RUHR

ECONOMIC PAPERS

Imke Rhoden
Daniel Weller
Ann-Katrin Voit

# Spatio-temporal Dynamics of European Innovation - An Exploratory Approach via Multivariate Functional Data Cluster Analysis

RUB

#926

# Imprint


## Ruhr Economic Papers

Published by

RWI – Leibniz-Institut für Wirtschaftsforschung
Hohenzollernstr. 1-3, 45128 Essen, Germany

Ruhr-Universität Bochum (RUB), Department of Economics
Universitätsstr. 150, 44801 Bochum, Germany

Technische Universität Dortmund, Department of Economic and Social Sciences
Vogelpothsweg 87, 44227 Dortmund, Germany

Universität Duisburg-Essen, Department of Economics
Universitätsstr. 12, 45117 Essen, Germany





## Ruhr Economic Papers #926

Responsible Editor: Thomas Bauer





ISSN 1864-4872 (online) – ISBN 978-3-96973-084-3

Ruhr Economic Papers #926

Imke Rhoden, Daniel Weller, and Ann-Katrin Voit

# Spatio-temporal Dynamics of European Innovation - An Exploratory Approach via Multivariate Functional Data Cluster Analysis

RUB

## Bibliografische Informationen der Deutschen Nationalbibliothek

The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available on the Internet at **http://dnb.dnb.de**

RWI is funded by the Federal Government and the federal state of North Rhine-Westphalia.

http://dx.doi.org/10.4419/96973084
ISSN 1864-4872 (online)
ISBN 978-3-96973-084-3

Imke Rhoden, Daniel Weller, and Ann-Katrin Voit[1]

# Spatio-temporal Dynamics of European Innovation - An Exploratory Approach via Multivariate Functional Data Cluster Analysis

## Abstract

*We apply a functional data approach for mixture model-based multivariate innovation clustering to identify different regional innovation portfolios in Europe. Innovation concentration is considered as pattern of specialization among innovation types. We examine patent registration data and combine them with other innovation and economic data across 225 regions, 13 years and 8 patent classes. This allows us to identify innovation clusters that are supported by several innovation- and economy-related variables. We are able to form several regional clusters according to their specific innovation types. The regional innovation cluster solutions for IPC classes for 'fixed constructions' and 'mechanical engineering' are very comparable, and relatively less comparable for 'chemistry and metallurgy'. The clusters for innovations in 'physics' and 'chemistry and metallurgy' are similar; innovations in 'electricity' and 'physics' show similar temporal dynamics. For all other innovation types, the regional clustering is different and innovation concentrations in the respective regions are unique within clusters. By taking regional profiles, strengths and developments into account, options for improved efficiency of location-based regional innovation policy in order to promote tailored and efficient innovation-promoting programs can be derived.*

*JEL-Code: O33, R12, C38*

*Keywords: Functional Data Analysis (FDA); innovation concentration; spatio-temporal cluster modeling; multivariate cluster analysis; European innovation; cluster algorithm*

*November 2021*

*1 Imke Rhoden, Institute of Energy and Climate Research; Daniel Weller, RUB; Ann-Katrin Voit, FOM. – The Idea and conception of the paper was developed by Imke Rhoden and Daniel Weller. Ann-Katrin Voit provided section 2.1. Daniel Weller developed and conducted the statistical analyses of section 3. All authors contributed to the text, editing, and literature research. - All correspondence to: Imke Rhoden, Institute of Energy and Climate Research – Systems Analysis and Technology Evaluation, Forschungszentrum Jülich, Wilhelm-Johnen-Str., 52428 Jülich, Germany, e-mail: i.rhoden@fz-juelich.de*

# 1 Introduction

Innovation is a key driver of Europe's sustainable economic success. There are various national and supranational approaches for delineating innovation profiles and analyzing the European innovation landscape on several levels, the most renowned being the European Innovation Scoreboard (EIS) and its regional equivalent (RIS). Ranking regions according to their innovation strength is important for identifying and analyzing the characteristics of innovation leaders, so conditions in lagging regions conditions can be improved, e.g. via regional or innovation policy (European Commission, 2021b, 2021c).

As innovation is a highly complex matter, it is crucial to focus on regional innovation profiles and align policy programs with regional characteristics, as there are several and significant difficulties in targeting an increase in innovation activity and possibly resulting economic growth. Furthermore, there is a need to look below national levels and investigate regional strengths and weaknesses for a more efficient adaptation of policy mixes better, as policies are often not able to address regional needs (Izsák et al., 2013). Insights into specific branches of innovation via patent analysis, supported by the inclusion of further knowledge indicators and regional characteristics, can provide levers to improving policies, thus harnessing not only innovation potentials but also regional potentials from a European cohesion policy (European Commission, 2021a). Thorough investigations and a precise understanding of the different types of innovation, their place of inception and their evolution over time are crucial for aiding Europe's path to a sustainable economic future.

In this paper, we apply a mixture model-based clustering analysis for multivariate functional data proposed by Schmutz et al. (2021) to explore the spatio-temporal dynamics of European regional innovation activities and uncover groups of regions with homogeneous innovation profiles. To achieve this, the analysis is based on the functional data analysis paradigm (FDA), which allows us to analyze latent functional forms, inherent dynamics and other features in time series of multiple innovation indicators too subtle to be captured by classical time series or clustering approaches. As innovation is a heterogenous phenomenon itself, we use several time series for main patent classes as proxies for innovation activity as well as other closely related indicators to generate individual innovation profiles for Europe's regions.

The paper is structured as follows: In the second section, a theoretical overview of relevant innovation literature with respect to general principles and related approaches for identifying innovation clusters is given. The third section provides a description of the time series data used for the statistical analysis, the general principles of the functional data paradigm and the mixture model-based multivariate functional clustering algorithm applied. A presentation of the clustering results is given in the fourth section, before the paper concludes with a discussion in the fifth section.

# 2 Theory

## 2.1 General innovation theory

In economics, the spatial dimension has played an increasingly important role since the beginning of the 1990s and the publication of 'Geography and Trade' (Krugman, 1991) widened the economic view for a better understanding of the global economy through its spatial dimension.

While some countries may experience lower growth or investment rates, other countries may suffer from higher unemployment rates. Krugman linked it globally and explained that global competition leads to more challenges that need to be considered. In order to have a sufficient number of qualified

jobs for their population, countries have to consider their advantages or disadvantages due to location and innovation. In modern economies, more companies are forced to export their products and are therefore subject to a higher level of international competitiveness. To manage this successfully, companies likely settle into industry clusters and may consider relocating to gain from location advantages.

Krugman also shows the crucial role of innovation in economies by saying '[t]he more you know, the more you can learn' (Krugman, 1979, p. 259). He pointed out that countries have the obligation to find regional strengths and weaknesses to ensure their success, which depends to a large extend on the development of innovation-promoting structures.

Accordingly, spatial factors and innovation activities are closely connected as e.g. seen in the Silicon Valley area in California, USA (Sturgeon, 2003). This leads to the following assumptions (Koschatzky, 2001):

- Regional factors influence the operational innovation process
- Innovation processes have a regional origin
- Innovation processes are spatially differentiated
- Spatial proximity promotes innovation-relevant interactions
- Regional innovation and technology policy support measures are effective

Following Christaller (1933), Lösch (1940) and their theory of central places for industrial activity, the clustering of companies is based on their spatial interdependencies. Additionally, Thünen (1966) described the minimization of transportation costs as an advantage of central locations and regional vicinity.

With regard to the innovation-related spatial factors, Krugman also considers transport costs to have an effect on regional growth rates (Koschatzky, 2001) as producers and consumers make spatial decisions based on prices and revenues to optimize their profits. Consequently, producers will try to increase profits by minimizing the costs of transportation, which can lead to a relocation of business. This logic applies to the end users as well, as they will relocate their demand to the regions with the lowest transport costs. As companies are more likely produce most efficiently where their required production factors are sufficiently available, they have a latent incentive to locate in their customer's geographical vicinity. This will result in low transport costs for both parties, the supply side and the demand side. However, it has to be considered that transport costs have been significantly reduced by transportation technology and have consequently become less important in many industries in recent years (Fernau, 1997).

In terms of spatial clustering, agglomeration effects are of crucial relevance as several advantages as well as disadvantages can arise from localized concentrations of companies. To reach optimal levels of competitiveness within clusters, both internal and external agglomeration effects have to be differentiated and considered. Internal agglomeration effects are also known as economies of scale and provide advantages reducing fixed costs by the production of larger quantities of goods, while external agglomeration effects refer to the proximity of companies in the same value chain. As clearly shown by (Marshall, 1920), companies with similar activities can profit from sharing access to skilled labor-by-labor pooling, sharing inputs from common suppliers, and benefiting from knowledge spillovers. Thus, companies can maximize their profits by shared use of workers, infrastructure, services as well as information. If a company has access to all their key resources in the vicinity, they can experience a competitive advantage (Sraffa & Dobb, 1951). Potential disadvantages of agglomeration effects can arise in form of higher environmental pollution (Breuste & Keidel, 2008), increasing property prices, higher competition in the local area, or overstrained infrastructure.

Further effects of agglomeration are localization and urbanization (Blume, 2004), where localization effects can be described as advantages arising from a company's proximity to other companies of its industry. Those advantages can be e.g. an industry-relevant job market in the area, research and development facilities and therefore patenting activity or the emergence of a specialized supply industry (Alcácer & Chung, 2014). Urbanization effects develop over time as different industries lead to more infrastructure as well as urbanization of affected areas and therefore to an increase in economic activity in general (Blume, 2004).

As some regions in Europe are highly successful and others are lacking behind, theories of innovation can provide means to identify and understand the disparities in terms of employment, infrastructure, availability of services and economic success in general. Often disparities arise due to the unequal distribution of natural resources, decisions of the public sector or other locational reasons. Within metropolitan areas, disparities can be seen between central and peripheral areas with peripheral areas increasing faster in value than the centers. Therefore, migration and relocations from centers to suburban areas can be seen, which can potentially lead to devaluation of certain areas and neighborhoods, thus creating social inequalities (Pflüger, 2019).

In the European Union, regional development is a key factor for equal living conditions in and between the member states. Therefore, the EU targets economic development via spatial planning, state planning and regional planning. Particularly the European Regional Development Fund (ERDF) and the Trans-European Networks (TEN) are means to detect national and regional disparities and to support the realization of equal living conditions. The newest European funding programs and a key focus of the European Commission is the European Green Deal, which aims to fight environmental degradation and climate change, while simultaneously searching for new and sustainable growth strategies to be competitive in the future. The European Green Deal concentrates on (European Commission, 2019):

- Investing in environmentally friendly technologies
- Supporting industries to innovate
- Introducing cleaner, cheaper and healthier forms of private and public transport
- Decarbonizing energy production
- Ensuring higher energy efficiency of buildings
- Improving global environmental standards via international cooperation

To achieve these goals, the EU must ensure a high level of labor skills as well as high levels of investment in research and development. Given this focus, it is vital to gain a precise understanding of the spatio-temporal dynamics within the European innovation system. The potential of innovation for mitigating climate change through new, efficient technologies that promote sustainable growth can help to avoid a lock-in and make innovation "green innovation" (Aghion et al., 2009). This applies not only to identifying the types and drivers of Member States' innovation strengths, but also to investigating locational differences in innovation. More knowledge about the structure of innovation and its place of inception can be used to understand innovation emergence as well as its inherent geographical nature, provide insights to the success of policy programs and furthermore help structure future policy programs for a sustainable innovation climate in the EU.

## 2.2 State of the Art

The topics of innovation, geography, clustering as well as their interdependencies can be investigated by a variety of approaches. If innovation is considered in the context with geography and economic growth, there is no single theoretical framework, as there are too many interlinkages between these topics find a universal approach (Acs & Varga, 2002). Thus, there are multiple schools of thought

regarding the temporal and spatial evolution of innovative activity (Rhoden, 2019). In this paper, we focus on approaches related to investigating innovation clustering like Fornahl and Brenner (2009), who find that differing types of innovation cluster differently, which points to the relevance of considering innovation as a differentiated subject.

Knowledge spillovers are another link between innovation and geography to consider, as knowledge (tacit or understanding) is often only transferred locally or regionally. Innovation is thus prone to spillover, as research shows (e.g. McCann and Simonen (2005), Constantini et al. (2013) and Aldieri et al. (2019)). This is confirmed by Bottazzi and Peri (2003), who correlate data on R&D and patents and find that R&D spending can increase innovation output, but only limited to a local scale. Giannitsis and Kager (2009) analyze links between technology and specialization as can determine market positions and competitive success. Thus, it is vital to know how static and dynamic conditions interact and how they contribute to emergence of innovation. They note that it is important for policy to adapt effectively and timely to changing circumstances, as technology specification can drive industry and thus competitive advantages. Here, policy can promote progress through innovation.

Capello and Lenzi (2013) search for patterns of knowledge, attitudes, and innovation behaviors in innovative European regions by means of a cluster analysis. They cluster the degree of knowledge and innovation that the selected regions produce, taking into account their different stages of the innovation process. Above all, the results indicate that policy measures at a regional level are useful and necessary, as innovation trajectories diverge due to regional characteristics. Innovation, the authors propose, is much more complex than just the divide between agglomerated and peripheral regions. Moreover, they suggest policies that are closely oriented towards the respective clusters and their specific innovation patterns, leading to a "smart" Europe.

Spielkamp and Vopel (1999) explicitly combine innovation systems and cluster theory to find innovation clusters in Germany. They assume the existence of agents in technological environment networks to create, use, and diffuse technology and apply this view together with further innovation variables so that a system of innovation and firms emerges that leads to certain patterns. Furthermore, they emphasize that due the extremely high complexity of innovation systems, a multitude of approaches are possible. Several variables are used in their clustering approach, the most important of which are innovation, knowledge, information, and industry characteristics.

Common among these approaches is the fact that innovation should not be considered without a spatial component, nor without a temporal component. Turkina and Van Assche (2018) examine innovation performance in clusters and find that linkages along the horizontal and vertical supply chain are key to increasing knowledge intensity and thus innovation. Pełka (2018) analyzes innovation clusters using symbolic density-based ensemble clustering, taking into account innovation policy. They investigate European countries and use the Regional Innovation Scoreboard as well as multiple innovation and other indicators. They calculate clusters with standard methodology (e.g. k-means) and investigate the heterogeneity of the clusters. The result is a ranking of innovation leadership.

Ionela-Andreea and Marian (2020) use data from the European Patent Office and calculate the Malmquist index for total factor productivity in knowledge performance (Caves et al., 1982). They also identify differences and similarities in the development of innovation capacities between the resulting clusters. Zabala-Iturriagagoitia et al. (2021) investigate the increasing territorial disparities in Europe using production theory and also apply the Malmquist index. They note that advances in innovation are not necessarily synonymous with technological progress and that there is no innovation convergence where lagging regions can catch up with leading regions.

Pelau and Chinie (2018) conduct a cluster analysis of European regions in relation to innovation and sustainable development, linking innovation and sustainability for an improved economic growth process. They use a static multivariate analysis to characterize regional clusters and find three major innovation-sustainability clusters ranked by degree of achievement. They also relate their approach to the literature on innovation systems, emphasizing the importance of the regional context of innovation. Kim and Bae (2017) apply clustering as a step in forecasting potentially promising technology. Based on the information contained in classified patents, which can give an indication of the technologies in development, they find technology-specific clusters. Their aim is then to derive potential trends for developing technologies based on the clusters.

A functional data analysis for multivariate innovation clusters taking into account different innovation types, multiple measures as well as temporal and spatial dimensions, which can furthermore explore innovation profiles, has not been conducted previously. We will illustrate the procedure in the following chapters.

## 2.3 Data origin and derivation of approach

In practice, several indicators can be used for approximating innovation, but we choose to mainly use patent data to indicate the type of innovation, with the patent classification scheme allowing a distinction between different types of inventions. The classification is based on the type of innovation group to which a patent belongs to and must be indicated when filing application. The patent classes we use are the eight major classes (A: *'human necessities'*; B: *'performing operations and transporting'*; C: *'chemistry and metallurgy'*; D: *'textiles and paper'*; E: '*fixed constructions'*; F: '*mechanical engineering, lighting, heating, weapons, and blasting'*; G: *'physics'*; H: *'electricity'* (WIPO, 2021)). As suggested by Griliches (1990) and noted by several other researchers, the inclusion of patents as an indicator of innovation is justified by the intentions pursued by filing a patent, i.e. an intended commercial use. Other innovation indicators are R&D personnel and researchers as well as internal R&D expenditures as percentage of the gross domestic product. In addition, we use a human capital indicator approximated by human resources in science and technology. These variables are suitable to support patents as indicator of innovation, as they are directly related to the emergence of innovation and can lead to patents or other forms of innovation.

In the model, variants of these variables are used. First, we compute Innovation Gini indicators according to Rhoden (2020) for each IPC class, which provide a measure of the degree of innovation variation in regions. Then, we calculate the labor density and relate it to the regional GDP, the share of R&D labor, the human capital density and the R&D expenditure per R&D labor. These measures are used to indicate labor productivity, human capital accumulation per worker and the R&D expenditure productivity. This step results in a set of five covariates that are included in the clustering process of the Innovation Ginis (see Table 1). In this way, multivariate spatio-temporal innovation dynamics of European regions can be aggregated into eight sets of clusters showing similarities and differences of the regional structures for each of the eight patent classes.

| Variable | Description | Eurostat datasets used |
|---|---|---|
| Innovation Gini | Innovation Gini for the relevant IPC class of patents (patent applications to the EPO by priority year); Normalization Factor: Economically active population in 1000 | PAT_EP_RIPC<br>LFST_R_LFP2ACT |
| Labor Density | Economically active population per square kilometer in 1000 | LFST_R_LFP2ACT<br>DEMO_R_D3AREA |
| Share of R&D Labor | R&D personnel and researchers directly engaged in R&D per economically active population in 1000 | RD_P_PERSREG<br>LFST_R_LFP2ACT |
| GDP per Labor | Gross Domestic Product at current market prices in Billion Euro per economically active population in 1000 | NAMA_10R_3GDP<br>LFST_R_LFP2ACT |
| Human Capital Density | Human resources in science and technology (Persons with education in science and technology) per economically active population in 1000 | HRST_ST_RCAT<br>LFST_R_LFP2ACT |
| R&D Investment per R&D Labor | Internal R&D investment in Billion Euro per R&D personnel and researchers directly engaged in R&D in 1000 | RD_E_GERDREG<br>RD_P_PERSREG |

Table 1: Variable Declaration, source: Own calculations (Eurostat, 2021; OECD, 2021; Office for National Statistics, 2021).

As our analysis focuses on European regions, our dataset consists mainly of data from Eurostat (2021) for the period of 2000 to 2012, with supplements from other statistical offices and organizations (i.e. (OECD, 2021; Office for National Statistics, 2021) used for filling missing values in the main datasets after checking for plausibility. However, there are still larger numbers of remaining missing values, which we choose to impute via natural spline interpolation using the annual cross-sections of our datasets as knots (Eubank, 1999; Simonoff, 2012). This imputation is applied when less than 30 percent of values for a region is missing and the pattern of missingness can be reasonably handled by spline interpolation, i.e. when there are enough values next to the missing values. Although this may seem like an arbitrary choice, sensitivity analyses have shown that this procedure strikes a more robust balance between the highest number of regions to cluster and the least amount of imputation bias compared to other approaches (e.g. Honaker and King (2010)).

Spatially, we focus on the European regions at NUTS-2 level, which necessitates the creation of a custom reference, as several revisions of the NUTS classification were made over the periods covered by our data. This reference is based on NUTS 2016, which corresponds to most of our data but adopts NUTS 2010 regions where later revisions differ from the regions in our dataset. We also create a custom shapefile to correctly represent the statistical geographical level, which we then apply throughout our calculations. In total, we use 225 distinct regions in our mixture model-based multivariate functional cluster analysis. All computations are realized in the software R (R Core Team, 2021)using the packages *fda* (Ramsay et al., 2021)and *funHDDC* (Schmutz et al., 2021).

# 3 Functional data paradigm

Although the concept of functional data dates back to Grenander (1950) and Rao (1958), the actual term *functional data* for objects that can naturally be viewed as smooth curves rather than a set of discrete observations was coined by Ramsay (1982), Ramsay and Dalzell (1991) and Rice and Silverman

(1991). In statistical terms, *functional data* are random variables usually observed at multiple discrete points on an infinite dimensional or functional continuum such as time, space or other variables describing continua (Ferraty & Vieu, 2006). Accordingly, a set of functional variables for multiple observations is called *functional dataset*. In line with Kokoszka and Reimherr (2017), we refer to functional data as

$$X_n(t_{n,p}) \in \mathbb{R}^p;\ t_{n,p} \in [T_{min}, T_{max}];\ n = 1, \dots, N;\ p = 1, \dots, P.$$

In this notation, functional data are given by a set of N independent curves $X_n$ observed in discrete sets of values $\{t_{n,p}, y_{n,p}\}$ along an interval $[T_{min}, T_{max}]$ over potentially infinite dimensions P. Functional data analysis can thus be performed not only with random curves, but also p-dimensional random surfaces. In most fields of research, however, the focus is still on the analysis of curves, which is why the term *curve data* (Gasser & Kneip, 1995; Gasser et al., 1984; Rice & Silverman, 1991) is often used for analysis of the special case of a one-dimensional continuum. A comprehensive review of the history of functional data analysis, its methods and applications in different fields of research is given by Wang et al. (2015).

In general, functional data are considered as independent and identically distributed samples from $L_2$-continuous stochastic processes whose mean and covariance estimators are given by $\hat{\mu}(t_p) = \frac{1}{n}\sum_{i=1}^{n} x_i(t_{n,p})$ and $\hat{v}(t_p) = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i(s_{n,p}) - \hat{\mu}(s_{n,p})\right)\left(x_i(t_{n,p}) - \hat{\mu}(t_{n,p})\right)$. As Deville (1974) has shown, both estimators converge to $\mu(t_p)$ and $v(s_p, t_p)$ in $L_2$-norm, which is consistent with the assumption of a latent functional form in the form of smooth curves rather than mere sequences of observations as basic principle of functional data analysis (Ramsay & Silverman, 1997).

As crucial smoothness may be for the analysis of functional data, it may not be obvious in raw datasets as observations are often contaminated or distorted by random noise, measurement errors or other types of bias (Ramsay & Silverman, 1997). These effects can be viewed as fluctuations in the smooth curves that we include by extending our earlier notion of functional data:

$$S_n(t_{n,p}) = X_n(t_{n,p}) + \epsilon_{n,p},$$

where $S_n(t_{n,p})$ is the realized and observable functional form and $\epsilon_{n,p}$ the representation of noise, disturbance or error. We would like to refer to Ferraty and Vieu (2006), Ramsay and Silverman (1997) and Kokoszka and Reimherr (2017) for a complete overview of the theoretical foundations of functional data analysis.

As our imputed data are still in their raw form, we use basis expansions to reconstruct their functional forms, which is necessary for any kind of functional data analysis (Aguilera et al., 2010). Ideally, this basis function is similar in shape and form to the observed functions, as the curves can then be easily approximated by a linear combination of the chosen basis function (Kokoszka & Reimherr, 2017). As there is no clear rule for choosing the most efficient shape and number of basis functions with respect to multivariate functional clustering (Jacques & Preda, 2014a), we follow the suggestions of Schmutz et al. (2020) and choose a set of B-spline functions whose size corresponds to the number of years for every variable, while applying a small roughness parameter to reduce potential biases due to our earlier spline imputation.

## 3.1 Multivariate functional clustering

Cluster analyses are used to find homogeneous groups of observations in datasets without prior knowledge of latent group relationships, which can be achieved with a wide variety of algorithms that have been proposed for clustering of functional data. However, due to the potentially infinite-dimensional nature of functional data, several issues arise that are of lesser importance for classical cluster

analyses, such as the reduction of functional dimensionality, which needs to be solved. To address these issues, several methodological approaches for clustering functional data have recently been published, ranging from the simple transfers of classical algorithms to the functional domain to complex model-based clustering after applying of statistical filtering (see Jacques and Preda (2014a) for a review).

However, most of these approaches focus on clustering univariate functional data (see e.g. Abraham et al. (2003); Bongiorno and Goia (2016); Bouveyron et al. (2015); Bouveyron and Jacques (2011); Chiou and Li (2007); Coffey et al. (2014); Jacques and Preda (2013); James and Sugar (2003); Li and Chiou (2011); Peng and Müller (2008); Serban and Wasserman (2005)), while there are still only few concepts dedicated to multivariate functional clustering. Among those concepts, model-based approaches have received more attention in recent years, as they have proven to be suitable for complex statistical structures and relationships (see e.g.Bouveyron and Jacques (2011); Ieva and Paganoni (2016); Jacques and Preda (2014b); Kayano et al. (2010); Schmutz et al. (2020); Traore et al. (2019)).

In our cluster analysis, we follow the mixture model-based approach proposed by Schmutz et al. (2020) to cluster multivariate functional data of regional innovation activities to investigate spatio-temporal similarities and differences in the European innovation system. This approach builds on previous work by Bouveyron and Jacques (2011) and Jacques and Preda (2014b) by circumventing the *curse of dimensionality* (Bellman, 1957) with a multivariate functional principal component analysis (MFPCA) and considers the analytical scores to be random variables with cluster-specific probability distributions. By reprojecting the previously infinite- onto a finite-dimensional problem, the cluster-specific probability distributions can then be approximated via expectation maximization (EM) (Dempster et al., 1977), which makes this approach highly flexible as additional assumptions can easily be imposed on the model.

## 3.2 Multivariate functional principal component analysis

The use of principal component analysis for functional data as a means for dimensionality reduction was already proposed by Ramsay and Silverman (1997). Multivariate functional data require more adaptive approaches, as shown by Jacques and Preda (2014b) and Schmutz et al. (2020). Specifically, MFPCA aims to find the eigenvalues and eigenfunctions to solve the decomposition equation of the covariance operator

$$\nu f_j = \lambda_l f_j$$

where $\lambda_j$ is a finite group of $j$ positive eigenvalues, *principal scores*, and $f_j$ is a group of corresponding multivariate eigenfunctions, *principal factors*. Following Schmutz et al. (2020), we assume that the latter are part of a linear space spanned by a matrix $\phi$:

$$f_j(t) = \phi(t) b_j'$$

Consequently, we can reformulate the eigenproblem using the covariance estimator

$$\hat{v}(s,t) = \frac{1}{n-1} \boldsymbol{\phi}(s) \boldsymbol{C}' \boldsymbol{C} \boldsymbol{\phi}'(t)$$

which leads to

$$\frac{1}{n-1} \phi(s) \boldsymbol{C}' \boldsymbol{C} \boldsymbol{W} \boldsymbol{b}_l' = \lambda_l \boldsymbol{\phi}(t) \boldsymbol{b}_l'$$

where $\boldsymbol{W} = \int_0^T \boldsymbol{\phi}'(t) \boldsymbol{\phi}(t)$ is a $R \times R$-Matrix containing the inner product of our basis functions.

The principal component analysis is then reduced to an eigenvalue decomposition of the matrix

$$\frac{1}{\sqrt{n-1}}\boldsymbol{CW}^{1/2}$$

allowing each multivariate curve $S_n(t_{n,p})$ to be identified by its scores $\delta_i = (\delta_{i_j})$ into the basis of multivariate eigenfunctions $(f_j)$ for $j \geq 1$ (see Jacques and Preda (2014b) and Schmutz et al. (2020) for proofs).

## 3.3 Mixture model-based clustering of multivariate functional data

Model-based clustering assumes that population data are a mixture of groups, so that the elements of this mixture can be modeled by their conditional probability distribution. Therefore, the latent finite mixture model for the approach by Schmutz et al. (2020) can be formulated as

$$g(s) = \sum_{k=1}^{K} \pi_k f_k(s_n)$$

where $g(s)$ is the probability density function of $s$, the mixture proportion of the $k$-th cluster is given by $\pi_k$ with $\sum_{k=1}^{k} \pi_k = 1$ and $f_k(s_n)$ being the conditional density function. However, a feature of functional random variables is the lack of general notion of probability density functions (Delaigle & Hall, 2010), which necessitates the use of a parametric approximation:

$$g(s) = \sum_{k=1}^{K} \pi_k f_k(s_n;\, \theta_k)$$

with $\theta_k$ being the parameter vector of the $k$-th mixture element. Given this approximation, the likelihood of the mixture model proposed by Schmutz et al. (2020) is then given by

$$l(\theta; s; z) = \sum_{n=1}^{N}\sum_{k=1}^{K} z_{kn} \log\left(\pi_k f(s_n; \theta_k)\right)$$

where $z_{kn}$ is a latent group variable equal to 1 if multivariate curves belong to cluster $k$ or 0 otherwise. Finally, we can obtain a fully parameterized form of the likelihood by including the gaussian density function $f(s_n; \theta_k))$ (see Schmutz et al. (2020) for proofs):

$$\begin{aligned} l(\theta; s; z) = -\frac{1}{2}\sum_{k=1}^{K} n_k \Bigg[ & -2\log(\pi_k) \\ & + \sum_{j=1}^{d_k} \log(a_{kj}) + \sum_{j=d_k+1}^{R} \log(b_k) + \sum_{j=1}^{d_k} \frac{q_{kj}^t W^{1/2} C_k W^{1/2} q_{kj}}{a_{kj}} + \sum_{j=d_k+1}^{R} \frac{q_{kj}^t W^{1/2} C_k W^{1/2} q_{kj}}{b_k} \Bigg] \\ & + \frac{nR}{2}\log(2\pi) \end{aligned}$$

where $a_{kj}$ and $b_k$ are a direct result of the MFPCA, since it is assumed that the scores of the $n_k$ curves of the $k$-th cluster ${\delta_n}^k$ follow a Gaussian distribution with mean function $\mu_k \in \mathbb{R}$ and a covariance matrix $\Delta_k$. The latter is crucial for both parameters as they are diagonal matrix elements:

$$\Delta_k = \begin{pmatrix} \begin{bmatrix} a_{k1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & a_{kd_k} \end{bmatrix} & 0 \\ 0 & \begin{bmatrix} b_k & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & b_k \end{bmatrix} \end{pmatrix} \begin{matrix} \left\{ d_k \right. \\ \left\{ R - d_k \right. \end{matrix}$$

Due to this covariance matrix, the variance of the first $d_k$ principal components can be modeled much more accurately, while the other components can be retained and modeled via the parameter $b_k$, which provides a model with much higher degrees of clustering flexibility (Schmutz et al., 2020).

### 3.4 Model inference via expectation maximization

An expectation maximization algorithm (Dempster et al., 1977) is used to estimate the parameters of the complete likelihood given in the previous section, as this type of algorithm has been shown to be reliable and reproducible in maximizing the likelihood of model-based clustering approaches. The algorithm uses two stages to estimate the model parameters and constantly alternates from one to the other until an optimal solution is found (Schmutz et al., 2020).

In the expectation step, the conditional expectation of the log-likelihood is calculated using the most recent parameter estimates. Then, the maximization step updates these estimates by maximizing the expected log-likelihood conditionally. The process is stopped when the difference of two successive estimations is smaller than $10^{-6}$ or a limit of $200$ iterations has been reached. However, the algorithm must first be initialized by either providing initial values or using random values. We choose to initialize the clustering analysis by applying a k-means algorithm by Hartigan and Wong (1979) with four partitions to the discretized values of our functional dataset to obtain initial values for the functional partitions. Although Schmutz et al. (2020) suggest to use multiple initialization strategies to prevent convergence to a local maximum, we found this approach to result in nearly identical cluster solutions as random initialization.

To obtain optimal cluster solutions for each IPC class, a series of models covering all parameter constraints provided by Schmutz et al. (2020) is estimated for a range of 2 to 10 clusters. We retain the solutions with the lowest Bayesian information criterion (BIC) (Schwarz, 1978) as our final cluster solution. The BIC is defined by

$$BIC = l(\theta; s; z) - \frac{m}{2} * \log(n)$$

where $l(\theta; s; z)$ is the maximum log-likelihood value, the number of model parameters is given with $m$ and $n$ is the number of individuals, which allows the log-likelihood to be penalized by model complexity. This procedure is in line with proposals by Schmutz et al. (2020).

## 4 Results

In the following sections, the results of our mixture model-based cluster analysis are presented for the eight IPC classes with respect to the innovation indicators. As these cluster solutions are the results of multivariate functional dynamics, differentiation of the clusters is based on a simultaneous evaluation of all modeling variables, i.e. Innovation Gini coefficients, labor density, share of R&D labor, GDP per labor, human capital density, and R&D investment per R&D labor (see table 1). This ensures that subtle spatio-temporal regional dynamics in the modeled indicators are captured and regional disparities can be shown more clearly. To optimize the cluster solution, a range of models with various parameter constraints is used for up to 10 clusters, with the lowest BIC indicating the best cluster solution for a given set of variables. Accordingly, the number of clusters varies across the eight patent classes, but the size of the clusters is not limited, i.e. the numbers of regions per cluster only depends on regional similarities in innovation dynamics.

## 4.1 IPC Class A

The clustering process results in ten distinct clusters of spatio-temporal innovation dynamics for the patent group *'human necessities'* (IPC A, see Figure 1, 1st row, left panel). While clusters 1, 2, and 6 are relative small and limited to few regions spread over Central European countries, the clusters 4, 5 and 10 consist mostly of neighboring regions in, with few exceptions, large parts of Eastern Europe and the Baltic Area (cluster 4), and Portugal and Spain (cluster 5). In contrast, cluster 10 is significantly less spatially concentrated, containing most parts of France, but also regions in Italy, Austria, Germany or Finland. Another large cluster is found in Scandinavian regions and Iceland, with regions in the United Kingdom, Germany and Italy also assigned to this group of innovation concentration. East Germany has similar innovation dynamics as regions in northern Spain, northern Ireland and southern Italy (cluster 8). These regions are often characterized as structurally weak, which seems to be reflected in innovation concentration potentials. Most regions in Central Europe, mainly Germany, Luxembourg, Belgium and the Netherlands are highly diverse with neighboring regions not being part of the same innovation cluster.

Regarding the mean curve of the Innovation Gini (see Figure 1, 1st row, right panel), it is noticeable that cluster 4 drops significantly compared to the other clusters until 2008, before a slight increase to a stable trajectory sets in. In comparison, the mean curve of cluster 5 rises very sharply from 2004 onwards, reaching a higher level of concentration than any of the other clusters. Most other cluster mean curves either show a stale trajectory or increase slightly until 2008 before declining. Despite these temporal dynamics, there are clearly noticeable crossings of most cluster mean curves, with cluster 5 being an exception, i.e. the clusters only develop in a relatively narrow range overall in terms of innovation concentration, but evolve highly variable in this range. The mean curves of the variables for labor density, share of R&D labor, GDP per labor and human capital density from 2000 to 2012 show a relatively even and almost linear increase. In terms of human capital density, the clusters do not differ much. In principle, these results hold for the mean curves of all indicators, but the decreasing function of cluster 10 deviates from the other stable or slightly increasing cluster trajectories of the share of R&D labor. The mean curves for R&D investment per R&D labor are very differentiated. While the curves for clusters 6 and 10 initially increase, then decrease until 2005, cluster 4 and 5 show a stable level, which is, however, well below all other clusters.

## 4.2 IPC Class B

Ten clusters are found for the patent group *'performing operations and transporting'* (IPC B). Again, Central Europe is quite fragmented in terms of cluster memberships, as the regions in this part of Europe are assigned to clusters 1, 2, 5, 7 and 10 (see Figure 1, 2nd row, left panel). Other regions in cluster 5 are located in England, large parts of Norway, Iceland and Finland, but also in Austria, northern Italy and France, while cluster 8 is mainly located in Spain, Portugal and East Germany, which corresponds to the clustering previously observed in these regions for IPC class A. Eastern Europe is largely composed of regions in clusters 4 and 8, with slightly more variation than for class A. Ireland and Northern Ireland are divided into three different clusters (6, 7 and 8), while Norway is divided into two large clusters (3 and 5). Most Regions belonging to cluster 2 are located in southern Italy and the northeastern parts of Spain and France, as well as parts of the Netherlands and Germany.

Regarding the temporal dynamics of the cluster for this IPC class, there are significant differences in the mean curves for innovation concentration (see Figure 1, 2nd row, right panel). For example, clusters 4 and 9, both located mainly in Eastern Europe, differ strongly. While the mean curves for most clusters decrease over time, cluster 3 seems to be an exception, as regions in this cluster seem to slightly increase their degree of concentration. In general, cluster 10 is the most stable in terms of innovation concentration. However, the mean curves of most clusters have slightly decreased since 2005, which

corresponds to a decrease in innovation concentration. The other covariates, with exception of R&D investment per R&D labor, mostly show stable or slightly almost linearly increasing over time. The mean curves of all clusters are very close to each other and similar in terms human capital density, which again is consistent with the results for IPC class A. The mean curves for GDP per labor show wide variation in terms of the level, with only cluster 6 and its exclusive focus on the UK showing a slight decline from 2007 onwards, while the regions of cluster 3, which are mostly located in Norway, show the highest overall mean values. In terms of the share of R&D labor, cluster 6 again diverges from the other clusters, showing a steady decline over time, while the other clusters remain largely stable. In terms of labor density, the mean curve for cluster 3 is significantly higher than all other curves, which are stable over time. The mean curves of cluster 5 and 6 show opposite trends in R&D investment per R&D labor, with one cluster increasing while the other decreases and vice versa. With the exception of the last four years, cluster 5 is mostly above the other clusters, which show a slight and steady increase over time.

## 4.3 IPC Class C

For the patent class *'chemistry and metallurgy'*, the spatio-temporal clustering process again resulted in ten clusters (see Figure 1, 3rd row, left panel). Essentially, the clustering appears to be similar to the results of the previous patent class, with a few exceptions. For example, cluster 6, 9 and 10 are mostly identical, with two Portuguese regions now belonging to the cluster mainly located in Eastern Europe. The latter is no longer divided, as all Eastern European regions have innovation profiles that make them part of the same homogeneous cluster. Compared to the results of the previous patent class, there are some changes affecting a few regions in cluster 1. While Central Europe is again fragmented compared to patent class B, and this also applies to a higher degree to Spain and Portugal, the homogeneous structure of Eastern Europe represents a clear contrast to the rest of Europe.

The mean curves for the Innovation Ginis are quite similar in their temporal dynamics, with several curves intersecting each other, but most remaining within a narrow, slightly declining corridor (see Figure 1, 3rd row, right panel). While cluster 3, with its focus on Eastern Europe, has the highest level of mean curves, but declines sharply from 2008 onwards, clusters 5 and 6 seem to develop comparably from 2008 onwards with slight time lags, whereas previously they had complementary trajectories. As in the previous cluster results, the labor-related covariates show a slight, but constant increase in the mean curves. In terms of human capital density, the mean curves are again close to each other and also increase linearly. There is an increase in GDP per labor for all clusters, with a slight dip in 2008 and the Eastern European regions of cluster 3 showing the lowest mean curve values. The share of R&D labor is more or less stagnant for all clusters, with cluster 5 again showing the highest mean curve values. In terms of R&D investment in R&D labor, the spatially-spread cluster 8 shows a strong increase in 2003, followed by a similarly long decline until 2008. Regarding labor density, all clusters show linear trajectories at very low levels, with the exception of cluster 5, which is spread over half a dozen regions across Europe and shows significantly higher and slightly increasing mean curve values. Across all co-variates, mean curves of cluster 3 are lower than all others, with the exception of Innovation Gini curves.

## 4.4 IPC Class D

In terms of the patent class for innovations in *'textiles and paper'*, a set of 9 distinct clusters was found in the clustering process, possibly due to missing data for some regions included in previous clustering results (see Figure 2, 1st row, left panel). The cluster with the highest number of regions is cluster 3, which includes regions in Finland, most of France and parts of Italy, Austria, Germany, Belgium, Luxembourg and the Netherlands. With the exception of Southern Germany, Belgium, Luxembourg and the Netherlands, neighboring regions are part of the same cluster. The United Kingdom is divided

between four different clusters, with cluster 9 occurring only in England. Most parts of Eastern Germany, Northern Ireland and parts of Spain are members of cluster 8. As the innovation profiles for this patent class seem to be more homogeneous than in previous results, most regions belong to a few larger clusters, while the remaining regions are divided into the highly distinct clusters 1, 4 and 6.

Compared to the previous results, the mean curves for the Innovation Ginis are at a very low level (see Figure 2, 1st row, right panel). Here, the regions of Eastern Europe and Portugal of cluster 4 show the highest values, but decline slightly after 2008. This contrasts with cluster 6, which consists of only two regions and follows a U-shaped trajectory, so that the curve only rises steadily after 2008 and shows the highest values of all clusters. Regarding the other variables, the mean curves show mostly linear trajectories, with a a few exceptions such as cluster 1, which mostly consists of Norwegian regions and shows the highest mean values with increasing trends. This is particularly noticeable for labor density and GDP per labor. Compared to the other clusters, cluster 6 varies the most over time, with its trajectory changing towards 2004 and even increasing non-linearly for both covariates and human capital density. As with the previous results, the most variation across all clusters is found for R&D investment per R&D labor. Here, cluster 5, which is scattered across Europe, shows a sharp increase to the highest mean curve value in 2008, before declining in a similar way. This is mirrored at a lower level in cluster 3, 7 and 9, with the first two reaching their maximum around 2002.

## 4.5 IPC Class E

The clustering for innovations in '*fixed constructions*' are quite similar to the clustering for the IPC class C, although only 8 clusters are found (see Figure 2, 2nd row, left panel). Essentially, clusters 2, 4 and 7 are evidence for this similarity. The fragmentation of Central Europe is shifted slightly to the west, as western German regions are members of the same cluster. The Scandinavian cluster also is found in Central European regions and is scattered across northern Italy, parts of the United Kingdom and Ireland. Another similarity to the innovation profiles of IPC classes C and E can be seen through due to cluster 2, which is exclusively found in the UK.

In comparison with the results of IPC class C, the mean curves for the Innovation Ginis for fixed constructions are on a much lower level, with the Eastern European regions of cluster 7 showing the highest mean curve values (see Figure 2, 2nd row, right panel). Furthermore, clusters 5 and 6 show complimentary trajectories and while the mean curves are steadily decreasing, at the same time the dispersion of all mean curves is decreasing over time. For labor density, share of R&D labor, GDP per labor and human capital density, the temporal dynamics of the mean curves are again comparable to the results of IPC class C, with the exception being that the dispersion across the mean curves is much smaller and no cluster has significantly higher mean values than all other clusters. In terms of R&D investment per R&D labor, clusters 5 and 6 show high maxima in the period from 2000 to 2004 and then converge to the overall corridor of cluster mean curves.

## 4.6 IPC Class F

For the patent class for '*mechanical engineering, lighting, heating, weapons, and blasting*' innovations, eight clusters are found, again showing noticeable similarities to the cluster results for IPC class C (see Figure 2, 3rd row, left panel). Especially the Eastern European regions (cluster 6), Scandinavia and parts of Central Europe (cluster 4), Spain, Portugal and East Germany (clusters 2 and 6) as well as France (clusters 6 and 7) are the reason for the similarities in the spatial cluster pattern. Nevertheless, some deviations from previous clustering results can be found in western Germany, northern Italy, Austria and parts of France. Compared to Western Europe, innovation profiles int the Northern and Eastern European regions seem to be more homogenous.

The cluster mean curves for the Innovations Ginis show some temporal variation and an overall increasing trend, with cluster 6 showing the highest level until 2004 before decreasing thereafter (see Figure 2, 3rd row, right panel). The strongest increase is shown by the mostly non-adjacent Central Europe regions of cluster 5 and the cluster 2 (East Germany and Spain), while the regions of cluster 4 stagnate at a stable level. With regard to the other covariates, similar temporal trajectories as for IPC class E are shown for the mean curves. Due to the curve maxima not standing out from the curves as in previous results, the cluster mean curves show smoother trajectories overall.

## 4.7 IPC Class G

As with most previous results, ten innovation clusters are found for the patent class for *'physics'* that resemble the clustering pattern of IPC class C, while sharing a few similarities with IPC classes E and F (see Figure 3, 1st row, left panel). With the exception of Denmark, which is now an independent cluster with a single region in northern Germany (cluster 7) and no longer part of the Scandinavian cluster (cluster 2). In addition, some smaller regions in the Netherlands are assigned differently compared to other IPC classes.

In the cluster mean curves of the Innovation Ginis, both the overall level and the curve maxima are very pronounced in comparison to IPC class C (see Figure 3, 1st row, right panel). The highest mean curve values are shown for cluster 10, which is located mostly in Norway, while cluster 8 increases sharply in 2004 before matching the temporal dynamic of cluster 10 from 2008 onwards. In contrast, most cluster mean curves remain stable for this variable, with cluster 7 being an exception that decreases over time. As far as the other variables are concerned, cluster 10 has the highest level of all mean curves for almost all of these variables. For labor density, share of R&D labor, GDP per labor and human capital density, the difference between cluster 10 and the other clusters in the mean values is very clear. Only in terms of R&D investment per R&D labor is cluster 10 surpassed by the maxima of clusters 6 and 7 until 2005, but as these curves decline again, the mean curve of cluster 10 reaches the highest level again in 2012.

## 4.8 IPC Class H

The last clustering found a set of ten clusters for electrical innovations (IPC H, see Figure 3, 2nd row, left panel). The East German regions are divided into three larger clusters (clusters 1, 6 and 7), with cluster 6 again consisting of regions in East Germany, Spain and Portugal that were assigned to the same cluster for other IPC classes. In addition, several regions in Ireland, England as well as southern Italy are members of this cluster. Scandinavia is also divided into three clusters, with members of cluster 4 found in Finland, France, Austria and other regions all throughout Central Europe. Another cluster is found in the southern regions of Norway as well as the central region of Paris. The rest of Scandinavia is clustered together with southern Ireland, southern Germany and the central London region (cluster 5). Most regions in England form their own cluster, with only few exceptions in Denmark, Italy and the southwest of France (cluster 10).

Consistent with all previous results, there is a high degree of variation in the mean curves of innovation concentration, with cluster 1 showing the highest overall mean values, but steadily decreasing over time, with a noticeable minimum in 2005 (see Figure 3, 2nd row, right panel). While most other clusters stagnate at a stable level, cluster 7 increases sharply after a minimum in 2001. In comparison, cluster 10 decreases until 2004, stagnates until 2009 before finally increasing. In terms of the other covariates, there are large similarities to IPC class G. Regarding labor density, cluster 8 shows the highest mean values and an increase with a clear gap to the other clusters, which remain constant over time and show similar values with minor variations. For all other variables, cluster 10 differs the most from the other clusters, as the mean curve for share of R&D labor decreases while all others increase constantly.

In addition, there is a dip in GDP per labor in 2007 and human capital seems to be gradually increasing for cluster 10. While the mean curves for most clusters are quite similar to the curves for IPC class G, high maxima for R&D investment per R&D labor are missing.

# 5 Discussion

Overall, innovation clusters in Europe differ by IPC class, although some regions are more similar than others and some IPC classes more interconnected in terms of innovation concentration. The Innovation Gini is mostly similar in the main regions in Eastern Europe, Spain, Portugal, and East Germany, resulting in these regions being in the same spatio-temporal cluster groups. Regarding the Innovation Gini and the different covariates used in the mixture model-based multivariate functional clustering, it is noticeable that some covariates seem to have opposite functional effects. This is the case when considering regions with the highest values of innovation concentration over times, which is usually accompanied by the lowest values in the covariates. This holds for all IPC classes except for class G (*'physics'*). The clustering results for the classes E, C and F are similar, with the pair E/F being more similar than the pair E/C. In addition, cluster solutions for the classes G and C as well as G and H show similar temporal dynamics.

If one relates the clustering results to analyses of innovation promoting policies from the same period, the clustering clearly shows the various efforts in innovation policy and general economic trends such as the economic crisis of 2008. The crisis is reflected in the functional curves and affects almost all IPC cluster solutions, with some being more affected than others. As Izsák et al. (2013) state in their final report for the European Commission, funding focused on innovation development slowed down during the period of our analysis, especially after the economic crisis. Nevertheless, funding shifted towards more collaborative projects which is one reason that our analysis showed the emergence of clusters not only of neighboring regions, but also at supra-regional level. Furthermore, the funding priorities have not shifted in their scientific and technological cores, so the FDA cluster model should be able to capture relevant effects to a large degree.

In general, our analysis would benefit from longer time series of data that could provide further insights into national and regional innovation dynamics. The time periods of funding programs often span several years or even decades, and it is possible that their impact is not fully captured by the analysis conducted in this paper. Similarly, it is possible that the impact of regional innovation policies has not been significant enough to have lasting effects related to innovation concentration (Izsák et al., 2013).

The concept of the European and Regional Innovation Scoreboards takes into account innovation developments over time and divides nations by regions, but policies derived from the European legislation are relatively inconsistent when. The innovation index generally shows little variation between countries, with most countries occupying the same or similar categories of innovation leadership. This is also true across regions, with exceptions due to highly specialized regions (e.g. Malta as a moderate innovator, is among strongest innovators in digitalization) (European Commission, 2021b).

Izsák et al. (2013) conclude that innovation policy should location-based and tailored to different conditions in order to take into account national characteristics. This idea is supported by the results of our analysis, as regional characteristics and differing conditions in the technological mix foster the emergence of heterogeneous innovation portfolios and thus suggest higher policy efficiency if properly taken into account.

# 6 Conclusion

Knowledge about regional innovation dynamics, leading to different Innovation Ginis that result in clustering regions differently across all of Europe, depending on the type of innovation activity is crucial when designing policies for supporting innovation in Europe.

In this paper, a mixture model-based multivariate functional clustering algorithm by Schmutz et al. (2020) has been adapted to analyze the spatio-temporal dynamics of European regional innovation activities at the NUTS-2 level from 2000 to 2012. Using multiple time series for the main patent classes as proxies for innovation activity as well as other closely related indicators, 225 Europeans regions were clustered according to their temporal innovation profiles. In this way, multi-characteristic innovation activity is taken into account, reflecting the political efforts of European policy programs. Our measurements for identifying the clusters are innovation- and economy-related variables including innovation concentration indicators which are based on Krugman's (1991) Innovation Ginis, with the distinction that as innovation indicator different IPC classes of patents are considered and regions are profiled according to their innovation portfolios.

The resulting innovative activity across the European clusters differs, although some regions in Eastern Europe and on the Iberian Peninsula are reliably constant across innovation type. Accounting for the differences in innovation, clustering for IPC classes E (*'fixed constructions'*) and F '*mechanical engineering, lighting, heating, weapons, and blasting'*) is almost identical, whereas similarities in regional clustering of classes E and C ('*chemistry and metallurgy'*) are relatively more distinct, but still comparable. Clusters of classes G ('*physics'*) and C are correspondent while classes H ('*electricity*') and G exhibit comparable dynamics over time. This supports a place based regional innovation policy approach that is not only able to account for differing regional potentials in innovation, but also for diverging specialization in innovation types.

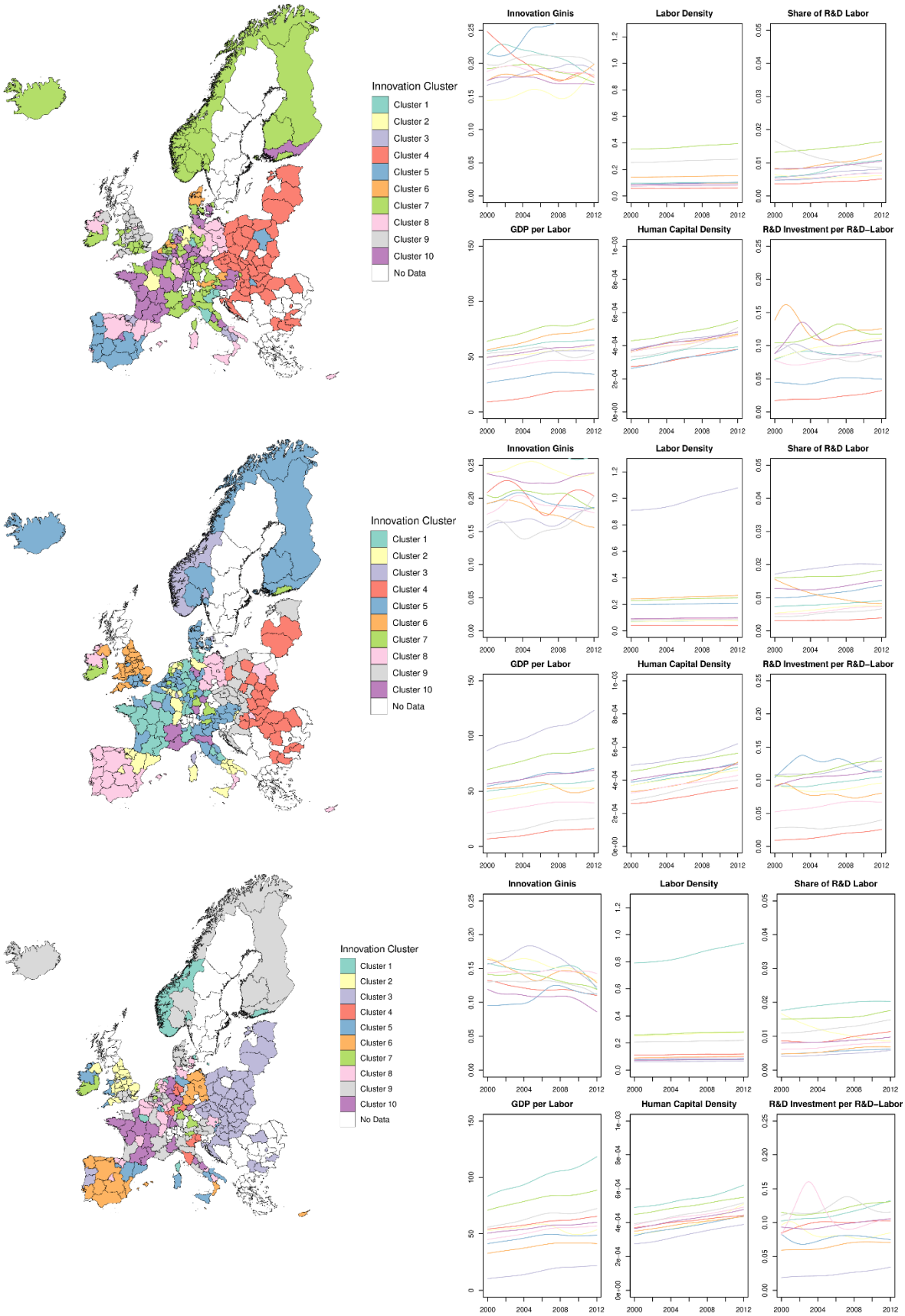


*Figure 1: Results of the mixture model-based multivariate functional clustering algorithm. Rows: IPC classes A, B, and C (A: 'human necessities'; B: 'performing operations and transporting'; C: 'chemistry and metallurgy'). Columns: left: Spatial cluster mapping, right: Temporal cluster dynamics. Source: Own calculations.*

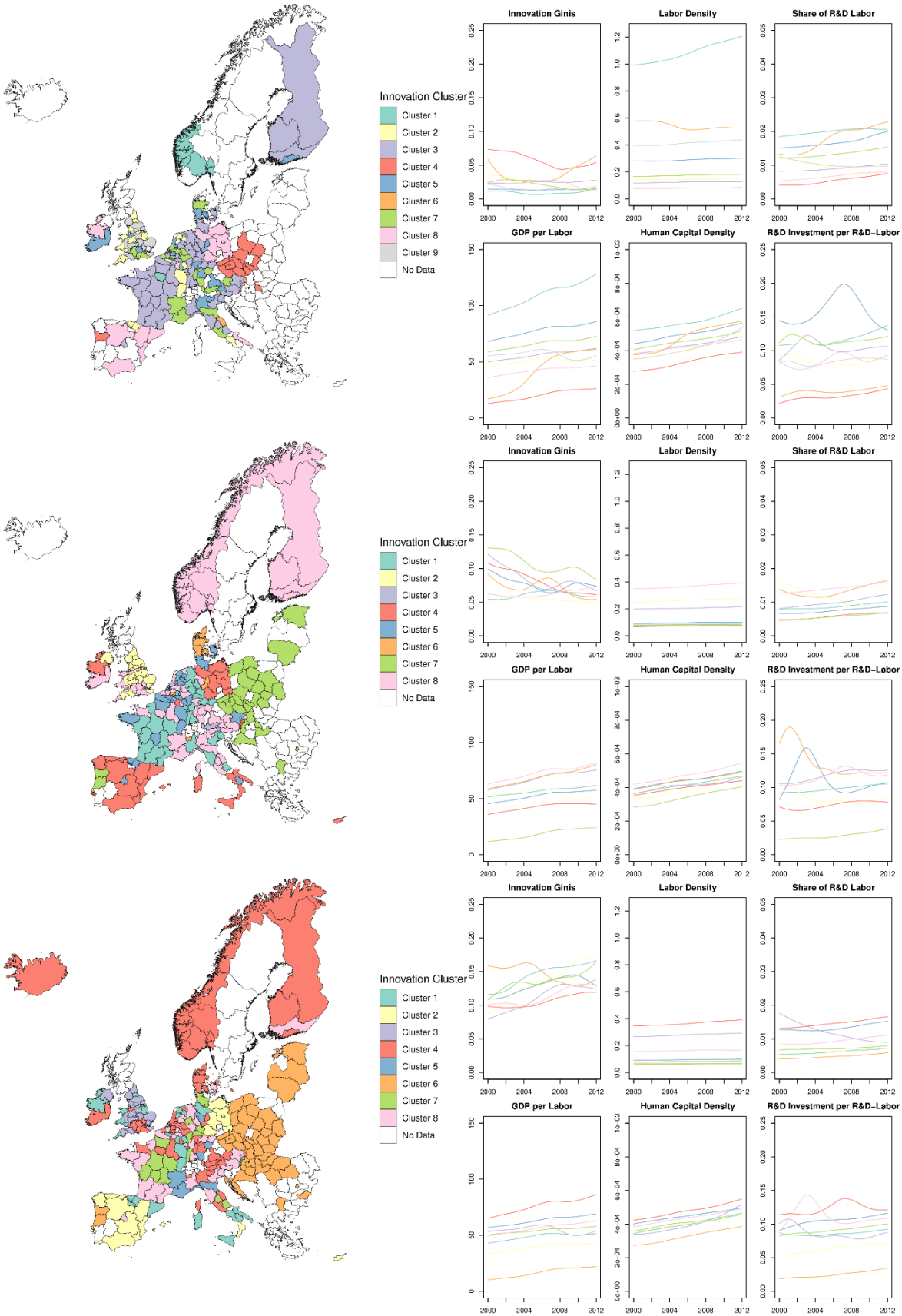


*Figure 2: Results of the mixture model-based multivariate functional clustering algorithm. Rows: IPC classes D, E, and F (D: 'textiles and paper'; E: 'fixed constructions'; F: 'mechanical engineering, lighting, heating, weapons, and blasting'). Columns: left: Spatial cluster mapping, right: Temporal cluster dynamics. Source: Own calculations.*

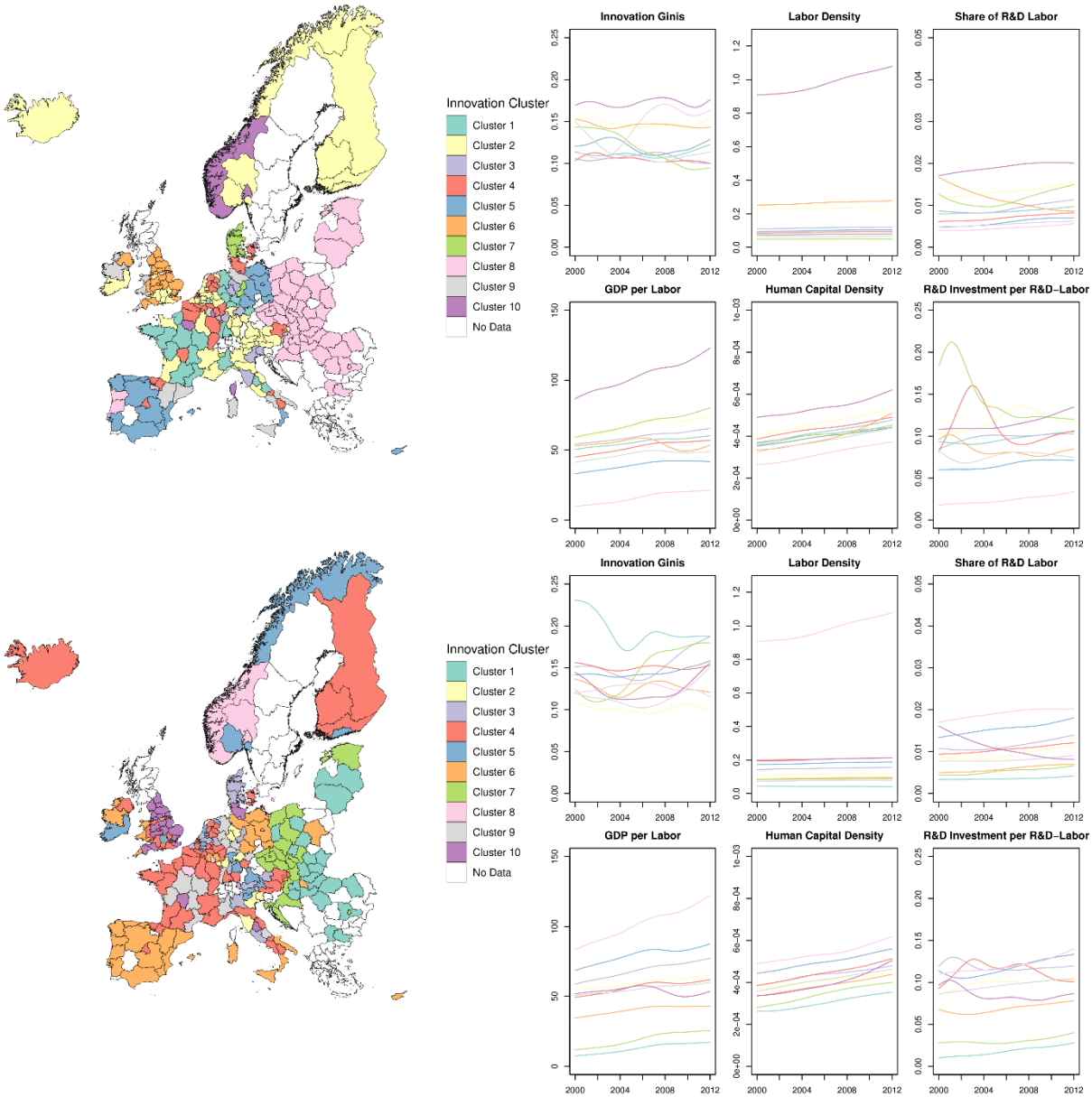


*Figure 3: Results of the mixture model-based multivariate functional clustering algorithm. Rows: IPC classes G and H (G: 'physics'; H: 'electricity'). Columns: left: Spatial cluster mapping, right: Temporal cluster dynamics. Source: Own calculations.*

# 7 References


Abraham, C., Cornillon, P.-A., Matzner-Løber, E., & Molinari, N. (2003). Unsupervised curve clustering using B-splines. *Scandinavian journal of statistics, 30*(3), 581-595.

Acs, Z. J., & Varga, A. (2002). Geography, endogenous growth, and innovation. *International Regional Science Review, 25*(1), 132-148.

Aghion, P., Hemous, D., & Veugelers, R. (2009). No green growth without innovation. *Bruegel Policy Brief-2009/07*.

Aguilera, A. M., Escabias, M., Preda, C., & Saporta, G. (2010). Using basis expansions for estimating functional PLS regression: applications with chemometric data. *Chemometrics and Intelligent Laboratory Systems, 104*(2), 289-305.

Alcácer, J., & Chung, W. (2014). Location strategies for agglomeration economies. *Strategic Management Journal, 35*(12), 1749-1761.

Bellman, R. E. (1957). *Dynamic Programming*: Princeton University Press.

Blume, L. (2004). Erfolgsfaktoren kommunaler Wirtschaftspolitik in Ostdeutschland. *Raumforschung und Raumordnung, 62*(1), 3-17.

Bongiorno, E. G., & Goia, A. (2016). Classification methods for Hilbert data based on surrogate density. *Computational statistics & data analysis, 99*, 204-222.

Bottazzi, L., & Peri, G. (2003). Innovation and spillovers in regions: Evidence from European patent data. *European economic review, 47*(4), 687-710.

Bouveyron, C., Côme, E., & Jacques, J. (2015). The discriminative functional mixture model for a comparative analysis of bike sharing systems. *The Annals of Applied Statistics, 9*(4), 1726-1760.

Bouveyron, C., & Jacques, J. (2011). Model-based clustering of time series in group-specific functional subspaces. *Advances in Data Analysis and Classification, 5*(4), 281-300.

Breuste, J., & Keidel, T. (2008). Urbane und suburbane Räume als Kulturlandschaften-planerische Gestaltungsaufgaben. *informationen zur raumentwicklung*(5), 279-288.

Capello, R., & Lenzi, C. (2013). Territorial patterns of innovation: a taxonomy of innovative regions in Europe. *The Annals of regional science, 51*(1), 119-154.

Caves, D. W., Christensen, L. R., & Diewert, W. E. (1982). The Economic Theory of Index Numbers and the Measurement of Input, Output, and Productivity. *Econometrica, 50*(6), 1393-1414. doi:10.2307/1913388

Chiou, J. M., & Li, P. L. (2007). Functional clustering and identifying substructures of longitudinal data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 69*(4), 679-699.

Christaller, W. (1933). Die Zentralen Orte in Suddeutschland, Jena. *Central Places in Southern Germany*.

Coffey, N., Hinde, J., & Holian, E. (2014). Clustering longitudinal profiles using P-splines and mixed effects models applied to time-course gene expression data. *Computational statistics & data analysis, 71*, 14-29.

Delaigle, A., & Hall, P. (2010). Defining probability density for a distribution of random functions. *The Annals of Statistics, 38*(2), 1171-1193.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological), 39*(1), 1-22.

Deville, J.-C. (1974). *Méthodes statistiques et numériques de l'analyse harmonique.* Paper presented at the Annales de l'INSEE.

Eubank, R. L. (1999). *Nonparametric regression and spline smoothing*: CRC press.

European Commission. (2019). The EU Green Deal. Retrieved from https://ec.europa.eu/info/strategy/priorities-2019-2024/european-green-deal_en

European Commission. (2021a). EU Cohesion Policy. Retrieved from https://ec.europa.eu/regional_policy/en/2021_2027/

European Commission. (2021b). European Innovation Scoreboard

Integrated European innovation scoreboard tool. Retrieved from https://interactivetool.eu/EIS/EIS_2.html

European Commission. (2021c). Internal Market, Industry, Entrepreneurship and SMEs - European Innovation Scoreboard. Retrieved from https://ec.europa.eu/growth/industry/policy/innovation/scoreboards_en

Eurostat. (2021). Eurostat Database. In.

Fernau, A. K. (1997). Standortwahl als Komponente der Wettbewerbsfähigkeit. *Wiesbaden. Deutscher Universitätsverlag*.

Ferraty, F., & Vieu, P. (2006). *Nonparametric functional data analysis: theory and practice*: Springer Science & Business Media.

Fornahl, D., & Brenner, T. (2009). Geographic concentration of innovative activities in Germany. *Structural Change and Economic Dynamics, 20*(3), 163-182.

Gasser, T., & Kneip, A. (1995). Searching for structure in curve samples. *Journal of the American Statistical Association, 90*(432), 1179-1188.

Gasser, T., Muller, H.-G., Kohler, W., Molinari, L., & Prader, A. (1984). Nonparametric regression analysis of growth curves. *The Annals of Statistics*, 210-229.

Giannitsis, T., & Kager, M. (2009). Technology and specialization: dilemmas, options and risks. *Knowledge for Growth. Prospect for Science, Technology and Innovation*.

Grenander, U. (1950). Stochastic processes and statistical inference. *Arkiv för matematik, 1*(3), 195-277.

Griliches, Z. (1990). *Patent Statistics as Economic Indicators: A Survey Part II*: NATIONAL BUREAU OF ECONOMIC RESEARCH (NBER).

Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A k-means clustering algorithm. *Journal of the royal statistical society. series c (applied statistics), 28*(1), 100-108.

Honaker, J., & King, G. (2010). What to do about missing values in time-series cross-section data. *American journal of political science, 54*(2), 561-581.

Ieva, F., & Paganoni, A. M. (2016). Risk prediction for myocardial infarction via generalized functional regression models. *Statistical methods in medical research, 25*(4), 1648-1660.

Ionela-Andreea, P., & Marian, N. (2020). Cluster Analysis of Regional Research and Development Disparities in Europe. *Studies in Business and Economics, 15*(3), 303-312. doi:doi:10.2478/sbe-2020-0060

Izsák, K., Markianidou, P., & Radošević, S. (2013). *Lessons from a Decade of Innovation Policy - What can be learnt from the INNO Policy TrendChart and The Innovation Union Scoreboard*. Retrieved from https://www.researchgate.net/publication/272566502_Lessons_from_a_Decade_of_Innovation_Policy_What_can_be_learnt_from_the_INNO_Policy_TrendChart_and_The_Innovation_Union_Scoreboard/link/57d7d2ef08ae601b39aebf36/download

Jacques, J., & Preda, C. (2013). Funclust: A curves clustering method using functional random variables density approximation. *Neurocomputing, 112*, 164-171.

Jacques, J., & Preda, C. (2014a). Functional data clustering: a survey. *Advances in Data Analysis and Classification, 8*(3), 231-255.

Jacques, J., & Preda, C. (2014b). Model-based clustering for multivariate functional data. *Computational statistics & data analysis, 71*, 92-106.

James, G. M., & Sugar, C. A. (2003). Clustering for sparsely sampled functional data. *Journal of the American Statistical Association, 98*(462), 397-408.

Kayano, M., Dozono, K., & Konishi, S. (2010). Functional cluster analysis via orthonormalized Gaussian basis expansions and its application. *Journal of classification, 27*(2), 211-230.

Kim, G., & Bae, J. (2017). A novel approach to forecast promising technology through patent analysis. *Technological Forecasting and Social Change, 117*, 228-237.

Kokoszka, P., & Reimherr, M. (2017). *Introduction to functional data analysis*: Chapman and Hall/CRC.

Koschatzky, K. (2001). *Räumliche Aspekte im Innovationsprozess: ein Beitrag zur neuen Wirtschaftsgeographie aus Sicht der regionalen Innovationsforschung*: LIT Verlag Münster.

Krugman, P. (1979). A Model of Innovation, Technology Transfer, and the World Distribution of Income. *Journal of Political Economy, 87*(2), 253-266. Retrieved from http://www.jstor.org/stable/1832086

Krugman, P. R. (1991). *Geography and trade*: MIT press.

Li, P.-L., & Chiou, J.-M. (2011). Identifying cluster number for subspace projected functional data clustering. *Computational statistics & data analysis, 55*(6), 2090-2103.

Lösch, A. (1940). Die räumliche Ordnung der Wirtschaft: eine Untersuchung über Standort, Wirtschaftsgebiete und internationalem Handel [The spatial order of the economy: An investigation of the location, economic areas and international trade]. *Jena, Fischer Publ*.

Marshall, A. (1920). Principles of economics an introductory volume Macmillan and Co. *London, UK*.

OECD. (2021). Database. Retrieved from https://data.oecd.org/

Office for National Statistics. (2021). Database. Retrieved from https://www.ons.gov.uk/

Pelau, C., & Chinie, A. C. (2018). Cluster Analysis for the Determination of Innovative and Sustainable Oriented Regions in Europe. *Studia Universitatis Vasile Goldiș Arad, Seria Științe Economice, 28*(2), 36-47.

Pełka, M. (2018). Analysis of innovations in the European Union via ensemble symbolic density clustering. *Ekonometria, 22*(3), 84-98.

Peng, J., & Müller, H.-G. (2008). Distance-based clustering of sparsely observed stochastic processes, with applications to online auctions. *The Annals of Applied Statistics, 2*(3), 1056-1077.

Pflüger, M. (2019). Regionale Disparitäten und Regionalpolitik: Treiber der Veränderung, Handlungsbedarf und Handlungsoptionen. *Wirtschaftsdienst, 99*(Sonderheft), 10-18.

R Core Team. (2021). R: A language and environment for statistical computing. In. Vienna, Austria: R Foundation for Statistical Computing.

Ramsay, J., & Silverman, B. W. (1997). Functional data analysis (Springer series in statistics).

Ramsay, J. O. (1982). When the data are functions. *Psychometrika, 47*(4), 379-396.

Ramsay, J. O., & Dalzell, C. (1991). Some tools for functional data analysis. *Journal of the Royal Statistical Society: Series B (Methodological), 53*(3), 539-561.

Ramsay, J. O., Graves, S., & Hooker, G. (2021). *fda: Functional Data Analysis*. Retrieved from https://CRAN.R-project.org/package=fda

Rao, C. R. (1958). Some statistical methods for comparison of growth curves. *Biometrics, 14*(1), 1-17.

Rhoden, I. (2019). *Space and Time in German Innovative Activity*: Springer.

Rhoden, I. (2020). Innovating in Krugman's Footsteps–Where and How Innovation Differs in Europe: Static Innovation Indicators for Identifying Regional Policy Leverages. *ZBW – Leibniz Information Centre for Economics*. Retrieved from https://www.econstor.eu/handle/10419/218875

Rice, J. A., & Silverman, B. W. (1991). Estimating the mean and covariance structure nonparametrically when the data are curves. *Journal of the Royal Statistical Society: Series B (Methodological), 53*(1), 233-243.

Schmutz, A., Jacques, J., & Bouveyron, C. (2021). *funHDDC: Univariate and Multivariate Model-Based Clustering in Group-Specific Functional Subspaces*. Retrieved from https://CRAN.R-project.org/package=funHDDC

Schmutz, A., Jacques, J., Bouveyron, C., Cheze, L., & Martin, P. (2020). Clustering multivariate functional data in group-specific functional subspaces. *Computational Statistics*, 1-31.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 461-464.

Serban, N., & Wasserman, L. (2005). CATS: clustering after transformation and smoothing. *Journal of the American Statistical Association, 100*(471), 990-999.

Simonoff, J. S. (2012). *Smoothing methods in statistics*: Springer Science & Business Media.

Spielkamp, A., & Vopel, K. (1999). Mapping innovative clusters in national innovation systems. *Boosting Innovation the Cluster Approach*, 91.

Sraffa, P., & Dobb, M. H. (1951). *The works and correspondence of David Ricardo* (Vol. 1): Cambridge University Press Cambridge.

Sturgeon, T. J. (2003). What really goes on in Silicon Valley? Spatial clustering and dispersal in modular production networks. *Journal of economic geography, 3*(2), 199-225.

Thünen, J. H. (1966). v.(1826): Der isolierte Staat in Beziehung auf Landwirtschaft und Nationalökonomie. *Perthes, Hamburg*.

Traore, O., Cristini, P., Favretto-Cristini, N., Pantera, L., Vieu, P., & Viguier-Pla, S. (2019). Clustering acoustic emission signals by mixing two stages dimension reduction and nonparametric approaches. *Computational Statistics, 34*(2), 631-652.

Turkina, E., & Van Assche, A. (2018). Global connectedness and local innovation in industrial clusters. *Journal of International Business Studies, 49*(6), 706-728.

Wang, J.-L., Chiou, J.-M., & Müller, H.-G. (2015). Review of functional data analysis. *arXiv preprint arXiv:1507.05135*.

WIPO. (2021). WIPO IP Portal. Retrieved from https://www.wipo.int/classifications/ipc/ipcpub/?notion=scheme&version=20210101&symbol=none&menulang=en&lang=en&viewmode=f&fipcpc=no&showdeleted=yes&indexes=no&headings=yes&notes=yes&direction=o2n&initial=A&cwid=none&tree=no&searchmode=smart

Zabala-Iturriagagoitia, J. M., Aparicio, J., Ortiz, L., Carayannis, E. G., & Grigoroudis, E. (2021). The productivity of national innovation systems in Europe: Catching up or falling behind? *Technovation, 102*, 102215.