder Nomaler & Bart Verspagen

Published 22 April 2022

Maastricht Economic and social Research institute on Innovation and Technology (UNU-MERIT)
email: info@merit.unu.edu | website: http://www.merit.unu.edu

Boschstraat 24, 6211 AX Maastricht, The Netherlands
Tel: (31) (43) 388 44 00

UNU-MERIT Working Papers
ISSN 1871-9872

Maastricht Economic and social Research Institute on Innovation and Technology
UNU-MERIT

UNU-MERIT Working Papers intend to disseminate preliminary results of research carried out at UNU-MERIT to stimulate discussion on the issues raised.

Canonical Correlation Complexity of European Regions

Önder Nomaler & Bart Verspagen

UNU-MERIT

Abstract:

In an earlier paper (Nomaler & Verspagen, 2022) we introduced a 'supervised learning' based alternative to the competing unsupervised learning algorithms (e.g., Hidalgo and Hausmann, 2009 vs. Tacchella et al, 2012) proposed in the so-called 'economic complexity' literature. Similar to the existing ones, our alternative, which we refer to as the "Canonical Correlation Complexity Method (CCCM)", also aims at reducing the high dimensionality in data on the empirical patterns of co-location (be it nations or regions) of specializations in products or technologies, while the ultimate objective is to understand the relationship between specialization, diversification, and economic development. In our alternative method which combines the toolkit of the Canonical Correlation Analysis with that of Principal Component Analysis, the data on trade or technology specializations and multiple dimensions of economic development are processed together from the very beginning in order to identify the patterns of mutual association. This way, we are able to identify the products or technologies that can be associated with the level or the growth rate of per capita GDP, and (un)employment. In this follow up paper, we use the CCCM to analyse the development patterns of European regions in relation to their respective technology specializations. Our findings provide insights for EU's industrial policies, especially those considered under the 'smart specialization' framework.

Keywords: Economic complexity; economic development; supervised learning; Canonical Correlation Analysis; Principal Component Analysis; technological specialization; revealed technological advantage; European regional development; smart specialization

JEL codes: F14; F63; O11, R11

22 April 2022

This paper is an outcome of a project funded by the Economic Complexity unit of the Joint Research Centre (JRC, Seville) of the European Commission through 2020-2021. The views expressed and any remaining errors are the sole responsibility of the authors.

1. Introduction

In this paper we describe the second stage of our research project on the Canonical Correlation Complexity Method (CCCM) which combines the long-established methods Canonical Correlation Analysis (CCA) with Principal Correlation Analysis (PCA) in the context of economic complexity research. Our first paper in this research project (Nomaler & Verspagen, 2022) has described the method in detail, and reported on the application of CCCM to the case of international trade (exports) and its relation to economic competitiveness. The scope of the first paper was a broad international one, covering the 100+ largest economies in the world.

In this paper, we shift the attention to the economic competitiveness of European regions, and the role of technological capabilities in this. The role of technology in economic development and growth has been well researched (e.g., Freeman and Soete, 1997; Nelson and Winter, 1982). Technological change provides both productivity gains, and new products and services that represent welfare gains to society. Technology can also be guided towards the solution of societal problems, such as global warming (Schot and Steinmueller, 2018).

In the present paper, we look at technology as a source of economic development as indicated by GDP per capita and its growth rate (for a similar approach, see, e.g., Fagerberg, 1987), as well as its relation to employment and unemployment (see, e.g., Freeman and Soete, 1987). The focus is on sub-national European regions (as in Fagerberg et al., 1997).

Our approach is rooted in the complexity literature (for an overview, see, e.g., Freire, 2021), which stresses the role of productive capabilities in competitiveness of, in this case, regions. By focusing on technology, we interpret productive capability as mainly influenced by technology. We use patents as the technology indicator, which means that our emphasis is on an output indicator of the inventive process (see, e.g., Pavitt, 1985 for an overview of the characteristics of patents as technology indicators). This means that we disregard the influence of technology through the diffusion of international or interregional technology flows. These flows have been considered as an important source of technological catching-up (e.g., Abramovitz, 1986, Fagerberg, 1987, Fagerberg et al., 1997). By focusing on patents, the emphasis is on the (global) technological frontier, rather than on technological catching-up.

The main aim of the analysis is to use the CCCM to assess the competitiveness of the regions of the European Union (plus the United Kingdom) in light of their technological specialization as indicated by European patents. The aim is not to introduce the CCCM as such, or explore its general workings, as this has been done in Nomaler and Verspagen, 2022. In fact, we refer to this earlier paper for most (formal) details of the method, and assume here that the reader is familiar with the basics as described therein.

The rest of this paper is laid out as follows. In the next section, we describe the data that will be used in the analysis. This concerns both the data on economic competitiveness (development level, growth, employment and unemployment) and the data on patents (technological specialization). Unfortunately, and due to changes in regional classifications, data on the economic competitiveness variables is available only for a single recent period (2015 – 2018), which means we cannot apply the panel perspective in Nomaler & Verspagen, 2022.

In Section 3, we present some details of the CCCM that are useful for the specific context of this paper. Section 4 presents the part of the analysis that selects a value for the threshold parameter *f* that is used in the CCCM. This parameter governs how much of the variation between regions in terms of the large number (5,000+) of patent variables that is retained in the later stages of the algorithm. Contrary to the case in Nomaler and Verspagen, 2022, we find that a high value of the *f* parameter can be used in the data set under consideration here. Section 4 also summarizes the basic estimation results (including parameter values) of the CCCM. Here we face a choice between using the raw estimated parameters, which correspond to composite factors of the competitiveness variables, or so-called rotated components, which correspond to pure competitiveness variables. This choice is representational only, i.e., it has no impact on the predictive power, or the relation between the patent variables and the predictions of the method. As in Nomaler and Verspagen, 2022, we opt for rotated components, because these turn out to be easier to interpret.

Section 5 presents some basic outcomes with regard to the predictive power of the method. We show that the quality of in-sample predictions is high, but that out-of-sample predictions are somewhat weaker. However, the quality of these out-of-sample predictions can only be assessed for two out of five competitiveness variables, due to data limitations. As in Nomaler and Verspagen, 2022, we find that in-sample prediction residuals add power to the out-of-sample predictions. This implies that deviations between actual and predicted values tend to be persistent over time within the region.

In Section 6 we present the bulk of our results. This section deals with the technological competitiveness profiles of the European regions in our sample, by exploring the relationship between the five competitiveness variables and technological specialization as related to the patent variables. We link technological competitiveness to six broad technology fields, and we show that in terms of the relation between technology and economic competitiveness, Europe is divided into two major parts. One of these parts can be seen as leading, in the sense that technology is related to above-average economic performance in these regions, and another part is more peripheral in the sense that technology is related to sub-average performance. Section 7 summarizes our main line of argumentation and presents the conclusions.

2. Descriptive statistics

Our data set consists of all NUTS-2 regions of the European Union Member States, plus the UK.[1] We opt for the NUTS-2 level because at this level sufficient data are available for our economic variables, and the majority of regions has a large enough number of patents in order to be able to calculate meaningful patent indicators. We exclude one region (FI20, Åland) because the economic data are incomplete. At the NUTS-2 level, five countries consist of only a single region (Cyprus, Estonia, Latvia, Luxemburg, and Malta), while another three countries (Croatia, Lithuania, and Slovenia) consist of two regions. All other countries have three or more regions in our data set. The total number of regions is 275.

[1] We use the 2016 version of NUTS.

Table 2.1 shows descriptive statistics for the five variables that we use in the economic competitiveness data set. These variables are GDP per capita in 2015 (measured at current market prices in PPS to the Euro), the growth rate of GDP per capita (in the same units) over the period 2015-2018, the unemployment rate in 2015, the growth rate of the unemployment rate over 2015-2018, and the growth of employment (in persons) over 2015-2018. We report the descriptives for GDP per capita in K€ as well the natural log (ln) version (ln of the € value). The latter is what we will use in the analysis.

The table shows values per country (unweighted average of the regions in the country), as well as values for the entire set of 275 regions, and the averages of the country values (as reported in the table). As is well known, there are substantial differences between countries, as well as within countries. The richest country (in terms of GDP per capita) is Luxemburg, followed by Ireland, while the poorest is Bulgaria. However, the within-country standard deviation of GDP per capita is very high in Ireland, although it is even higher in the United Kingdom.

Growth of GDP per capita is highest in Romania, followed by a number of Eastern-European countries (Bulgaria, Estonia, Croatia, Lithuania, Latvia, Slovenia) as well as Cyprus. Unemployment is highest in Greece and Spain, followed by Croatia, Portugal, Slovenia and Italy. All these countries have double-digit unemployment rates. Germany and the United Kingdom have the lowest unemployment rates. The growth rate of the unemployment rate is negative in all countries (i.e., the unemployment rate declined), with many countries showing double-digit growth rates. However, countries showing high unemployment rates in 2015 (such as Greece, Italy and Spain) are not the ones that show the largest decline in unemployment. Finally, employment grows at a positive rate in all countries, although the rate is modest (in most cases at a slower rate than GDP per capita).

Figure 2.1 shows maps of these variables, as well as the number of patents. For the number of patents, we use totals over the 2010 – 2015 period. We use these cumulative numbers because the yearly numbers are rather volatile, especially at the high level of disaggregation that we will use in the CCCM. Using the cumulative number of patents over a somewhat longer period is also consistent with the strongly cumulative nature of technology (e.g., Nelson and Winter, 1982; Freeman and Soete, 1997).

Table 2.1. Descriptive statistics for variables of the economic competitiveness data set

		Averages						Standard deviation					
	N	Y€	Y	gY	U	gU	gE	Y€	Y	gY	U	gU	gE
Austria	9	36.9	10.5	1.8	5.0	-6.78	1.3	6.31	0.18	0.55	2.13	2.98	0.53
Belgium	11	33.1	10.4	1.6	8.7	-10.35	1.5	11.28	0.31	0.83	4.10	3.30	0.38
Bulgaria	6	12.2	9.3	4.4	9.8	-13.12	0.7	4.74	0.32	1.13	1.62	5.88	1.01
Cyprus	1	23.8	10.1	4.9	14.9	-14.54	3.6						
Czech Republic	8	25.4	10.1	3.2	5.1	-18.32	1.4	10.64	0.31	0.74	1.68	1.90	0.54
Germany	38	34.5	10.4	1.8	4.6	-8.77	1.1	7.59	0.21	0.84	1.67	3.42	0.80
Denmark	5	34.1	10.4	2.6	5.9	-6.40	1.4	7.44	0.20	0.61	0.35	1.13	0.53
Estonia	1	22.1	10.0	4.5	6.1	-5.46	0.9						
Greece	13	17.7	9.8	0.9	23.8	-5.93	2.1	3.62	0.18	1.67	4.37	6.32	1.27
Spain	19	25.3	10.1	2.2	21.3	-9.69	2.5	5.12	0.20	0.62	5.95	3.23	1.39
Finland	4	31.2	10.3	2.6	9.0	-7.52	1.3	6.52	0.19	0.80	0.91	1.60	0.27
France	22	26.7	10.2	1.0	9.7	-5.04	0.7	5.76	0.17	0.87	1.61	3.09	1.22
Croatia	2	17.2	9.7	4.0	15.7	-15.43	1.3	0.45	0.03	0.10	0.50	1.02	0.50
Hungary	8	18.9	9.8	2.8	6.7	-15.51	1.8	8.52	0.36	1.16	2.23	3.07	0.84
Ireland	3	47.9	10.7	1.4	9.7	-14.84	2.8	16.86	0.40	4.99	0.49	0.33	0.69
Italy	21	27.6	10.2	2.1	11.9	-4.37	0.8	7.57	0.28	0.53	5.11	3.36	0.72
Lithuania	2	24.7	10.1	4.3	8.8	-11.38	1.3	6.80	0.28	0.13	1.10	1.61	1.87
Luxemburg	1	78.0	11.3	1.2	6.3	-5.29	2.9						
Latvia	1	18.6	9.8	4.5	9.9	-8.08	0.2						
Malta	1	27.1	10.2	3.6	4.9	-10.88	6.2						
Netherlands	12	34.7	10.4	1.7	6.5	-15.62	1.6	7.23	0.20	1.11	0.98	0.96	0.64
Poland	17	18.7	9.8	2.7	7.6	-16.10	0.5	6.75	0.27	0.39	1.64	2.69	3.66
Portugal	7	21.8	10.0	2.4	12.8	-14.34	2.2	3.41	0.14	1.01	1.50	1.78	0.55
Romania	8	16.7	9.6	7.6	6.8	-12.88	0.6	8.68	0.39	1.11	2.44	1.80	0.89
Sweden	8	34.4	10.4	1.0	6.7	-5.70	1.6	6.88	0.17	0.96	1.00	2.60	0.49
Slovenia	2	24.0	10.1	4.1	9.0	-13.85	2.2	4.30	0.18	0.24	1.45	1.85	0.00
Slovak Republic	4	27.4	10.1	0.0	10.6	-14.79	1.8	16.19	0.49	0.81	3.41	2.43	0.41
United Kingdom	41	32.2	10.3	0.4	4.5	-6.21	1.0	24.23	0.36	1.04	1.43	6.89	1.18
Total sample	275	28.3	10.2	1.9	9.0	-9.37	1.3	13.79	0.38	1.74	6.19	5.73	1.45
Between countries		28.3	10.1	2.7	9.4	-10.61	1.7	5.55	0.14	0.92	1.74	1.90	0.73

Variable names: N = number of regions, Y€ = GDP per capita in K€PPS, Y = ln(GDP per capita), gY = growth rate (%) of GDP per capita, U = unemployment rate, gU = growth rate (%) of U, gE = growth rate of employment

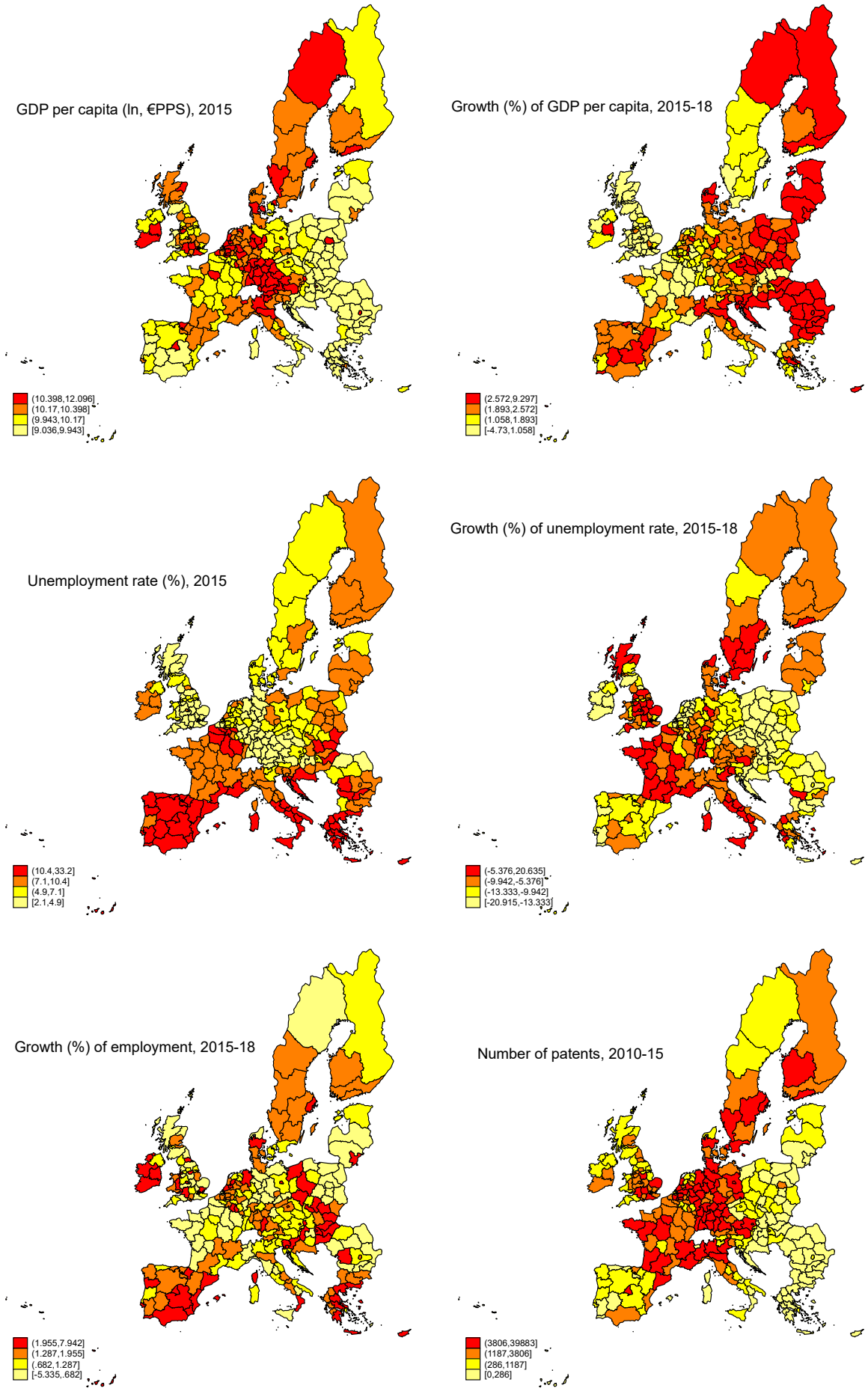


Figure 2.1. Maps of the variables in the economic competitiveness data set, and number of patents (patents are totals over 2010-2015 period)

The maps in Figure 2.1 show that all of these variables show spatial concentration. GDP per capita is (still) high in the so-called blue banana regions (although in our case, it is a red banana), which stretches from the south of the United Kingdom, through Belgium and the Netherlands, south Germany and the Alps to North Italy (see, e.g., Faludi, 2015). South of Ireland, Denmark and the Southern Scandinavian regions also show high GDP per capita.

The blue (red) banana does not, however, generally show high values of growth of GDP per capita. This is highest in the East (almost all of the post-EU-15 countries show high growth rates), as well as in Spain and Portugal. Unemployment is highest in the South, including France (especially the Northern French regions), and is generally low in the blue banana area. Unemployment grows rapidly in France and Italy, employment grows rapidly in Spain, Greece and part of Eastern Europe.

Finally, the number of patents is especially high in the blue banana area, as well as in Southern Sweden and North Denmark, North-East Spain, and large parts of Austria, France and Finland. On the other hand, most of Eastern Europe has low levels of patenting, with the exception of the Warsaw and Budapest areas.

Next, Figure 2.2 shows revealed technological advantages (RTA) of the regions. We define these in a similar way as the revealed comparative advantage (RCA) indicator that was used in the analysis of trade in our earlier paper (i.e., the indicator is scaled to [0,1] with ½ as the neutral value). Total global patents are used as the reference category, i.e., our RTA indicators indicate specialization of the European regions vis-à-vis the global totals, not intra-European specialization. Although we will use much more detailed technology classes in the CCA complexity analysis, for the purpose of the maps, we classified patents in just six technology groups. Five of these are derived from the classification proposed by Schmoch (2008). This classification is based on IPC (International Patent Class) codes, both at the 8-digit and 4-digit level. Every IPC code is assigned to one of the technology classes, which means that the classes as defined by Schmoch are mutually exclusive (an IPC code will be assigned to only one of the Schmoch classes). We also add a few IPC classes to the classification that were not included originally.[2]

The sixth broad technology class consists of all patents with a CPC classification with codes Y02 or Y04. The CPC classification is used by the patent office to tag patents on either climate change mitigation (CPC code Y02) or so-called smart electricity grids (CPC code Y04). Because the smart grids play a large role in sustainable energy use and distribution, these two CPC codes together represent technologies to combat climate change. Note that every patent that has a CPC tag also has an IPC class assigned to it. This means that while the five Schmoch classes, the CPC Y02/4 class overlaps with the five Schmoch classes.

The maps show somewhat of a dichotomy between, on the one hand, electrical engineering and instruments, which have relatively few regions with technological specialization, and, on the other hand, chemicals, mechanical engineering and other fields, where we find relatively many regions that are specialized in these fields.

[2] These are A61P (added to chemicals/pharmaceuticals), B33Y (additive manufacturing/3D-printing, added to mechanical engineering), C13B (added to food chemistry), G01Q (added to instruments/measurement), G16B (added to electrical engineering/computer technology), G16H (added to instruments/medical technology), G21B (added to mechanical engineering/other special purpose machines) and H04W (added to electrical engineering/digital communication).

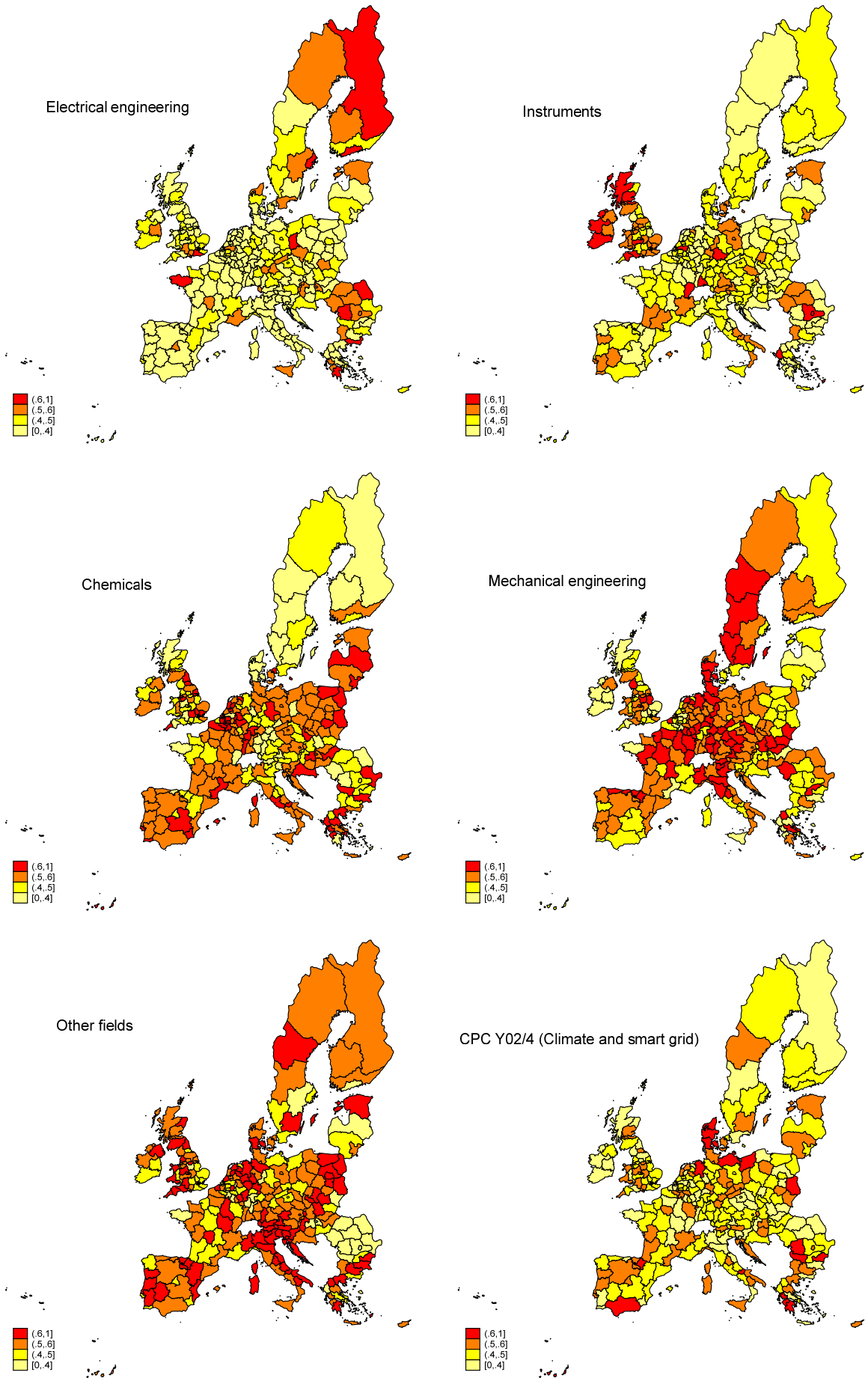


Figure 2.2. Maps of technological specialization (by patents) in six large areas (2010-2015)

In the map for CPC Y02/4 the regions with strongest specialization are found on the outskirts of Europe (most of Denmark. South France, East Poland and a few regions in Bulgaria and Romania. On the whole, this map shows relatively few regions with positive specialization in these CPC codes, which indicates that the majority of European regions is not specialized in climate-related technologies.

3. A brief overview of methods

For a general description and (mathematical) details of the CCCM, we refer to Nomaler and Verspagen (2022). Here we provide a brief overview, in line with the specifics of this second paper.

In implementing the CCCM, we use patent data at the finest possible aggregation level, which is 8-digit IPC classes and 8-digit CPC classes (Y02 and Y04). In our baseline estimations, we use cumulative patent data for the period 2010 – 2015, as was the case in the previous section. In order to reduce the impact of random noise, we exclude observations with few patents. This concerns both regions, i.e., regions with less than 10 patents over the period are excluded from the sample, and patent classes, i.e., patent classes (IPC or CPC) with less than 10 patents over the period are excluded. This leaves us, in the 2010 – 2015 period, with a sample of 267 (out of 275) regions, 5,019 (out of a possible 7,327) 8-digit IPC codes, and 48 (out of a possible 49) 8-digit CPC codes.

With five indicators of economic competitiveness, 267 regions and 5,067 (=5,019+48) patent-based variables, the raw data that we use in our analysis consists of the matrix **L**, a 267x5 sized matrix of the (de-meaned) competitiveness data set, and **M**, the 267x5,067 sized matrix of de-meaned RTA figures. The first step of our procedure is to reduce the dimension of **M** with principal component analysis (PCA):

$$\mathbf{M}^{*} = \mathbf{MV} \qquad (1)$$

where **V** is a 5,067x*n* matrix of loadings (i.e., this matrix contains the weights used to obtain the reduced-dimension scores). Essentially, **V** consists of the eigenvectors of the matrix product $\mathbf{M}^{\mathrm{T}}\mathbf{M}$ that are associated with its largest *n* eigenvalues (that account for no more than a fraction *f* of the variance in the RTA data set), as stacked horizontally in decreasing order of the associated eigenvalues. We will elaborate below on how we choose the value for *f*.

The standard Canonical Correlation Analysis (CCA) procedure takes the matrices $\mathbf{M}^{*}$ and **L** as input and computes the weight matrices **A** (of size 5x5) and **B** (of size *n*x5), as well as a vector **r** that contains the five canonical correlation coefficients. The elements of **r** can be seen as a measure of goodness of fit for each composite competitiveness dimension that is derived in the CCA. The weight matrices **A** and **B** are chosen by the CCA in such a way that the correlation coefficient between the first columns of **LA** and $\mathbf{M}^{*}\mathbf{B}$ is maximized, and given this relationship, the correlation between the second columns is maximized, etc.

The matrix product **VB** (5,067x5) contains the 'complexity' scores for the patent classes, separately in each of the five competitiveness dimensions. The matrix product **MVB** (267x5), contains the region-level complexity scores, again, separately in each five competitiveness dimensions.

In predicting, we can choose to either predict the values of the composite factors (**LA**) generated by the CCA, or the underlying (five) individual indicators that make up the matrix **L**. Like in Nomaler and Verspagen, 2022, and as will be explained in Section 5 below, we will opt for the

latter, because this gives a more direct picture in the form of the indicators that we are interested in. In order to obtain this prediction at the indicator level, we need to perform what we call a rotation of the CCA results. The rotation post multiplies the patent class complexity scores and the region-level complexity scores by the matrix $\mathbf{rA}^{-1}$. Accordingly, our in-sample predictions (indicated by a hat above the matrix variable) of the five original competitiveness indicators populate the 267x5 sized matrix

$$\hat{\mathbf{L}} = \mathbf{MVBrA}^{-1} \quad (2)$$

and the patent class complexity scores that can directly be associated by the five original competitiveness indicators populate the 5,067x5 sized matrix are

$$\mathbf{C} = \mathbf{VBrA}^{-1} \quad (3)$$

Thus, $\hat{\mathbf{L}} = \mathbf{MC}$ (4)

4. Selecting the *f* threshold: stability vs predictive power

We will now proceed to present results on stability and predictive power of the CCCM for these data, with the aim to select a value for the threshold *f* as used in the CCA stage of the algorithm. The threshold parameter *f* governs how much of the variation in the patent data (as transformed into RTA values) we retain in the second (CCA) stage of the procedure. On the one hand, including more variation (a higher value for *f*) will lead to higher predictive power (at least in-sample), but on the other hand, a higher value for *f* may lead to instability (across time periods) of the scoring parameters for the RTA variables. In Nomaler and Verspagen, 2022, which dealt with data on international trade instead of patent data, we selected a value *f* = 0.65.

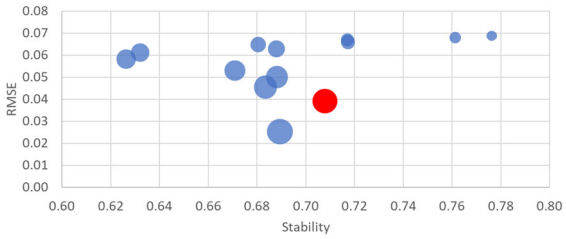
We have 5,067 patent-based variables in the data set that is used for the baseline estimations. Each one of those is an RTA indicator, scaled on the interval [0, 1] as before, indicating the region's revealed technological advantage in the related class. Also as before, we use the global patent numbers as the comparison, which implies that the RTA indicators represent specialization relative to the global totals, not intra-European specialization. All IPC classes together form one specialization domain, and the same holds for all CPC codes. In other words, the RTA indicators attached to IPC classes represent specialization the particular class relative to all other IPC classes, and the RTA indicators attached to CPC classes represent specialization the particular class relative to all other CPC classes.

Figure 4.1 shows the resulting relation between predictive power and stability of scores. The size of the dots corresponds to the *f* value that was used, with larger dots corresponding to larger values. We use values 0.45 to 0.95 in steps of 0.05, and add the value 0.99 as the last one. On the vertical axis, we show, for each of the five variables in the competitiveness data set, the root mean squared error (RMSE) of the in-sample predictions for the variable. Obviously, the aim is to have minimal RMSE, which corresponds to the best predictions. On the vertical axis, we show the stability of the estimated patent score parameters. This is not straightforward to measure, as we do not have a panel data set as in Nomaler and Verspagen, 2022. With the panel data set, we could split the sample into two periods, and compare the score parameters between those two periods.

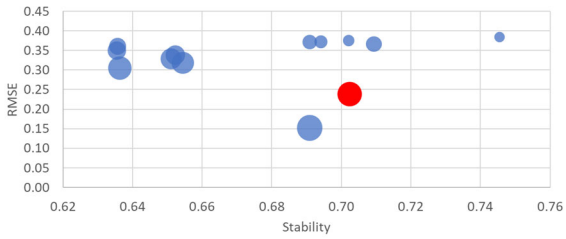
This is impossible in the current data set, as we have data on our competitiveness variables for only one period (2015-2018). Therefore, we resort to another way of judging stability of the scores. We implement two slightly different data sets, which both use the same data for the competitiveness variables, but have two different patent data sets. In the first data set (the baseline as we described it above), we use cumulative patent data for the period 2010 – 2015, as

was the case in the previous section. In the second data set, we use patent data for a shorter period, i.e., 2013 – 2015. We then compare the resulting score parameters (for patent classes) for the two data sets by calculating the correlation coefficient between the scores.

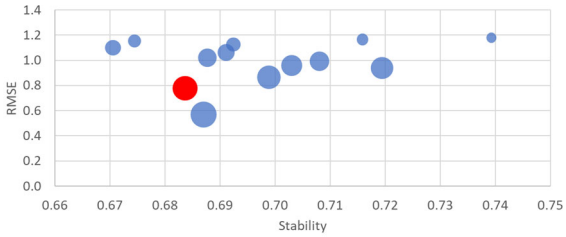
Obviously, because we use cumulative numbers of patents over the period and the shorter period is embedded in the longer one, this threshold is more restrictive in case of the shorter period. Therefore, we have a different number of regions and patent classes in the two samples. We deal with this by calculating correlation (as an indicator for stability) between the classes that are present in both estimations.



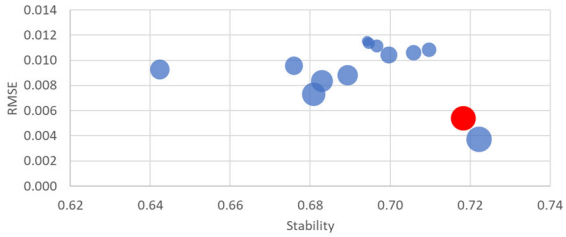
GDP per capita



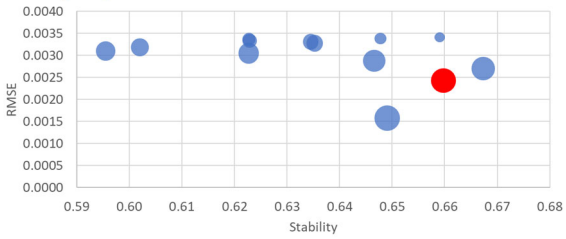
Growth of GDP per capita



Unemployment rate



Growth of unemployment rate



Growth of employment

Figure 4.1. Stability and RMSE of in-sample predictions

The figure shows a rather different trade-off than what we presented in Nomaler and Verspagen, 2022, where we used data on trade specialization, as well as a different set of competitiveness variables. In the present case, the expectations about predictive power (RMSE) are by-and-large confirmed: we find that the largest f values (largest dots) always yield the lowest RMSE. And although the picture is slightly mixed over the five variables, we find no substantial stability penalty for higher f values. In two cases (growth of the unemployment rate and growth of employment), the largest f values also show the highest stability. In two other cases, GDP per capita and its growth rate, the highest f values yield intermediate stability, and in one case (the unemployment rate), the highest f values yield stability values on the lower side.

This suggests that (in)stability is not something to worry about in this data set, and therefore we proceed to use a high f value. We pick f = 0.95, mostly because the higher value (f = 0.99) sometimes brings us close to the limit of the number of variables (PCA components), which is

equal to the number of regions. With $f = 0.95$, our baseline run uses 205 PCA components, which is well below the 267 regions. The observations in Figure 3.1 with $f = 0.95$ are colored red.

We now proceed to the details of the CCA estimation with $f = 0.95$. Table 4.1 shows the loading matrix **A** for the five competitiveness variables. Each column presents the loadings for one particular factor, and the corresponding canonical correlation of these factors is documented in the last row. In the interpretation of these coefficients, it is important to keep in mind that a high unemployment rate, as well as (positive) growth of the unemployment rate are generally considered non-desirable features of the regional economy, while the three other variables are generally considered as desirable properties. Thus, the first factor, with its high loading on GDP per capita and relatively high loading on the growth rate of unemployment, combines "positive" and "negative properties of the regional economy.

This is different for the second factor, which also has a high loading on GDP per capita, a relatively high loading on the growth rate of GDP per capita, and relatively strongly negative loading on the unemployment rate, which are all "positive" properties. But the second factor also has a relatively strongly negative loading on the growth rate of employment, which is a negative property.

Table 4.1. The loading matrix (A) for the competitiveness data set, and canonical correlations

	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5
GDP per capita	0.819	2.827	0.727	-0.745	-0.696
Growth of GDP per capita	-0.012	0.212	0.443	0.186	0.360
Unemployment rate	-0.074	0.044	0.108	-0.109	-0.087
Growth of unemployment rate	0.129	-0.117	0.108	0.036	0.016
Growth of employment	0.055	-0.203	0.092	0.694	-0.308
Canonical correlation	0.956	0.917	0.875	0.808	0.748

Also, the third factor mixes positive properties (high positive loadings on GDP per capita and its growth rate, and mildly positive on the growth rate of employment) with negative properties (mildly positive loadings on unemployment and its growth rate). The fourth factor loads high on the growth rate of employment, but strongly negative on GDP per capita. Thus, also the fourth factor mixes positive and negative properties of the regional economy. Finally, also the fifth factor is mixed: strongly negative on GDP per capita and growth of employment, but positive on the growth rate of GDP per capita. This pattern seems indicative of a catching-up region (initially with low GDP per capita, but growing relatively rapidly).

The mixed character of all five factors makes them hard to interpret in a straightforward way. Thus we opt, as we did in Nomaler and Verspagen, 2022, to continue the interpretation of the results and our predictions in terms of the so-called rotated components, which provide a pure interpretation in terms of the five individual variables in the competitiveness data set. For completeness, we document these rotated loadings in Table 4.2.

Table 4.2. The rotated loading matrix (rA^{-1}) for the competitiveness data set

	Factors (labeled by variable)				
	GDP per capita	Gr of GDP pc	Unempl. rate	Gr of unempl. rate	Gr of empl.
GDP per capita	0.250	-0.640	-2.993	0.042	-0.002
Growth of GDP per capita	0.233	0.369	-1.068	-0.021	0.001
Unemployment rate	-0.016	0.764	3.071	0.019	0.002
Growth of unemployment rate	0.002	0.458	-1.459	-0.012	0.009
Growth of employment	-0.090	0.861	-2.103	-0.003	-0.006

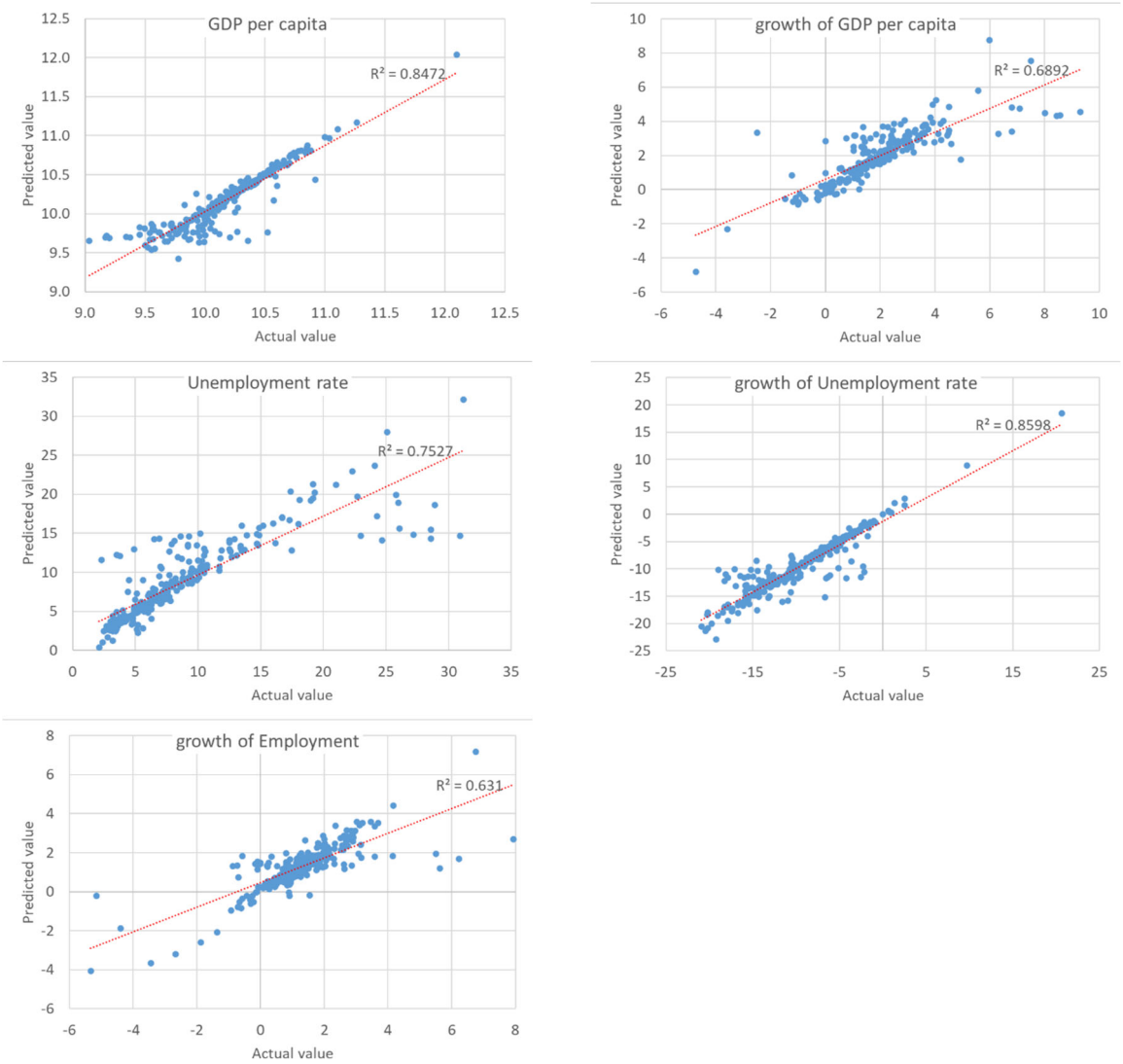


Figure 5.1. In-sample prediction

5. Prediction

We now move to look in detail at how well the CCCM predicts the five competitiveness variables. Figure 5.1 shows the correlation plots between the actual value of the variables (horizontal axis) and the in-sample predicted value (vertical axis). All of these correlations are fairly strong, with the highest R^2 obtained for the growth of the unemployment rate (0.86), closely followed by GDP per capita (0.85). Some of these relations indicate a degree of misspecification, indicated by a non-linearity in the plots. For example, the model predicts that a range of regions is at a relatively low level of GDP per capita, while in fact these regions do vary in regard of this variable (this is indicated by a more or less horizontal cloud of points at the bottom of the graph). Another example is a set of regions that has rather high unemployment, but is predicted to have lower unemployment rates. However, overall, the in-sample predictions are of fairly high quality (higher than in Nomaler and Verspagen, 2022 where we looked at data on exports and countries).

Table 5.1 documents some summary statistics of the in-sample predictions. The comparison between the standard deviation of the actual regional deviations from the average and the standard deviation of regional predictions is an interesting feature. Remember that we use centered (de-meaned) data, which implies that we predict deviations from the average. The comparison between the two standard deviations provides additional information to the correlations documented in Figure 5.1. We see that the standard deviation of the predicted deviations is close to, although smaller than the standard deviation of the actual deviations.

The fact that these standard deviations are relatively close is a positive characteristic of the predictions. Relatively smaller standard deviations of the predictions would yield very flat slopes in Figure 5.1, indicating that the variation of predictions is much smaller than the variations of the actual values, and this would make the predictions less accurate. The standard deviations are, roughly, in the range of 80 – 90% of the actual standard deviations, with the highest value (92%) obtained for GDP per capita and the lowest value (78%) for the growth rate of employment.

Table 5.1. Prediction-related summary statistics

	GDP pc	gr GDP pc	Unemp	gr Unemp	gr emp
Average of the variable	10.171	1.932	9.009	-9.371	1.300
StDev of regional deviations	0.385	1.743	6.192	5.733	1.451
StDev of regional predictions	0.353	1.440	5.097	5.163	1.129

In terms of assessing out-of-sample predictions, we are severely limited by the data. As already indicated, we have data for the variables in the competitiveness data set only for the period 2015-2018. Therefore, we have no possibility to investigate the quality of any out-of-sample predictions for the growth rate variables in the data set, simply because the actual data are not available. However, with some adjustments, we can investigate the quality of the 2018 out-of-sample predictions for (log of) GDP per capita and the unemployment rate. While the in-sample predictions of these variables refer to 2015, we used data for 2018 to calculate growth rates of these variables, and we can also predict the 2018 values using the CCCM. In order to maximize the variation in our patent-related variables (RTA) between estimation and prediction, we implement an estimation with patent data for 2013-2015 (cumulative number of patents overt the period, as before) and just GDP per capita and the unemployment rate for 2015 as the competitiveness variables. We use these estimations (the rotated patent class score variables) on 2015-2018 patent data (again, RTAs) to generate 2018 predictions for the two variables.

These predictions are documented in Figure 5.2 (correlations). Both correlations are (much) lower than the corresponding ones for in-sample predictions. Thus, the out-of-sample predictions are worse than the in-sample predictions, as was the case in Nomaler and Verspagen, 2022. The correlation for GDP per capita is the highest of the two (R^2 = 0.47). For the unemployment rate, we see that some of the predictions are negative, which is actually an economic impossibility, and which therefore detracts from the quality of this prediction.

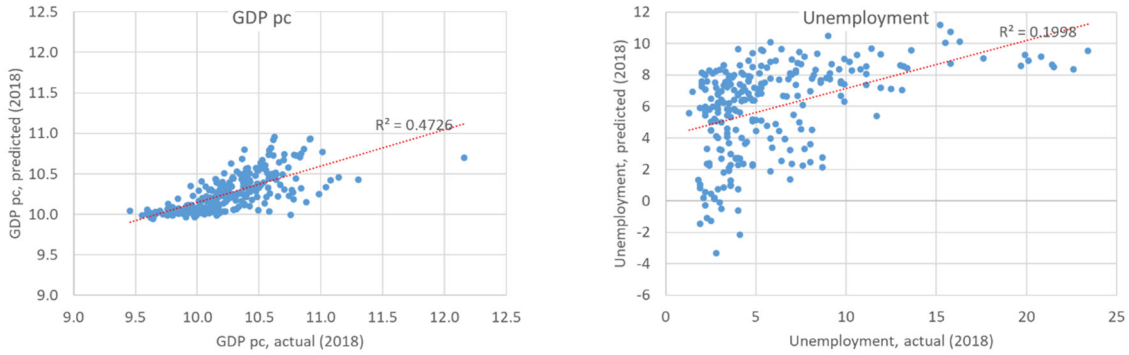


Figure 5.2. Out-of-sample prediction, without in-sample residuals

As in Nomaler and Verspagen, 2022, we also present predictions in which we add the in-sample residual from the estimation on which the predictions are based. This assumes that the prediction errors are somehow persistent within a region over time. These out-of-sample predictions are documented in Figure 5.3. The observed R^2 values are indeed somewhat higher than in Figure 5.3, which suggests that the residuals indeed have predictive power. However, there are still some regions with negative predicted unemployment.

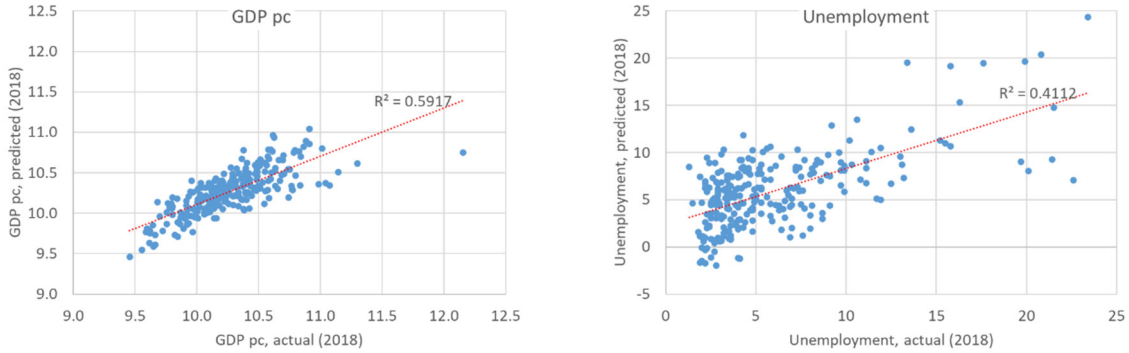


Figure 5.3. Out-of-sample prediction, with in-sample residuals

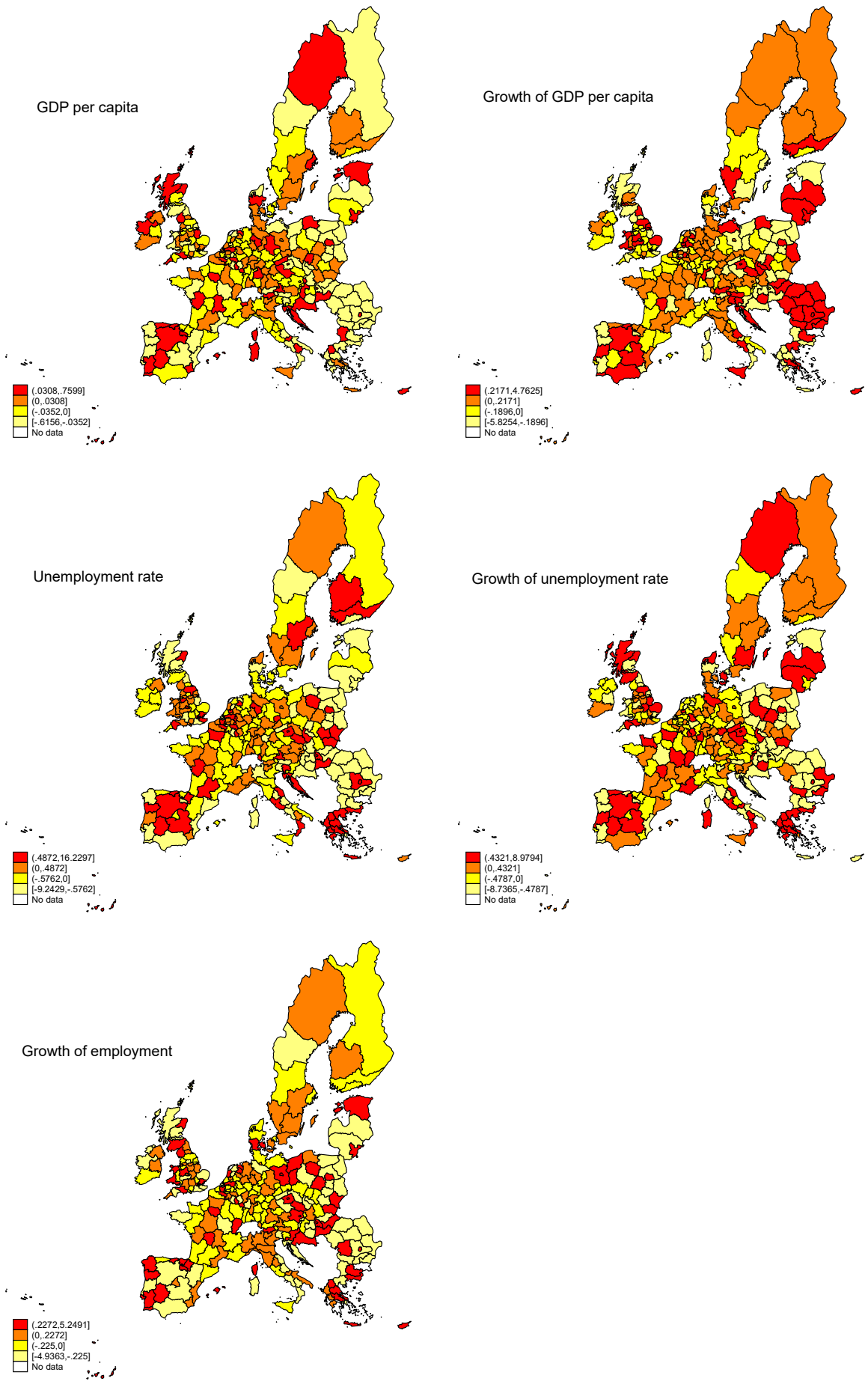


Figure 5.4. Maps of the in-sample residuals, per variable

The fact that these in-sample residuals are important for the out-of-sample predictions also suggests that there are variables related to regional economic performance that are not included in our data set, and hence have been left out of the CCCM. If such variables exist, and if these variables change slowly over time, then we would indeed expect that the in-sample residuals are persistent over time. And although technology is an important determinant of economic performance, it is to be expected that there are other variables that are important for prediction (but have not been included in the analysis).[3]

Figure 5.4 provides a geographical overview of the value of these residuals. Each of these maps uses four colors, where the yellows indicate negative residuals, and the orange and red hues indicate positive values. The borderlines in the positive or negative domain is formed by the median value, of either positives or negatives.

The overall impression from these maps is that positive and negative values of the residuals are pretty much spread-out over the European Union. There are some clusters of especially the red values, e.g., for Bulgaria and Romania for the growth of GDP per capita, and for a part of Spain for all variables except the growth of employment. But it seems to be the case that such geographical patterns are a minor part of the explanation of the residuals of the in-sample predictions of the CCCM.

6. Which technologies are related to economic performance?

We now turn to investigating how the patent class variables (regions' RTAs in 8-digit IPC and CPC Y02/4 codes) are related to economic performance, as indicated by the outcomes of the CCCM (with $f = 0.95$). First, we look at the rotated scores of the patent (IPC+CPC Y02) variables for the five variables in the competitiveness data set. It is important to remember that whereas other approaches in the 'economic complexity' literature provide a single-dimensional indicator for the fine-grained technology classes, our approach generates several such indicators (fives in this case), each associated with a different dimension of economic performance of the regions. These are the respective columns of the 5,067x5 sized matrix **C** given by equation 3.

Because the five economic competitiveness variables are only weakly, and sometimes negatively, correlated, we would generally expect that the patent class scores (columns of matrix **C**) would also not be strongly correlated. Figure 6.1 shows, as an example, the scatter plot between the first and third columns of matrix **C**, which are respectively associated with per capita GDP and the unemployment rate. The key thing to observe is that the mildly negative relation is, most of all, highly scattered. Only a fraction of the patent classes that are associated with relatively higher (lower) performance are also associated with higher (lower) unemployment rates. Table 6.1 gives the correlation coefficients between the five columns of matrix **C**, as well as the correlation coefficients between the original variables (between brackets). All of these correlations are relatively low. In terms of the scores of the patent class variables (RTAs), this means that the five dimensions that the CCCM identifies indeed measure different things.

[3] We also ran the algorithm with country dummy variables included in the "right hand side" data, alongside the patent variables. This did not yield any major differences as compared to the results reported here. Thus, it seems that country dummy variables are generally not good candidates for "explaining" the persistence of the in-sample residuals.

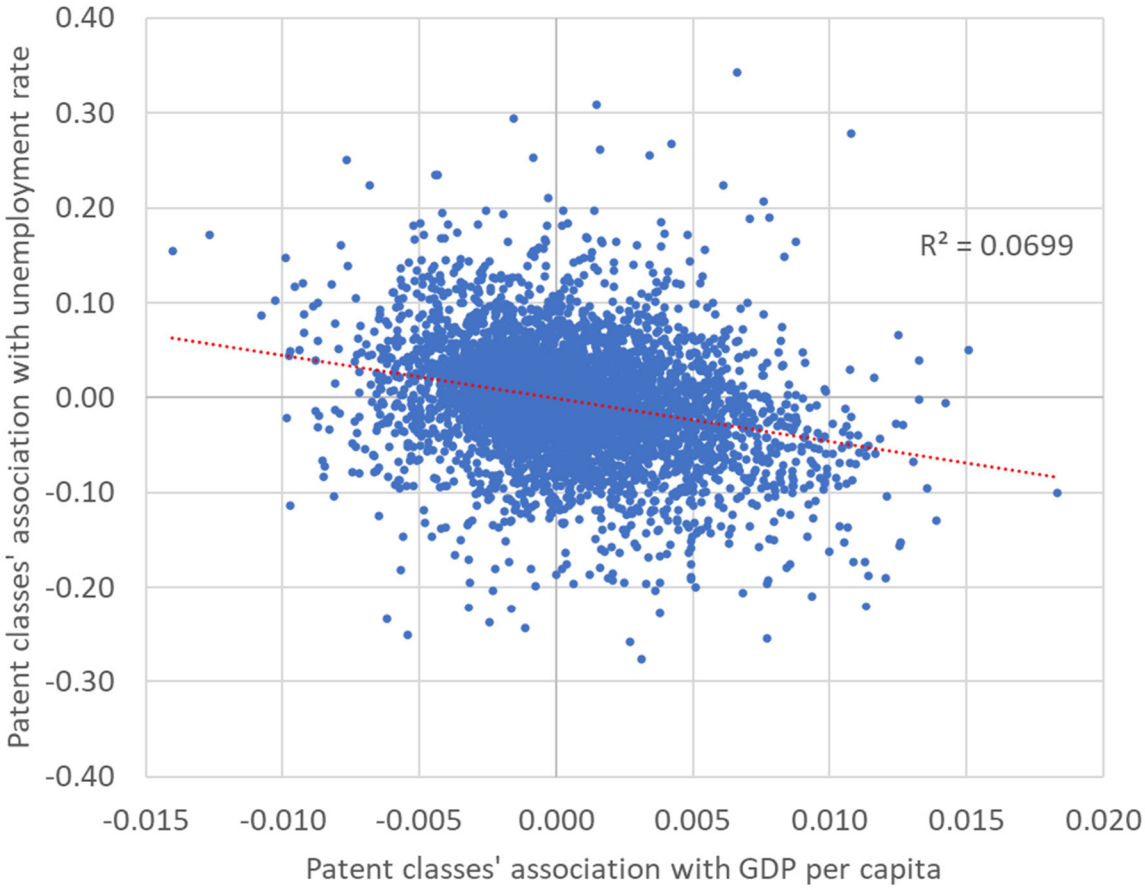


Figure 6.1. The first two dimensions of the totated scores of patent class variables plotted agians each other: Scores associated with GDP pc vs. those associated with the growth rate of GDP pc.

Table 6.1. Correlation matrix of the patent class scores (columns of matrix C), and of the five competitiveness variables

	gr GDP pc	Unemp	gr Unemp	gr Emp
GDP pc	-0.118 (-0.338)	-0.264 (-0.374)	0.165 (0.260)	0.034 (0.112)
gr GDP pc		-0.004 (0.043)	-0.210 (-0.334)	0.054 (0.002)
Unemp			0.069 (0.009)	0.242 (0.212)
gr Unemp				-0.312 (-0.282)

Note: values between brackets are the correlation coefficients between the original competitiveness variables (over 267 regions).

As we have 5,067 patent variables, there is too much information to consider each individual patent class. Therefore, we analyze the patent score variables at the level of the aggregate classes as defined in the Schmoch classification that was used above in Section 2. In this case, we look, in first instance, at the lower-level classes instead of the five classes of Section 2. There are 33 of such classes, and these are presented in Table 6.2, which also documents how these 33 classes are grouped into the five classes of Section 2.

Table 6.2. Technology fields based Schmoch's scheme

Electrical engineering		**737**
1	Machinery, apparatus, energy	354
2	Audio-visual technology	70
3	Telecommunications	76
4	Digital communication	36
5	Basic communication processes	84
6	Computer technology*	101
7	Semiconductors	16
Instruments		**608**
8	Optics	87
9	Measurement**	267
10	Control	114
11	Medical technology	140
Chemicals		**1814**
12	Organic fine chemistry	307
13	Biotechnology	60
14	Pharmaceuticals	45
15	Macromolecular chemistry, polymers	187
16	Food chemistry	105
17	Basic materials chemistry	435
18	Materials, metallurgy	242
19	Surface technology, coating	119
20	Micro-structure and nano-technology	19
21	Chemical engineering	223
22	Environmental technology	72
Mechanical engineering		**3151**
23	Handling	236
24	Machine tools & additive manufacturing***	476
25	Engines, pumps, turbines	366
26	Textile and paper machines	450
27	Other special purpose machines	580
28	Thermal processes and apparatus	267
29	Mechanical elements	287
30	Transport	489
Other fields		**1017**
31	Furniture, games	267
32	Other consumer goods	432
33	Civil engineering	318

Notes: * This field merges the fields "Computer technology" and "IT methods for management" from the original scheme; ** This field merges the fields "Measurement" and "Analysis of biological materials" from the original scheme; *** "Additive manufacturing" which was not in the original scheme was added to this field. The right column gives the number of 8-digit IPC classes in the field.

Figure 6.2 presents boxplots describing the distribution of the patent class scores (i.e., the five respective columns of the matrix **C** according to equation 3) over the 33 technology fields, with the set of CPC variables aggregated in an additional (34th) class. There is one set of boxplots for each variable in the competitiveness data set, as there is also one set of scores for each of these variables. We use rotated scores, which result from the loadings in the PCA that summarizes the information of the 5,067 patent variables, and the rotated score variables of the principal components. The colors in the boxplots indicate the six large technology groups (five Schmoch classes plus the CPC class), and the boxplots are presented left-to-right in the order of the technology fields in Table 6.1 (the CPC class on the right-hand side).

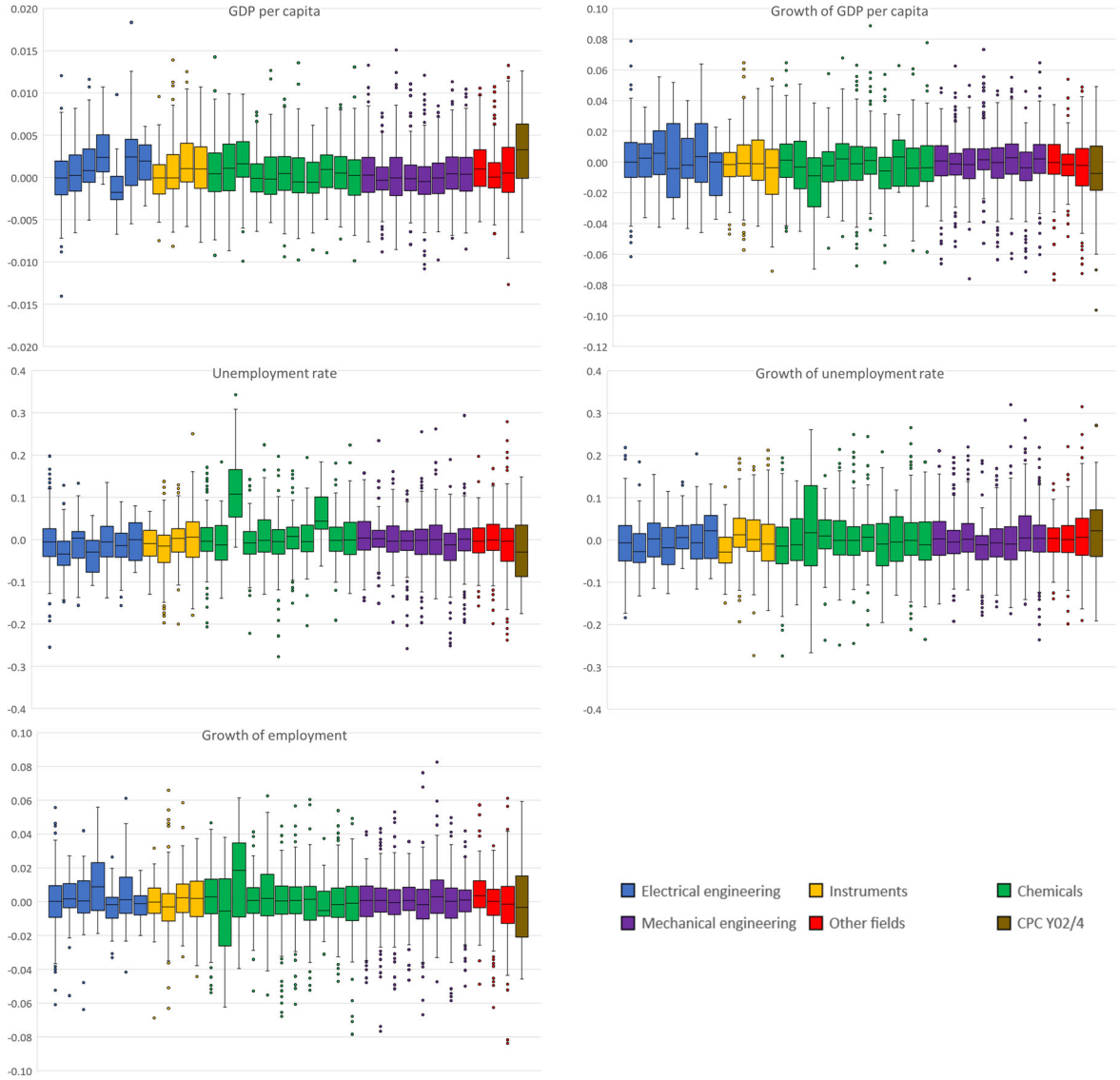


Figure 6.2. Boxplots of patant class score variables, five competitiveness variables, by technlogy field

The boxplots provide one overall impression, which holds almost without exception for all five variables: dividing up the 5,067 patent classes over the 34 groups does not discriminate the high from the low scores in terms of averages. This is seen by the fact that the distributions for the 34 groups are very similar: the boxes indicating the 1st to 3rd quartile range almost always overlap

for all 34 classes, and also most of the classes have outlier observations on the low as well as the high side. There are only two proper exceptions to this pattern, which are the pharmaceuticals class and the micro-structures and nano-technology class, both in the chemicals group, for the unemployment rate variable. For these two classes, the distribution of scores lies somewhat higher than for the other classes, indicating that, on the whole, specialization in these classes tends to be associated with higher unemployment rates.

However, the general tendency that the boxplots indicate is that all of these 34 technology classes contain patent codes that are associated with high values of the competitiveness variable, as well as low and middle-range values. In other words, which of the 34 technology field a specific 8-digit IPC (or CPC) code fits into, is not indicative of the for the nature of that 8-digit class in terms of the relation between technology specialization and economic performance.

This outcome is not specific to the technology classification of Schmoch. We also looked at the distribution of the scoring coefficients over technology fields defined in different ways. For example, we aggregated the 8-digit IPC codes into 4-digit IPC codes, of which there are 605, and into 3-digit IPC codes (there are 125 of these) as well as into NACE codes (we used the PATSTAT concordance between IPC and NACE, which yields 83 classes in our data set). Despite the fact that all these alternatives have more sub-categories than our 34 fields in Figure 6.2, they all yield the same basic conclusion: these ways of aggregating 8-digit classes do not help in distinguishing between low or high scoring coefficients of the detailed classes.

However, this does not necessarily mean that the large technology fields are not helpful in distinguishing between economic performance of the regions in our sample, because economic performance as predicted by the CCCM depends on the scoring coefficients combined with the regional technology specialization pattern (both at the lowest level of aggregation). The regional predictions are obtained by the product sum of the 5,067 technology specialization variables and their associated regional RTA scores (see equation 4). Thanks to the linear nature of the analysis framework, these regional predictions can be fully decomposed down to their 5,067 constituent contributors, or to any aggregation of these 5,067 classes, such as the Schmoch categories in Table 6.2. These aggregations can also be used for an 'analysis of variance' type of exercise.

To put it formally, for each *i* of our five different competitiveness variables, we rewrite equation 4 in a decomposed form as

$$\hat{\mathbf{L}}_i = \mathbf{M}\widehat{\boldsymbol{C}}_{\boldsymbol{\iota}} \; for \; \forall i = 1,2,3,4,5 \qquad (5)$$

where $\widehat{\boldsymbol{C}}_{\boldsymbol{\iota}}$ is the diagonalized form (of size 5,067x5,067) of the vector that is the i^{th} column of matrix **C**.[4] Clearly, for all practical purposes, there is too much detailed information in each 267x5067 sized matrix $\hat{\mathbf{L}}_i$. We can aggregate this (column-wise) into the Schmoch categories (or any other set of aggregate categories of choice, such as 3-digit IPC) by using the according concordance matrix **F** where $\mathbf{F}_{hg}$=1 if the 8-Digit that corresponds to row *h* belongs to the aggregate (Schmoch) category that corresponds to column *g* and *0* otherwise. Say the aggregation scheme comprises *q* categories. Then, the aggregated decomposition (facilitated by the 5067x*q* sized matrix **F**), for each competitiveness variable *i*, will be given by the 267x*q* matrix

$$\hat{\mathbf{L}}_{\mathbf{i}}^{\mathbf{Agg}} = \mathbf{M}\widehat{\boldsymbol{C}}_{\boldsymbol{\iota}}\mathbf{F} \qquad (6)$$

For example, for *i*=1, which corresponds to the per capita GDP indicator, and opting for the five broad Schmoch classes, plus the CPC Y02/4 classes aggregated into one Greentech category (thus *q*=6), the 6 elements on the h^{th} row of the 267x6 matrix $\hat{\mathbf{L}}_{\mathbf{1}}^{\mathbf{Agg}}$ will give the respective contribution

[4] $\widehat{\boldsymbol{C}}_{\boldsymbol{\iota}}(k,m) = \mathbf{C}(k,i)$ if $k = m$, and 0 otherwise.

of Electrical Engineering, Instruments, Chemicals, Mechanical Engineering, Other fields, and Greentech to our (in-sample) prediction of the deviation of the per capita GDP of region *h* from the mean value of this variable over all regions in our sample (i.e., $\hat{\mathbf{L}}_{\boldsymbol{h}}$). We can take it one step further and add to this matrix one more (i.e., a 7th) column that gives our in-sample prediction errors (i.e., $\mathbf{e}_{*,\boldsymbol{i}} = \mathbf{L}_{*,\boldsymbol{i}} - \hat{\mathbf{L}}_{*,\boldsymbol{i}}$). This gives a full decomposition of the actual (as opposed to predicted) deviations of the per capita GDP of regions from the mean. For each of our competitiveness indicator *i*, let us refer to this full decomposition as $\mathbf{L}_{\mathbf{i}}^{\mathbf{Agg+}}$.

Further note that, by construction, the average for each individual column of $\mathbf{L}_{\mathbf{i}}^{\mathbf{Agg+}}$ is zero, and just as our total predictions, reflects a dimension as a vector of deviations from a mean. Therefore, the (q+1)x(q+1) sized matrix product

$$Cov(\mathbf{L}_{\mathbf{i}}^{\mathbf{Agg+}}) = \left(\mathbf{L}_{\mathbf{i}}^{\mathbf{Agg+}}\right)^{\mathbf{T}}\left(\mathbf{L}_{\mathbf{i}}^{\mathbf{Agg+}}\right) \tag{7}$$

is the (q+1)x(I+1) sized covariance matrix of the decomposition of our de-meaned prediction of the *i*th competitiveness indicator, and it holds that

$$Var(\mathbf{L}_{*,\boldsymbol{i}}) = \sum_{m=1}^{q+1}\sum_{k=1}^{q+1} Cov(\mathbf{L}_{\mathbf{i}}^{\mathbf{Agg+}})_{mk} \tag{8}$$

That is, just as $\mathbf{L}_{\mathbf{i}}^{\mathbf{Agg+}}$ is a full decomposition of the *i*th (de-meaned) competitiveness indicator, $Cov(\mathbf{L}_{\mathbf{i}}^{\mathbf{Agg+}})$ is a full decomposition of the variance of the indicator. For better interpretability, the variance decomposition can also be expressed in terms of shares, and given the symmetry of a covariance matrix around the diagonal, in a diagonal plus upper diagonal form matrix

$$VarDecompShr\left(\mathbf{L}_{*,\boldsymbol{i}}\right)_{mk} = \begin{cases} 2\frac{Cov(\mathbf{L}_{\mathbf{i}}^{\mathbf{Agg+}})_{mk}}{Var(\mathbf{L}_{*,\boldsymbol{i}})}, if\ m < k \\ \frac{Cov(\mathbf{L}_{\mathbf{i}}^{\mathbf{Agg+}})_{mk}}{Var(\mathbf{L}_{*,\boldsymbol{i}})},\ if\ m = k \end{cases} \tag{9}$$

This variance decomposition is shown in Table 6.3 for all five variables. The six diagonal values in each panel (indicated in orange) are the respective variances of the contributions of the six technology categories to the overall variance of the variable. Of these six variances, the highest one is always associated to the mechanical engineering field (which is also the field that has most technology classes in our sample). For example, we see a contribution of 8.2% for GDP per capita, and 13.2% for growth of the unemployment rate. This suggests that mechanical engineering is the field that adds most to the regional predictions of per capita GDP. On the other hand, the contribution of the CPC variable is to predictions is only marginal, in each of the five panels.

These diagonal values always add up to a relatively minor share of the total variance of the variable (i.e., in each panel of Figure 6.3). For example, for per capita GDP, the diagonal values contribute 23.7% of the total variance, and for growth of the unemployment rate, they add 27.5%, which is the highest share among the five variables. The table also shows the share of the variance of the residuals, and we can note that one minus this share is the R^2 of the regression of actual values on the predicted values. The share of the residuals in total variance is smallest for growth of the unemployment rate, closely followed by GDP per capita, which suggests that these are the variables that the method predicts best (in-sample).

Table 6.3. Decomposition of the variance in each competitiveness indicator down to 5 Schmoch groups, CPC Y02/4 and (in-sample) prediction residuals

GDP per capita

	Elect Eng	*Instruments*	*Chemicals*	*Mech Eng*	*Other fields*	*CPC Y02/4*	*Residual*
Elect Eng	**0.020**	0.030	0.039	0.045	0.031	0.009	0.001
Instruments		**0.026**	0.059	0.057	0.046	0.013	0.002
Chemicals			**0.071**	0.098	0.064	0.018	-0.003
Mech Eng				**0.082**	0.074	0.015	0.000
Other fields					**0.034**	0.014	0.001
CPC Y02/4						**0.003**	-0.001
Residual							**0.153**

Share Variance (diagonals, orange): 0.237
Share Covariance (off diagonals blue): 0.610
Total Explained (R2): 0.847

growth of GDP per capita

	Elect Eng	*Instruments*	*Chemicals*	*Mech Eng*	*Other fields*	*CPC Y02/4*	*Residual*
Elect Eng	**0.015**	0.010	0.032	0.036	0.007	0.001	0.002
Instruments		**0.013**	0.044	0.042	0.019	0.004	0.002
Chemicals			**0.084**	0.128	0.056	0.011	0.003
Mech Eng				**0.092**	0.058	0.009	-0.002
Other fields					**0.021**	0.005	-0.003
CPC Y02/4						**0.002**	-0.002
Residual							**0.311**

Share Variance (diagonals, orange): 0.227
Share Covariance (off diagonals blue): 0.463
Total Explained (R2): 0.689

Unemployment rate

	Elect Eng	*Instruments*	*Chemicals*	*Mech Eng*	*Other fields*	*CPC Y02/4*	*Residual*
Elect Eng	**0.026**	0.038	0.029	0.060	0.024	0.007	0.005
Instruments		**0.027**	0.023	0.066	0.026	0.007	-0.001
Chemicals			**0.071**	0.110	0.041	0.003	0.008
Mech Eng				**0.099**	0.062	0.011	-0.010
Other fields					**0.019**	0.004	-0.002
CPC Y02/4						**0.001**	-0.001
Residual							**0.247**

Share Variance (diagonals, orange): 0.244
Share Covariance (off diagonals blue): 0.509
Total Explained (R2): 0.753

growth of the Unemployment rate

	Elect Eng	*Instruments*	*Chemicals*	*Mech Eng*	*Other fields*	*CPC Y02/4*	*Residual*
Elect Eng	**0.013**	0.010	0.040	0.040	0.013	0.001	-0.001
Instruments		**0.020**	0.049	0.075	0.027	0.007	0.001
Chemicals			**0.086**	0.162	0.051	0.010	0.001
Mech Eng				**0.132**	0.075	0.015	0.001
Other fields					**0.023**	0.006	-0.002
CPC Y02/4						**0.002**	0.000
Residual							**0.140**

Share Variance (diagonals, orange): 0.275
Share Covariance (off diagonals blue): 0.585
Total Explained (R2): 0.860

growth of Employment

	Elect Eng	*Instruments*	*Chemicals*	*Mech Eng*	*Other fields*	*CPC Y02/4*	*Residual*
Elect Eng	**0.015**	0.011	0.032	0.038	0.014	0.001	-0.002
Instruments		**0.008**	0.027	0.035	0.011	0.001	-0.002
Chemicals			**0.092**	0.115	0.043	0.008	0.015
Mech Eng				**0.098**	0.049	0.008	-0.011
Other fields					**0.020**	0.002	-0.001
CPC Y02/4						**0.001**	0.001
Residual							**0.369**

Share Variance (diagonals, orange): 0.235
Share Covariance (off diagonals blue): 0.396
Total Explained (R2): 0.631

For all variables, the largest share of the total variance is brought about by the covariances among the six technology classes, which is the sum of the off-diagonal values as indicated in blue. We observe marginal covariance values between the technology classes and the residuals (the area in green), however, by definition these values add up to zero. The share of the covariances in total variance is highest for GDP per capita, at 61%, closely followed by the growth of the unemployment rate (58.5%).

In order to investigate the nature of the covariances between technology fields in terms of their predictive power further, we apply PCA on the covariance matrix of the components of regional predictions by technology field. This is done for every of the five individual economic competitiveness variables, and the results are documented in Table 6.4, which presents the loadings on the six technology fields of the first component, as well as the explained variance of this first component[5]. The latter ranges between about 72% (for the growth rate of employment) and 80% (for the growth rate of unemployment).

From the PCA in Table 6.4, we would be able to construct five compositive factors (using the factor loadings) that can be used as (in-sample) predictions of the regional deviations in each variable. Obviously these five "technology fields PCA predictions" are of a lesser quality than the full (in-sample) predictions that were displayed in Figure 5.1, yet still rather good predictions as indicated by the large share of variance captured by the first component.

The loadings show a consistent pattern between the five variables, with the highest loading always found for the mechanical engineering field, followed by chemicals. The CPC field always has the smallest loading. To a large extent, this pattern mimics the size of the groups, as indicated by the last column in the table. Mechanical engineering is the largest field with about 41% of the 5,067 detailed 8-digit classes, followed by chemicals with 26.5%, and these are also the two fields with the highest loadings, across all five economic competitiveness variables. In fact, the correlation between the last column of the table and each of the other five columns is very large. This correlation coefficient ranges between 0.94 (for GDP per capita) and 0.99 (for the unemployment rate).

Table 6.4. Principal Components Analysis to maximize the share of explained variance of regional predictions

	GDP pc	gr GDP pc	Unemp	gr Unemp	gr emp	Sh IPC/CPC
Electrical engineering	0.2323	0.1534	0.2565	0.1388	0.1698	0.104
Instruments	0.3191	0.1953	0.2655	0.2287	0.1444	0.088
Chemicals	0.5596	0.6411	0.5176	0.5587	0.6487	0.265
Mechanical engineering	0.6178	0.6750	0.7259	0.7473	0.6918	0.408
Other fields	0.3760	0.2639	0.2591	0.2360	0.2229	0.125
CPC Y02/4	0.0893	0.0453	0.0409	0.0471	0.0341	0.009
Share of variance	0.731	0.768	0.704	0.802	0.717	

Note: the last column gives the share of the technology field in the total number (5,067) of fully disaggregated patent classes (CPC and IPC).

[5] I.e., the first 6 figures on i^{th} column Table 6.4 gives the leading eigenvector of the covariance matrix $\left(\mathbf{L_i^{Agg}}\right)^{\mathbf{T}}\left(\mathbf{L_i^{Agg}}\right)$ and the last figure is the leading eigenvalue, normalized by the sum of all 6 eigenvalues.

This reinforces the impression that the aggregation of the 5,067 8-digit codes into six broad fields resembles a more or less random process of grouping, an impression that was already emerging from Figure 6.2, which re-shuffles the 5,067 detailed classes in 34 technology fields, and the 34 technology fields into the six fields in Table 6.4. Therefore, the relatively high quality of the (in-sample) predictions that was documented in Section 5 above, must be seen as the result of variation in the scores of the 8-digit patent classes, not so much the variation at the level of 34 or 6 technology fields.

This result is, in a sense, discouraging for policymakers, because it suggests that the question which technology classes are related to economic performance can only be answered properly at the lowest level of aggregation, and this is not a very insightful level of analysis for identifying policy options. The (8-digit) technology classes with the strongest relation to economic performance are found in a wide range of technology fields, and the policymaker who would want to target these classes would have to be very specific and selective. If, on the other hand, the policymaker would target a broad class, like mechanical engineering, she would only reach a relatively small part of the detailed classes that are strongly related to economic performance.

In summary, we may picture the regions, or rather their specific technological specializations, as samples from the distribution of scores at the detailed level of 5.067 classes. Aggregating the sample to just six technology fields, provides a picture in which the covariances between the six fields dominate. But these covariances do not inform policymakers very well, because they provide little information on which specific technologies should be targeted. Moreover, it is clear that in terms of the six large technology fields, the ultimate result of this sampling in terms of regional variation is much dominated by the relative size of the six fields in terms of these classes: the larger classes (chemicals and mechanical engineering) account for the largest part of the variance.

This reasoning on the importance of variation at the 8-digit level technology classes vs. the importance of covariance at the level of six classes alerted us to the possibility of performing the analysis on the basis of more aggregated patent data. Therefore, we repeated the entire analysis as described so far (keeping also the value $f = 0.95$ for detailed results) with just 42 patent classes in the basic underlying data set (i.e., matrix **M**). These 42 classes are the 33 Schmoch classes as in Table 6.2, plus nine CPC classes (at the 4-digit level). Note that this data set defies the basic idea of the complexity literature that the "product space" (in this case patent classes) usually contains many more entities than the number of geographical units: we have 42 patent classes, which is smaller than the 267 regions.

We found that the outcomes of this analysis differ in two major ways relative to the 5,067 x 267 data set. First, the predictions that result from the 42 x 267 data set are by far inferior to those resulting from the larger data set. This can be seen, for example, from the fact that the prediction residuals account for 67.5% of the total variance of the regional deviations (this is an average across the five variables) in the results of the smaller data set, while they account for only 24.4% when using the larger data set.

Second, when we use the smaller data set, the pure variances of the six technology fields (i.e., disregarding the covariances between them) explain a much larger share of the variance of the total predicted deviations: on average 98.4% vs. 32.5% when using the larger data set. This means that if we wanted to provide policymakers with broader technology fields to base their action on instead of very specific (8-digit) patent classes, then we would have to accept a major loss in terms of predictability.

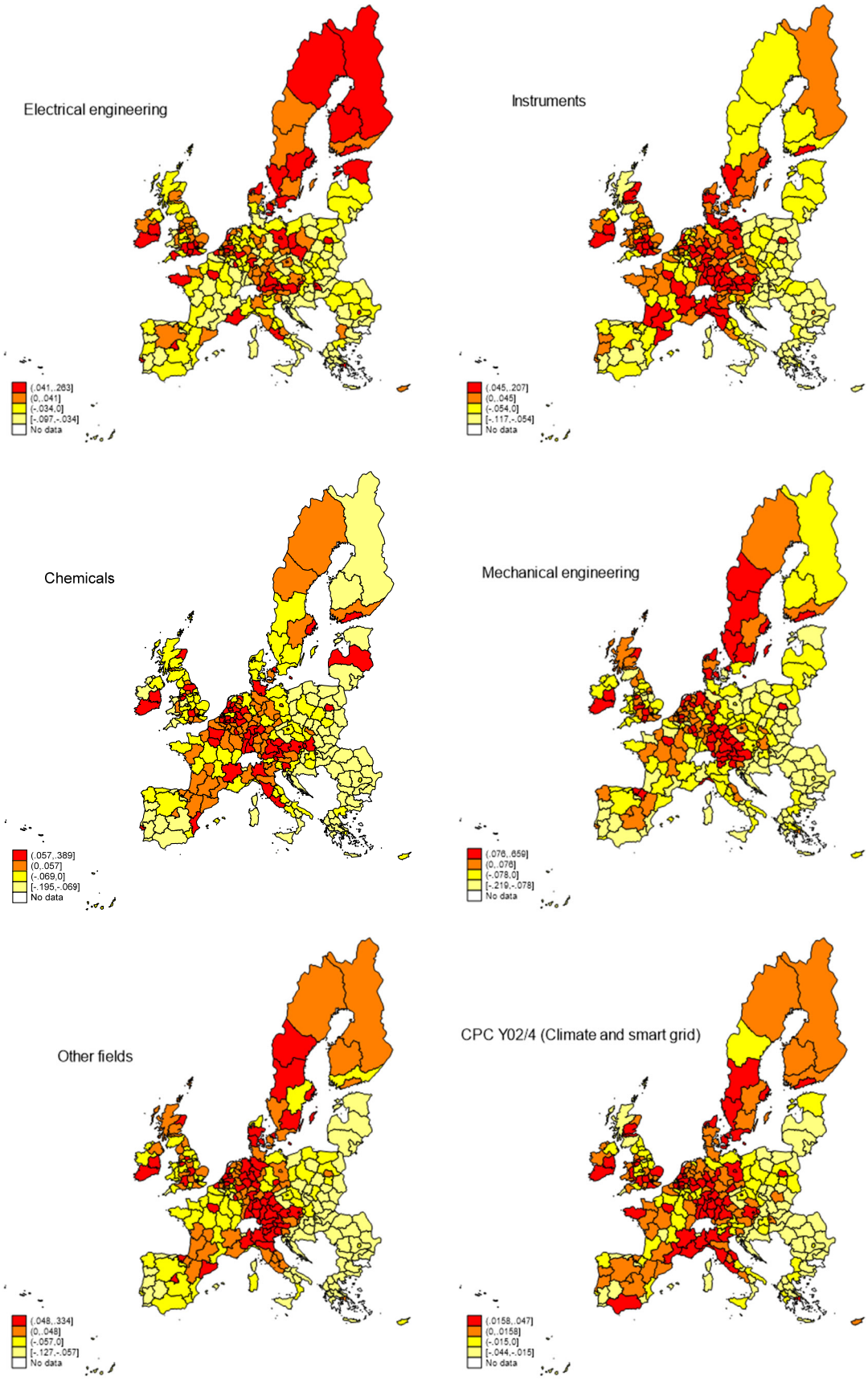


Figure 6.3. Predictions for regional disparities of GDP per capita, by six technology fields

To gain further insight into the predictions using the full (8-digit) detail, we explore the geographical dimension of the six-way (as in Table 6.3) technological decomposition of our predictions, by displaying the geographical variation given by each column of our 267x6 decomposition matrix $\mathbf{L_i^{Agg}}$, for each of the five competitiveness indicators. That makes 30 maps. The first six maps, that decompose the variation of our predictions of per capita GDP, are shown in Figure 6.3, while the other 24 maps can be found in the appendix. In each map, the predictions data have been arranged in four intervals. Two of these contain negative values, and two contain positive values. The two intervals with positive (negative) data have been separated by the median values of all positive (negative) values in the map.

These maps provide us with a valuable lesson on an important methodological issue regarding the level of aggregation in the economic complexity literature, and which also has policy implications. We illustrate the issue with the maps for GDP per capita in Figure 6.3, and the top-left panes of Figures 2.1 and 2.2 above. Figure 2.1 shows a clear concentration of relatively high levels of welfare (i.e., GDP per capita) in the so-called blue banana regions, as discussed above. Figure 2.2 suggests that, with RTAs computed at the level of the five Schmoch categories (plus CPC Y02/4), none but a few blue banana regions appear to be (strongly) specialized in technologies related to Electrical Engineering. This combination of observations suggests that there is no or only a weak association between technological specialization in Electrical Engineering and welfare. However, in all maps in Figure 6.3, including the one on Electric engineering, it is possible to identify (though broadly) the blue banana regions, which indicates that higher welfare is actually associable with certain subcategories (at 8-digit resolution) under each of the five Schmoch fields, certainly not excluding Electric engineering. The devil is in the detail.[6]

This suggests that (selective) 'diversification' is a better concept key to understand economic competitiveness/performance than 'specialization'. By its very nature, and as a concept for the policy maker, specialization must refer to rather aggregate entities, such as the Schmoch categories, or a sectoral scheme such as NACE. Our analysis shows that these classifications are not very useful to identify the potential sources of competitiveness (of European regions). On the other hand, diversification as a (policy) strategy seems to make more sense, because the highly disaggregated (8-digit) technology classes that are associated with a specific policy target variable (as GDP per capita, in Figure 6.3), can be targeted by diversification, even if it is along with other classes that are not associated to high levels of GDP per capita, or are related to high values of other policy target variables. From this point of view, the idea of 'smart specialization' (see, e.g., Balland et al., 2019) might better be coined 'smart diversification'.

Beyond this, the maps suggest that the predicted impact of the technology specialization variables on the economic competitiveness variables in European regions shows a clear and marked spatial pattern. In order to investigate this further, we conclude our analysis by summarizing the similarities between European regions in terms of the relationship between their technological specialization and economic variables in a network analysis.

To do this, we start from our decomposition of the predicted regional disparities in terms of the six technology fields (the five aggregated Schmoch fields plus CPC), That is, once again (i.e., as used to produce the maps in Figure 6.3 and in the appendix), we use the five 267x6 matrices $\hat{\mathbf{L}}_\mathbf{i}^{\mathbf{Agg}}$ (as defined by equation 6, for each of the competitiveness variables), each of which decompose the respective dimension of the total predicted regional disparity into six sub-category values. We stack these five matrices horizontally in order to construct a 30-dimensional space (6 technology

[6] Our discussion here is in terms of only one of our five variables, GDP per capita, but the conclusions extend to the other variables as well.

fields x 5 competitiveness variables), in which we can position the regions by their scores on each of the dimensions. In order to map this space, we *z*-score the 30 dimensions (for normalization), and then calculate the Euclidean distance between the regions in this *z*-scored space. Finally, we calculate the similarity between two regions as $10/d$, where d is the Euclidean distance in 30-dimensional space (the value 10 is an arbitrary scaling factor that doesn't influence the results). These similarity values are used to graph a network between regions.

We use the Linlog method (Noack, 2009) to graph this method, and use the VOS software (Waltman et al., 2010) to display the network. We also use the procedure in Newman (2004) to distinguish clusters of regions.[7] In order to stress the largest similarities in the network, we cut all regional connections with a value below 1.95. This threshold keeps 252 of the 267 regions connected to each other in the network, and the 15 non-connected regions become isolates. Threshold values larger than 1.95 break the network into components of which the second-largest contains more than one region. At the 1.95 threshold, 90% of all connections in the original network (which is fully connected) are cut.

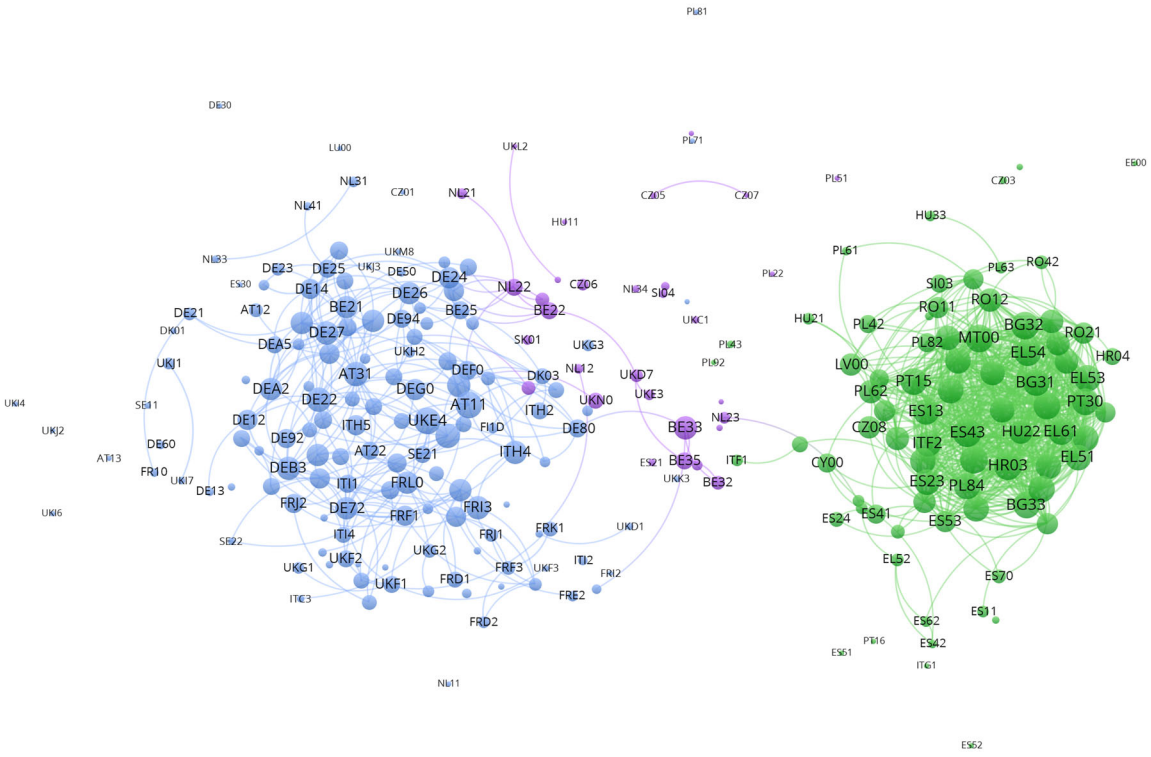


Figure 6.4. Network map of 252 European regions, based on similarity of their distribution of predicted regional disparities over technology fields

The network is displayed in Figure 6.4. There are three clusters, distinguished by color, which show a clear coherence. On the right side of the figure, we see the green cluster with many Eastern and Southern European regions. The blue cluster on the left-hand side contains many regions from

[7] We set the attraction parameter in VOS to 6, repulsion to 0, and resolution to 1. Minimum cluster size is set to 4 and clusters smaller than this size are merged with the nearest larger cluster.

the blue banana zone, as well as (other) regions from Germany, the Netherlands, France, and Scandinavia. The purple cluster is in between.

This network and its partition suggest a European regional divide. Figure 6.5 presents the European regional map with regions colored according to the clusters in Figure 6.5. The colors are the same between the two figures, with the added red color for the regions that dropped out of the largest network component due to the thresholding. The map in Figure 6.5 brings out clearly that in terms of regional predictions of the five competitiveness variables by the CCA complexity method, Europe is divided in a center-periphery pattern.

The center consists of West/Central Europe, Scandinavia, the South of the UK, and North Italy. The periphery is South Europe except North Italy, and Eastern Europe. We find isolated parts of the periphery that are similar to the center, e.g., parts of Poland and the Madrid regions. The purple regions are in between center and periphery, and are mostly found in Belgium, the Netherlands, the Czech Republic, Hungary and Poland. The red regions, which are dissimilar to the rest of Europe, are found scattered over the map.

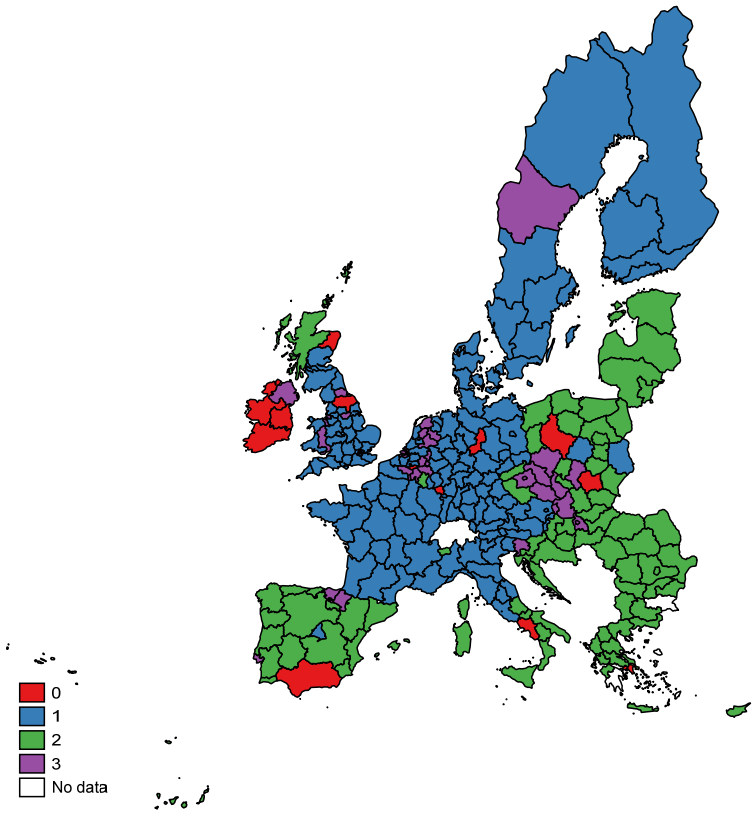


Figure 6.5. Network clusters of 252 European regions, geographical representation

7. Summary and conclusions

In this second paper of the Canonical Correlation Complexity project, we analyzed European (sub-national) regions, taking their technological specializations in terms of detailed 8-digit patent classes as the "predictors" of five economic competitiveness variables: GDP per capita and its growth rate, the unemployment rate and its growth rate, and the growth rate of employment. One basic conclusion is that the CCCM generates relatively good predictions, both in-sample and out-of-sample.

Like in Nomaler and Verspagen, 2022, which used trade data to predict a slightly different set of economic competitiveness variables for countries, we find here that the in-sample prediction errors (residuals) add to the out-of-sample predictive power. This suggests that the part of economic competitiveness that is not related to technological specialization is persistent over time. Thus, while it is entirely reasonable that technology specialization alone cannot make prefect predictions, we also learn that the part of the economic variables that is not related to technological specialization changes only slowly over time.

The CCCM identifies a number of patent classes that are positively (or negatively) related to each of the economic competitiveness variables. It would be tempting to use lists of detailed (8-digit) technology classes to inform policymakers about which technologies are related to strong economic performance, so that they could target these technologies in their policies. But such a recommendation has several major pitfalls.

First, and this is generally applicable to any industrial or innovation policy that targets specific goals, what may work in one regional context may not work in another, and, moreover, several regions targeting the same technology classes at the same time may induce competition that diminishes the effect of the policy ("not every region can be a Silicon Valley"). Thus, we need to be very careful in drawing simple but far-reaching policy conclusions from the CCCM results. Based on our analysis in Nomaler and Verspagen, 2022, which provides a detailed comparison of the CCCM with other complexity algorithms, we feel that this caution should be applied to the entire set of complexity algorithms.

Second, our results suggest that the scores of the 8-digit patent classes (i.e., their weights in predicting the economic competitiveness variables) cannot very easily be aggregated to broader technology fields with the aim to inform policy. If we aggregate the 5,067 individual 8-digit classes to six broad technology fields, the variation within the six fields accounts for only a small fraction of the total variation of the regional predictions (typically around or slightly below one third of the total variance). This means that it is very hard to formulate policy recommendations that suggest stimulating one or a few of the broad technology fields across European regions.

The reason for this result is that each of the six broad technology fields contains several 8-digit patent classes with low scores/weights, as well as several classes with high scores. Targeting the broad technology classes does not distinguish enough between the technology classes that are positively or negatively related to economic performance. This leaves the policymaker who wants to use our results with two options. On the one hand, she could target very specific technology classes, and on the other hand, she could try to stimulate technological diversification. The first of these options (targeting detailed technology classes) is difficult because they are very specific (and because of the first pitfall identified above). The second option, diversification, may target broad technology fields (in specific regions), with the aim to make a broad range of technologies accessible, among which are technologies related to specific economic policy variables, such as growth and employment.

In this respect, we suggest that the various decompositions[8] that we proposed in Section 6 above, and which are largely complementary to our core method (canonical correlation-based complexity analysis), may prove to be a useful input in the 'smart specialization' policy discussion. These decompositions, which are likely also applicable to other complexity algorithms, are a potential policy toolkit that combines the methods of the economic complexity and the product space literatures.

In terms of further research, one option that we want to pursue concerns the tradeoff that we identified between the level of aggregation of the basis patent data set and the nature of predictions obtained using CCCM. In our main analysis, we used a very detailed patent data set (8-digit technology codes), and this yields good predictions, but little usefulness of the broad technology fields, as summarized above. On the other hand, when we used a more aggregated patent data set from the start (42 technology fields instead of 5,067 8-digit classes), the usefulness of the six technology fields increased, but the quality of the predictions dropped considerably. We expect that such a tradeoff will also manifest itself for other complexity algorithms (as discussed in detail in Nomaler and Verspagen, 2022). But this expectation remains to be investigated in further research.

References

Abramovitz, M. (1986). Catching Up, Forging Ahead, and Falling Behind. The Journal of Economic History, 46(2), 385-406.

Balland, P.A., Boschma, R., Crespo, J., and Rigby, D.L. 2019. Smart specialization policy in the European Union: relatedness, knowledge complexity and regional diversification. Regional Studies, 53(9): 1252-1268.

Fagerberg, J. 1987. A technology gap approach to why growth rates differ, Research Policy, 16 (2-4): 87-99.

Fagerberg, J., Verspagen, B. and Caniëls, M. 1997. Technology, Growth and Unemployment across European Regions, Regional Studies, 31 (5): 457-466.

Faludi, A. (2015). The 'Blue Banana' Revisited, *European Journal of Spatial Development*, vol. 56.

Freire, C., 2021, Economic Complexity Perspectives on Structural Change, in: Foster-McGregor, N., Alcorta, L., Szirmai, A. and B. Verspagen (eds), New Perspectives on Structural Change. Causes and Consequences of Structural Change in the Global Economy, Oxford: Oxford UP.

Freeman, C. and L. Soete, 1987, Technology and full employment, Wiley-Blackwell.

Freeman, C. and L. Soete, 1997, The Economics of Industrial Innovation, Pinter Publishers.

Nelson, R., & Winter, S. (1982). An evolutionary theory of economic change. Cambridge, Mass. and London: Belknap Harvard.

Newman, M.E.J. (2004). Fast algorithm for detecting community structure in networks. Physical Review E, 69, 066133.

Noack, A. (2009). Modularity clustering is force-directed layout. Physical Review E, 79, 026102

[8] I.e., the matrices $\hat{\mathbf{L}}_i$ and $\hat{\mathbf{L}}_{\mathbf{i}}^{\mathbf{Agg}}$ introduced respectively by equations 5 and 6.

Nomaler, Ö and B. Verspagen, 2022, The Canonical Correlation Complexity Method, UNU_MERIT Working Paper Series, #2022-015.

Pavitt, K., 1985 Patent statistics as indicators of innovative activities: Possibilities and problems. Scientometrics 7, 77–99.

Schmoch, U. (2008). Concept of a Technology Classification for Country Comparisons. Final Report to the World Intellectual Property Organisation (WIPO), downloaded from https://www.wipo.int/export/sites/www/ipstats/en/statistics/patents/pdf/wipo_ipc_technology.pdf

Schot, J. & Steinmueller, W.E. (2018). Three frames for innovation policy: R&D, systems of innovation and transformative change. Research Policy 47 (9), 1554-1567.

Waltman, L., Van Eck, N.J., & Noyons, E.C.M. (2010). A unified approach to mapping and clustering of bibliometric networks. Journal of Informetrics, 4(4), 629-635.

Appendix. Additional maps

This appendix documents a number of additional (to Figure 6.3 of the main text) maps, each of which breaks down the regional predictions of the competitiveness variables by six technology fields (i.e., respective columns of the matrices $\hat{\mathbf{L}}_i$ as given by equation 5).

In each map, the predictions data have been arranged in four intervals. Two of these contain negative values, and two contain positive values. The two intervals with positive (negative) data have been separated by the median values of all positive (negative) values in the map.

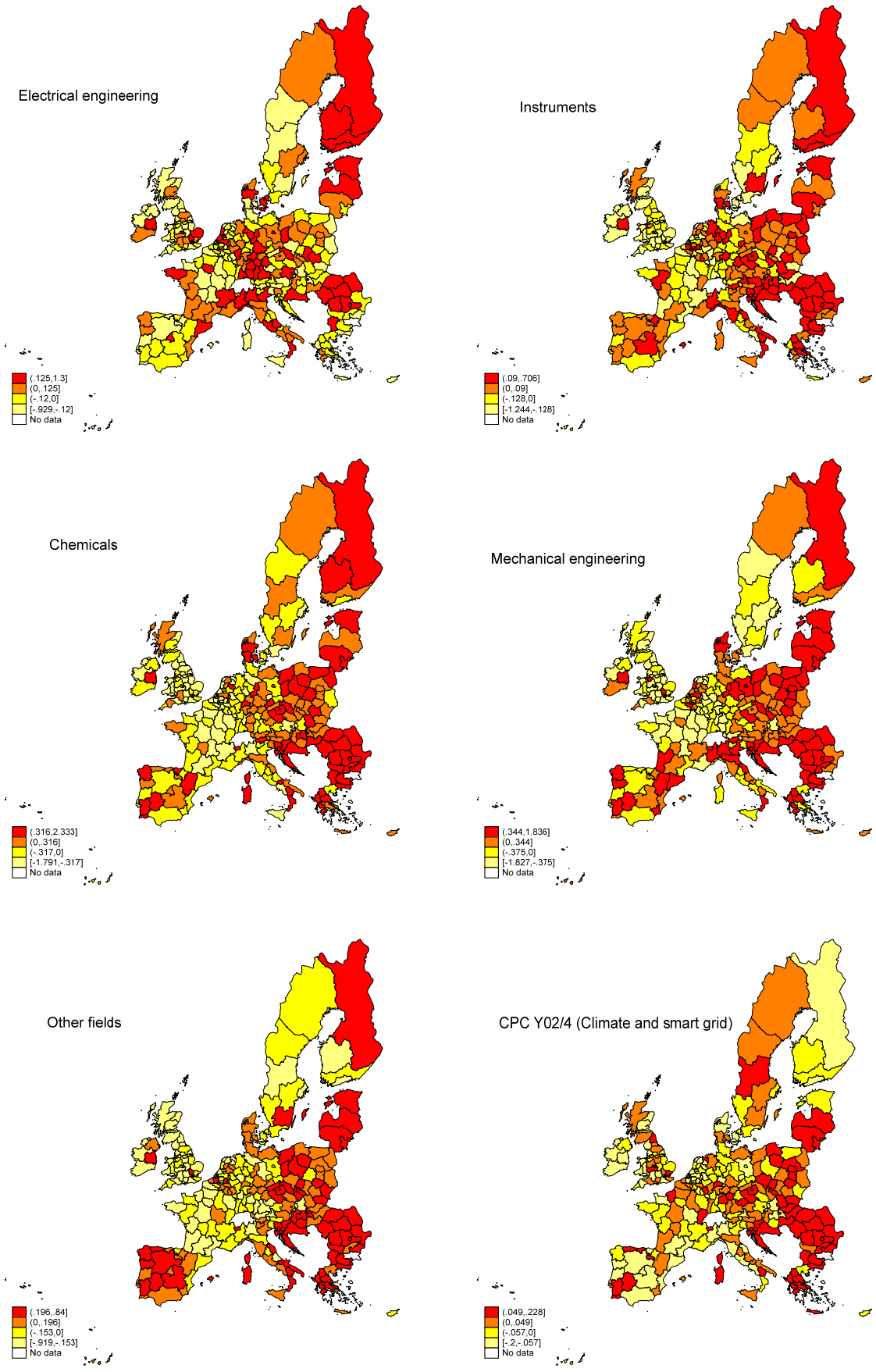


Figure A1. Predictions for regional disparities of the growth of GDP per capita, by six technology fields

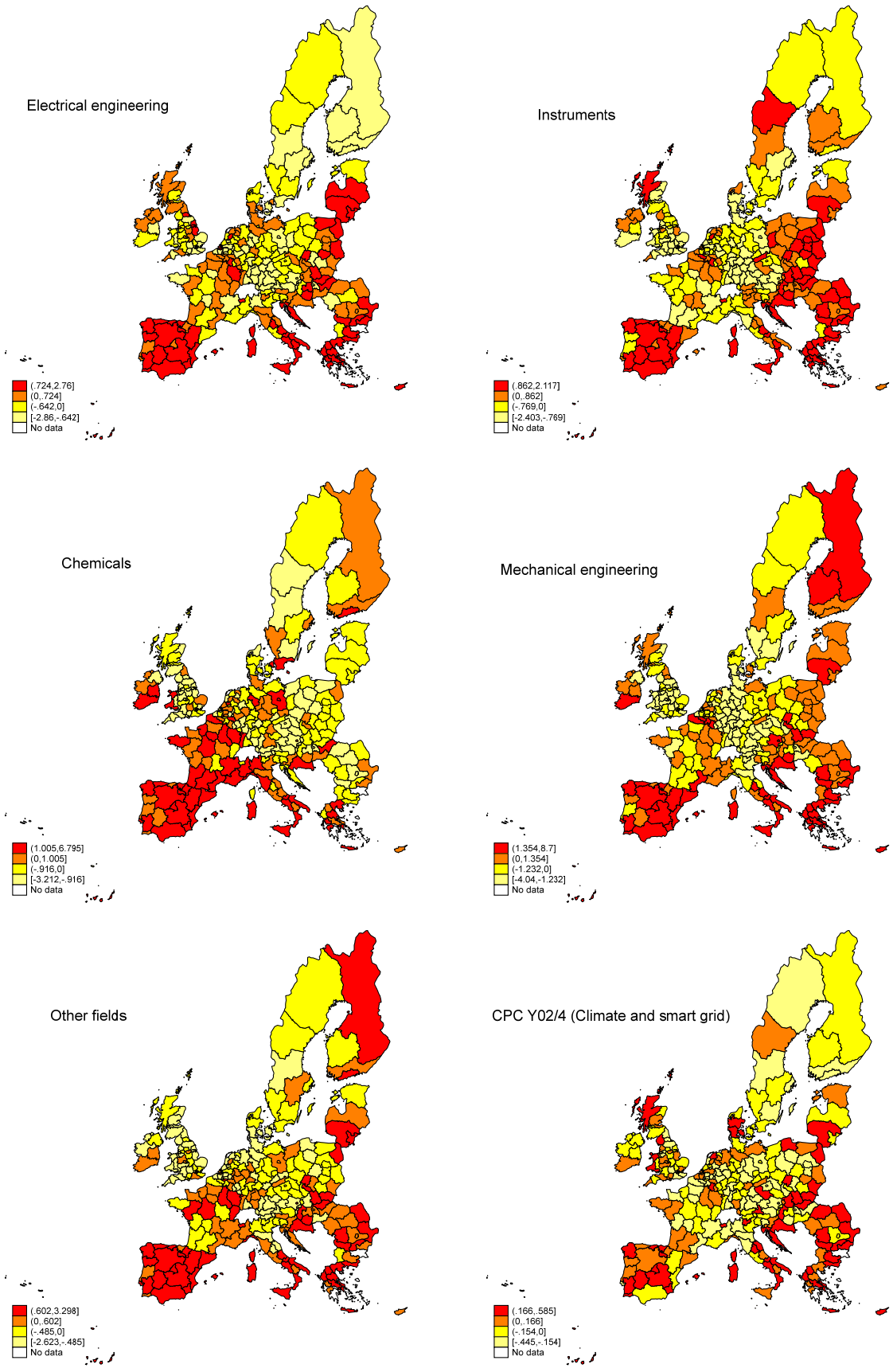


Figure A2. Predictions for regional disparities of the unemployment rate, by six technology fields

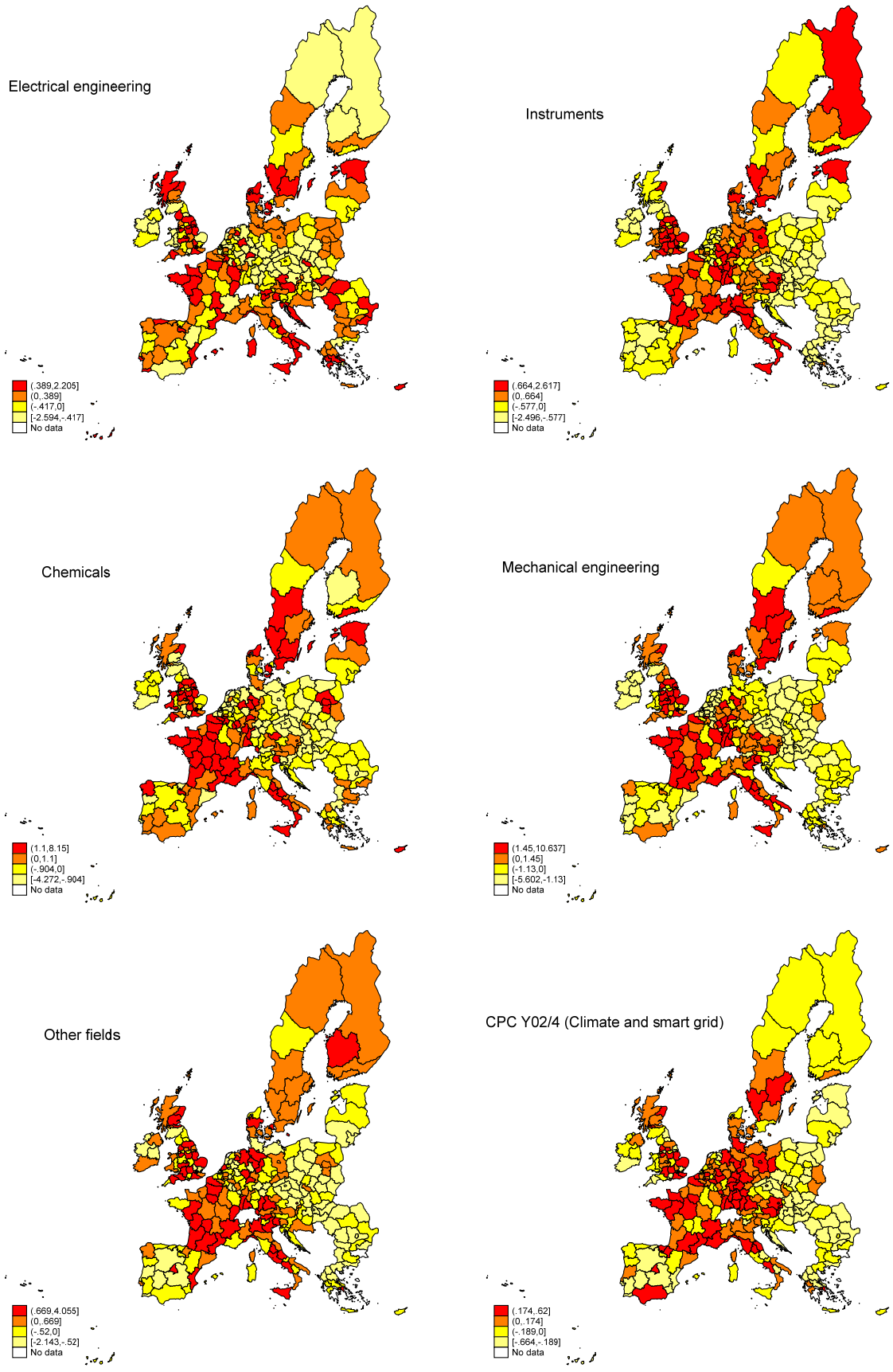


Figure A3. Predictions for regional disparities of growth of the unemployment rate, by six technology fields

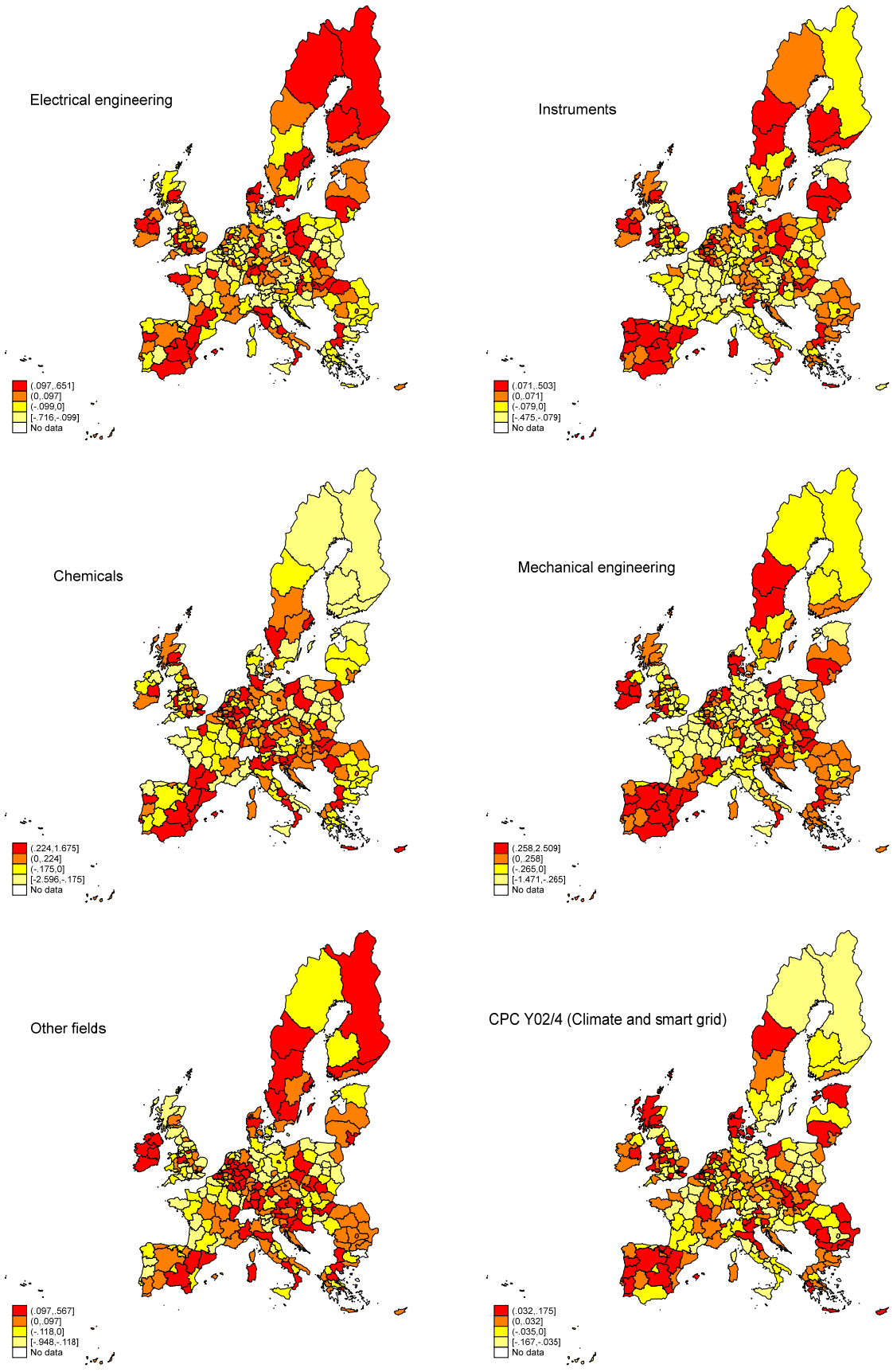


Figure A4. Predictions for regional disparities of the growth of employment, by six technology fields

The UNU-MERIT WORKING Paper Series

2022-01 *Structural transformations and cumulative causation towards an evolutionary micro-foundation of the Kaldorian growth model* by André Lorentz, Tommaso Ciarli, Maria Savona and Marco Valente
2022-02 *Estimation of a production function with domestic and foreign capital stock* by Thomas Ziesemer
2022-03 *Automation and related technologies: A mapping of the new knowledge* base by Enrico Santarelli, Jacopo Staccioli and Marco Vivarelli
2022-04 *The old-age pension household replacement rate in Belgium* by Alessio J.G. Brown and Anne-Lore Fraikin
2022-05 *Globalisation increased trust in northern and western Europe between 2002 and 2018* by Loesje Verhoeven and Jo Ritzen
2022-06 *Globalisation and financialisation in the Netherlands, 1995 – 2020* by Joan Muysken and Huub Meijers
2022-07 *Import penetration and manufacturing employment: Evidence from Africa* by Solomon Owusu, Gideon Ndubuisi and Emmanuel B. Mensah
2022-08 *Advanced digital technologies and industrial resilience during the COVID-19 pandemic: A firm-level perspective* by Elisa Calza Alejandro Lavopa and Ligia Zagato
2022-09 *The reckoning of sexual violence and corruption: A gendered study of sextortion in migration to South Africa* by Ashleigh Bicker Caarten, Loes van Heugten and Ortrun Merkle
2022-10 *The productive role of social policy* by Omar Rodríguez Torres
2022-11 *Some new views on product space and related diversification* by Önder Nomaler and Bart Verspagen
2022-12 *The multidimensional impacts of the Conditional Cash Transfer program Juntos in Peru* by Ricardo Morel and Liz Girón
2022-13 *Semi-endogenous growth in a non-Walrasian DSEM for Brazil: Estimation and simulation of changes in foreign income, human capital, R&D, and terms of trade* by Thomas H.W.Ziesemer
2022-14 *Routine-biased technological change and employee outcomes after mass layoffs: Evidence from Brazil* by Antonio Martins-Neto, Xavier Cirera and Alex Coad
2022-15 *The canonical correlation complexity method* by Önder Nomaler & Bart Verspagen
2022-16 *Canonical correlation complexity of European regions* by Önder Nomaler & Bart Verspagen