International Institute for Applied Systems Analysis
IIASA www.iiasa.ac.at

**Address:** IIASA, Schlossplatz 1, A-2361 Laxenburg, Austria
**Email:** repository@iiasa.ac.at **Telephone:** +43 (0)2236 807 342

gm

YSSP Report
**Young Scientists Summer Program**

# Evolving the Knowledge Space:
## Towards a Selection Dynamics Model of Patent Classes

Bernardo Sousa Buarque
(bbuarque@gmail.com)

## Approved by

**Supervisor**: Gergely Boza
**Co-Supervisor**: Gerald Silverberg

**Program**: Evolution and Ecology Program (EEP)
October 31st, 2020

This report represents the work completed by the author during the IIASA Young Scientists Summer Program (YSSP) with approval from the YSSP supervisor.

It was finished by _____ and has not been altered or revised since.

This research was funded by IIASA and its National Member Organizations in Africa, the Americas, Asia, and Europe.

# Table of Contents

# Abstract

The current report seeks to understand the selection dynamics of patent classes (CPCs) in Europe by employing the methodology developed by Acemoglu et al. (2016) to predict future patenting. Their research focuses on citation networks measuring the knowledge flows across technologies and uses theses to estimate future volumes of patents per CPC during 1995-2004 in the United States. In our current analysis we replicate their results using the European patent database for the years 2005-2014, and likewise demonstrate that the innovation networks have significant predictive power over future patenting in Europe. Furthermore, we improve their methodology by accounting for more complex interactions between CPCs. Finally, we discuss their implications for developing a selection-dynamics model grounded in evolutionary theory.

# Acknowledgments

I start by thanking both my supervisors - Gergerly Boza and Gerald Silverberg. Their assistance was vital to the development of this research. And my study would not have been possible without them.

I also want to thank the organizers of this summer programme and those supporting our every step. It was an exceptionally challenging summer, but their help made everything effortless and rewarding.

Likewise, I need to acknowledge all those who participated in the workshops and presentations, those who contributed with valuable insights and comments during my presentations.

I must acknowledge as well my YSSP fellows. Not only did they contribute to the success of my research, but also made the entire experience more stimulating and enjoyable.

Finally, I want to thank those in my home institution - the University College Dublin. My doctoral supervisor, Dieter Kogler, and all my colleagues at the Spatial Dynamics Lab.

# About the authors

**Bernardo Buarque** is a fourth-year PhD candidate for the Spatial Dynamics Lab at the University College Dublin (Contact: bbuarque@gmail.com)

# Introduction

How can we understand and model the evolution of technology? Ironically, the answer to this question will never be categorical and will likely forever change in time. Still, Arthur and Polack (2006) offer perhaps the best definition to the problem when they assert:

> *"new technologies are never created from nothing. They are constructed – put together – from components that previously exist; and in turn these new technologies offer themselves as possible components – building blocks – for the construction of further new technologies" (p.23).*

Inspired by this notion that knowledge begets knowledge, recent theoretical studies sought to display and often predict the direction of technological change. Grounded on the theory of the adjacent possible (Kauffman, 2000), these models describe the creation of new products as the combination, recombination, modification, and gradual adaptation of an existing set of components. They show the process of innovation as a constant search for improved combinations of ingredients and whereby the new mixes descend from the existing ones (Auerswald et al., 2000; Silverberg, 2002; Silverberg & Verspagen, 2005, 2007; McNerney et al., 2011; Korhonen & Kasmire, 2013; van Dam & Frenken, 2020).

While these theoretical models set the groundwork for how we examine the evolution of technology, we also witness a swell of empirical papers striving to complement, apply, and validate the ideas from these models. Foremost, making use of patent data and its inherent classification system, past authors tried to predict the pace of technological change. They tried to describe the creation of new patents or to anticipate novel combinations of codes (Youn & Magee, 2018; Tacchella et al., 2020).

Acemoglu et al. (2016) stand as the first to use citation patterns across patent applications to predict future rates of innovation by code. Indeed, as the authors suggest: "the interaction of this pre-existing network structure with patent growth in upstream technology fields has strong predictive power on the future of innovation" (P.11483). Besides, over the years, other scholars equally employed the "interdependence between technologies to predict innovation dynamics" (Pichler et al.,2020).

Along these lines, this paper aims to advance the related literature. Focusing on the work by Acemoglu et al. (2016), we wish to understand if we can reproduce their findings using a different data source. Namely, we seek to use their innovation networks method to predict innovations for the Cooperative Patent Classes (CPC) in Europe. In doing so, we hope to both confirm and extend their model. Chiefly, we expect that we will learn possible paths to advance the existing literature.

Therefore, like Acemoglu et al. (2016), we created a network graphing the knowledge flows across the CPC classes between 1985-2004. Then, we used these matrices along with the patenting volumes for that time to forecast the number of documents produced per CPC in 2005-2014. Ultimately, we ran a regression to compare these predicted volumes to the actual amount of patent outputs between 2005 and 2014. And we find the methodology presented by them equally describes patent development in Europe. Indeed, for several model specifications and data sample, we consistently find a positive and significant correlation between the predicted and real patenting volumes.

We also provide and demonstrate two modification to their original method. Foremost, we highlight the importance of accounting for the complex interactions of CPCs within every invention. Prior research, such as Acemoglu et al. (2016), considered the CPCs as atomic, isolated entities. Yet, we understand there are emergent properties unique to their combinations. And we show that we can

reproduce the innovation networks approach using the distinct CPC sequences as nodes. Further, doing so seems to improve the quality of our predictions.

Also, we propose an alternative to evaluate the creation of new patents at the regional level. The growing literature on Evolutionary Economic Geography (EEG) examines the dynamics of knowledge production across and within regions. It highlights the influence of spatial and cognitive proximity in the ability to collaborate and innovate (Boschma et al., 2015). Thus, we wanted to examine whether the innovation networks method could help improve this literature too. Foremost, using the method, we sought to reproduce the findings by Kogler et al. (2017) – who studied the dynamics of knowledge production in Europe. And we showed that applying the methodology for the individual regions can produce robust results for most NUTS2 in Europe. More importantly, we show that including citation flows between regions improves the predictions for most places.

In summary, we propose that it is possible to improve the innovation networks methodology in several ways. And, as we shall discuss later, we hope the outputs from this report can help to develop robust selection dynamics models which can predict future patenting. Indeed, most theoretical models on innovation look exclusively at the creation of new ideas. They seldom study how the distribution of a fixed number of technologies change over time. So, inspired by the outputs of this report, we shall propose some alternatives to implement a selection dynamics model that can explain the incidence of CPCs across regions. Namely, we will highlight how the insights from the combined CPCs strategy and the citation flows across places could contribute to building such a model.

We expect these selection models will strengthen how we examine the evolution of technologies. We understand they can assist us when measuring and investigating regional specialization. Ultimately, we hope such data will be useful for assessing the causes and consequences of technological change. And it can be employed to evaluate and propose smart policies.

Social scientists long defended the importance of innovation for economic growth and performance (Lucas, 1988). Likewise, they are aware of technologies potential to disrupt markets and societies and its potential to create winners and losers. Thus, we want a model that predicts the direction of technological change. A model that can anticipate which patent classes ought to become more popular, and which regions are likely to dominate the production of these codes.

Secondly, we imagine these models will help policymakers to design so-called Smart Specialization Strategies. As the European Commission (2020) writes, smart specialization is a "place-based approach characterized by the identification of strategic areas for intervention based both on the analysis of the strengths and potential of the economy." Hence, the insights from a sturdy selection dynamics model will be fundamental for testing and finding the CPCs with higher potential returns for each region. Likewise, the model prediction will be vital for assessing the consequences of these smart policies – i.e., we can compare a region's patenting development to what we expected from the model and see if the policy is boosting a more complex economy.

At last, in this report, we will focus on patents and their classification system. However, we expect the methods proposed here could readily be adapted to study the selection of other product classes. Indeed, the complexity economics model started using the co-occurrence of exports to study regional development and complexity (Hidalgo et al., 2007). Since then, researchers repurposed the approach to focus on industries and occupations (Jara-Figueroa et al., 2019) – even to Olympic medals (Knuepling & Broekel, 2020). Thus, the selection dynamics method could equally be implemented to study the regional shares of occupations – enhancing our grasp of local opportunities and gaps.

After this introduction, the remainder of the paper is organized as follows. Session two presents the patent data used throughout the analysis, how to build the innovation networks, as well as the method borrowed from Acemoglu et al. (2016) to predict patenting in Europe. Session three highlights our top results. And Session four concludes with a discussion on modelling alternatives.

# Data and Methods

Our primary data source is PATSTAT. We retrieved from the database all patent applications filed with the European Patent Office (EPO) between 1985 and 2014. We do restrict our sample to the earliest application per family and those with at least one inventor residing in Europe. Hence, we obtained a data sample consisting of about 1.4 million files. And for each said patent, we collected essential information to construct our analysis. Namely, we took the geolocation of their inventors (European NUTS2) and their filing years.

We collected all front-page references to other EPO documents. If the record cited files from another office (e.g., USPTO or JPO) but which has an equivalent European patent – viz, from the same family – we used this information to retrieve the European one. As before, we restrict our sample to one patent cited per family. We also limit references to those within at most 10-years from publication. Ultimately, we recovered nearly 1.1 million patent citations across the EPO documents. Because not all files refer to others, in the end, we get 600,000 citing patents – which, on average, cited two other EPO documents.

We also retrieved the Cooperative Patent Classification (CPC) assigned to every record. These CPC codes indicate the fundamental building blocks that contributed to the creation of the invention. Hence, they are often used in the literature to study similarity and knowledge flows across technological domains (Leydesdorff et al., 2017). Every patent gets allocated to at least one CPC, but most have more than one. When this is the case, we listed all the CPC belonging to the patent – i.e., we do not limit our data to a single CPC per patent.

Like other patent classification systems, the CPC is a hierarchical arrangement – sorted into sections, which in turn can be divided into classes, sub-classes, and groups. There are more than 150,000 different patent codes within the CPC structure. In this analysis, we will be focusing on the 3-digits (class) and 4-digits (sub-class). These amount to 128 and 650 different codes, respectively. We opted to operate at this level because they are the closest to the USPTO classes used by Acemoglu et al. (2016). Moreover, these are standard levels of granularity used across similar studies (Pichler et al., 2020; Kogler et al., 2017).

To summarize, we obtained data on over 1.4 million European patents between 1985-2014. Out of those patents, we picked 600,000 documents that cite other EPO files. Our data contains the year, location, and classes of each citing patent. Furthermore, we know which records they refer to and have the same set of information about all those documents. We may then use this data to measure knowledge spillovers across technological fields and, like Acemoglu et al. (2016), to predict patenting volumes per CPC for both the European sample and the NUTS2 regions of the continent.

## Innovation Network

There is more than one option to build so-called knowledge networks and use them to measure CPC proximity. For example, inspired by the economics of complexity (Hidalgo et al., 2007), previous studies graphed the co-occurrence of CPCs within patents. At their core, these expositions admit that codes that often occur together exhibit high levels of "cognitive proximity" (Nooteboom, 2000).

Therefore, earlier literature used this insight to compute knowledge relatedness and regional specialization (Kogler et al., 2013; 2017). They employed it to model the dynamics of CPC use and recombination (Youn & Magee, 2018; Tacchella et al., 2020).

We chose to use citation across patents to build a measure of knowledge spillover between the technologies (Acemoglu et al., 2016). That is, we seek to create a network that captures the citation flows between patent classes. A graph that shows the connections between CPCs grounded on how often patents holding one CPC cites documents with another code.

There is still more than one method to construct the said network. And one must be aware that ignoring factors inherent to the patent system might bias our intended measures of proximity (Alstott et al., 2017). Therefore, we decided to build the citation matrix following the straightforward algebra method demonstrated by Pichler et al. (2020).

The CPC citation network is not observable. And we must derive it from the structures formed by the patent citations. Thus, our first step is to build a citations matrix that takes the value of one whenever a patent *p* cites another document *q.* That is, we make a citation matrix ($H_{p,q}$) where each node is a patent and, whenever a file refers to another, we draw an edge between the two.

Next, we use the information enclosed by each patent to establish a patent-technology link matrix. So, we build a bipartite network ($B_{p,i}$) where we write an edge between a patent and a class when the document carries the CPC. Because patents often include more than one CPC code, to avoid inflating the total number of citations each class receives, we decided to row-normalize matrix $B_{p,i}$. Hence, we attribute shares to each CPC within a patent. For example, if a patent contains three different codes, the row-normalized matrix will record a 1/3 share for each CPC.

As demonstrated by Pichler et al. (2020), we can use both matrices described in the above paragraph to make a graph linking CPC classes. That is, we make an algebraical projection of the citation network across the CPC classes as:

$$C_{i,j} = B_{p,i}^{T} H_{p,q} B_{p,i} \tag{1}$$

We produce, thus, a matrix whereby each element ($C_{i,j}$) depicts the sum of all citations from one CPC to another at a given period. Therefore, it measures the knowledge flows across technology codes at the time. Once again, we can row-normalize this $C_{i,j}$ matrix to obtain an adjacency matrix where each element represents a weighted direct edge between the CPC nodes. Mathematically, we express these as a function of the total citations:

$$W_{i,j} = \frac{C_{i,j}}{\sum_{j=1} C_{i,j}} \tag{2}$$

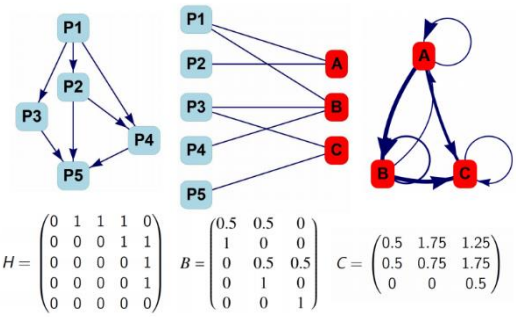
We made the citation network for all three decades in our sample. So, for example, we build a graph for the first decade containing all citing patents filed between 1985-1994. The period restriction applies only to the citing documents. Concerning the cited files, as explained, we limit them to be at most 10-years older than its source.

To further illustrate the method used to build these CPC citation networks, the figures below describe the process in detail as formulated by Pichler et al. (2020) – Panel A. Likewise, in Figure 1.B, we graph an example network which we made using the 3-digits CPCs from 1995 to 2004. To keep the plot simple, we ignore self-citations. The size of the nodes is proportional to their in-degree centrality,

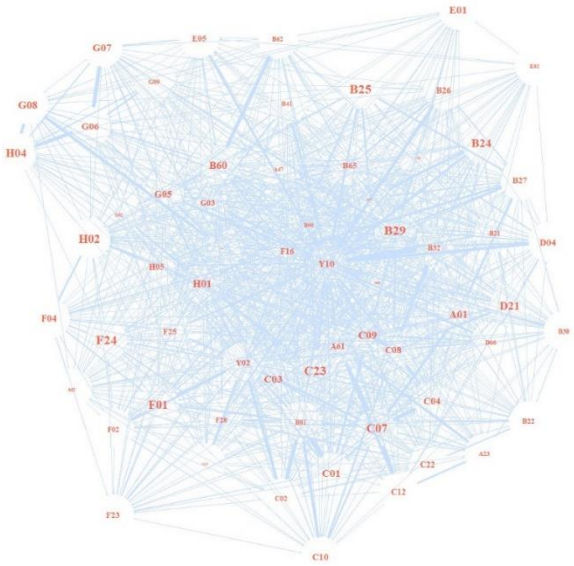
and we only included those nodes with incoming citations above the median. In turn, the edge's widths depend on their weights and, once more, we only added those links with values above the median.

FIGURE 1 – The Innovation Network

PANEL A – Building the Network




PANEL B – Example Network

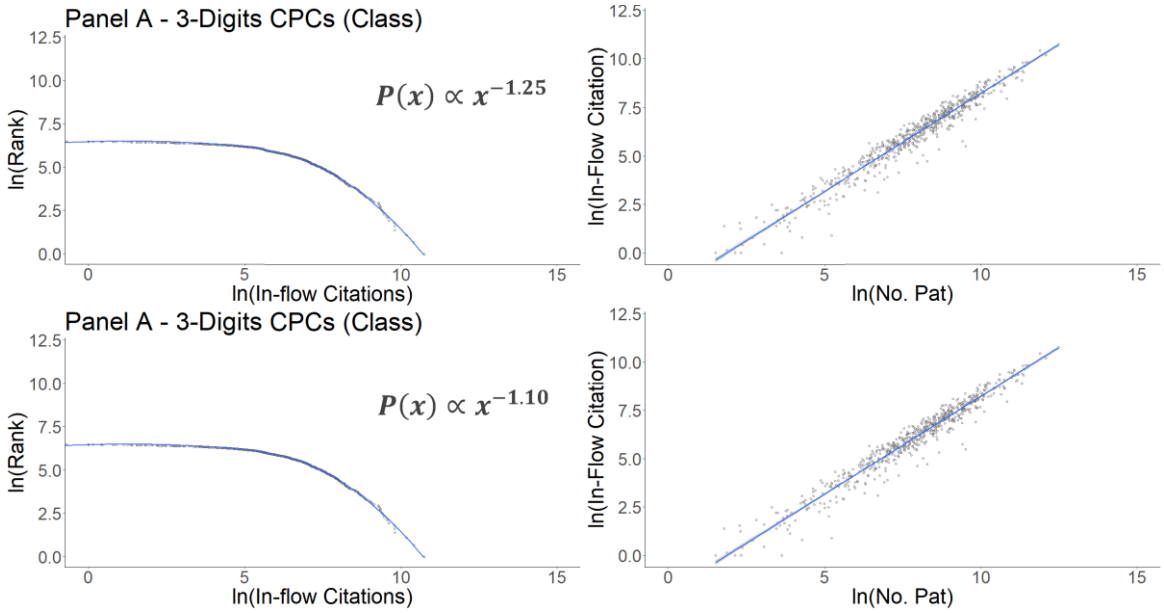


*Note:* The figure on Panel A is an extension of that in Pichler et al. (2020). It describes the algebra used to build the CPC citation matrix (C). The first matrix (P) graphs the citations across patents. For example, because the patent P1 cites P2, we draw a directed edge between the two. The second graph (B) illustrates the links between patents and CPCs. Since the document P1 includes codes A and B on its front-page, we draw an edge between the patent and the two classes. Graph (B) contains weighted edges, so whenever a patent includes two CPCs – like P1 – each link receives a weight of 1/2. We calculate the last network on Panel A using the algebraic method proposed by Pichler et al. (2020) and highlighted on Equation 1. Panel B, in turn, plots an example innovation network – which we build using the patent applications and its 3-digits CPCs from 1995 to 2004. The nodes are CPCs, and the edges are weighted according to their citation flows (i.e., elements of the C matrix). We ignore the self-citation loops, and we only show the CPCs with in-degree centrality above the median. The same holds for the edges.

One might ask how stable these knowledge-flows networks are. To address this question, after creating the matrices for each period, we estimated the correlation amongst them – Appendix I. The test shows the graphs to be consistent and stable for the period under examination. Looking at the 3-digits CPCs, it displays a correlation rate of over 0.85 for the 10-years horizons. And it is above 0.7 when studying at the 20-year range. Comparing the incoming citation ranks, in turn, reveals an even stronger correlation – with values all above 0.98. We find the 4-digits CPC matrices are less stable – as one might expect given its larger number of nodes – but still quite correlated. For the 10-years window, we measured a correlation rate above 0.6.

Another question one could ponder regarding our graphs concerns their degree distribution. Namely, one might wonder if the in-flow citations show evidence of scale-free networks that are prevalent in many social, biological, and physical systems – including the citations among academic and patent documents (Newman, 2005). Therefore, Figure 2 displays the degree distribution for the 3-digits and 4-digits networks between 1995-2004. The x-axis shows the natural logarithm of in-flow citations to a CPC, and the vertical axis the natural logarithm of their rank.

FIGURE 2 – Innovation Networks Scaling




The figure shows a quasi-linear relationship between the two natural logarithms – as one tends to find when examining scale-free distributions (Newman, 2005). The plots include a formula for their tail distributions. We calculated the tail indices using the Hill estimator (Hill, 1975; Jia, 2018). To do so, we first had to define cutoff points – i.e., where the tails start/end. For the 3-digits CPCs, we employed a maximum cutoff equal to 8; whereas, for the 4-digits, we used the natural logarithm values between 2 and 7. The estimated indices suggest the degree distribution has very heavy tails – perhaps due to the short number of classes/nodes.

Figure 2 also presents the relationship between the number of incoming citations and total patenting volumes per CPC – displayed on the right side of the image. The log-log plot describes a superlinear relationship between the two variables. In other words, the number of citations received by each CPC rises with its total patenting; yet this association is larger than one to one. Therefore, it illustrates a so-called Mathew effect whereby those CPCs with most patent ought to receive more citations and thus produce even more variants.

The scaling analysis displayed in Figure 2.B suggests an increasing dynamics of knowledge specialization (Boschma et al, 2015). It hints the presence of subsystems with self-sustaining growth dynamics, often referred to as autocatalytic sets[1] by the innovation literature (Napolitano et al., 2018). In other words, when a given CPC becomes more prevalent, it grows as an upstream source for those classes closest to it in the network. As such, we expect those CPCs to expand as well. Yet, as these downstream classes increase, it feedbacks itself, the original growing code, and those closest to them. Hence, we expect a, virtuous cycle whereby those closely related technologies enforce one another.

Along these lines, Jain and Krishna (2006) put forward a simple mathematical proof for the presence of autocatalytic sets. Indeed, based on their test, we know that if our graph $W_{ij}$ has an autocatalytic set, then its Perron-Frobenius eigenvalue $\lambda_1(W_{i,j})$ must be larger than zero. So, to show

[1] The term "autocatalytic sets" is often used in social sciences as a metaphor – borrowed from its application in biology and chemistry. We understand the terminology might confuse some, but to keep in line with the related literature on technology evolution, we also employ the phrase here.

the existence of such cliques, we need only to compute the eigenvalues for the 3-digits and 4-digits matrices. And, focusing on the decade between 1995 and 2004, we find that $\lambda_1(W_{i,j})$ is indeed larger than zero, namely 0.0078 and 0.0015. Thus, we may assume the CPC citation network has auto-catalytic sets – which might explain the patterns of regional specialization observed in Europe.

## Predicting Future Patent Volumes

We created a database linking patents grounded on their citations and also connecting the patents to their CPCs. We used this data to plot innovation networks measuring the references flows across CPCs and collected a few initial characteristics regarding these graphs. Nevertheless, our ultimate goal is to use these innovation networks to anticipate patenting growths. To achieve so, we used the method established by Acemoglu et al. (2016) and employed it to the European patents.

The method proposed by the authors consists of three stages. First, we calculate the citation flows across the CPCs for each year passed after the target's invention. Because there are significant differences at the speed at which knowledge diffuses – i.e., how long it takes for a CPC to receive citations – making the innovation networks independently for each year allow us to control for this heterogeneity. Moreover, it supports a more complex understanding of knowledge diffusion, which accounts for the age of invention. Thus, instead of creating the matrices for the 10-year window as before, we can calculate the citation flows as:

$$CiteFlow_{i\to j,a} = \frac{C_{i,j,a}}{Patents_j} \tag{3}$$

where $CiteFlow_{i\to j,a}$ measures the rate of citations from class *i* to *j* at a given diffusion lag $a = [1,10]$.

We computed the $CiteFlow_{i\to j,a}$ variable using the data on citing patents between 1985 and 2004. Next, we employed the citation flows to predict forward patenting for the subsequent period of 2005-2014. To do so, we multiplied the $CiteFlow_{i\to j,a}$ variable by the observed patenting volume ($P_{j,t-a}$) for each CPC within a 10-year window before the focal year – the year we wish to predict. That is, we define the expected number of patents for a given year as:

$$\hat{P}_{i,t} = \sum_{j=1}\sum_{a=1} CiteFlow_{i\to j,a}\, P_{j,t-a} \tag{4}$$

where $\hat{P}_{i,t}$ represents the predicted number of patents belonging to the CPC *i* at the year *t*. And $P_{j,t-a}$ is the actual patent volumes for the class *j* at the year *t* minus the lag *a*. So, when we attempt to predict the patent volumes for the class *i* in 2007, for example, we model it as the average impact from code *j* that occurred with a 7-year diffusion lag from 2000.

Finally, after we calculated forward patenting for each CPC between 2005-2014, we ran a linear regression between these values and the actual observed volumes. That is, to test the quality of our predictions we estimate the following regression:

$$\ln(P_{i,t}) = \beta \ln(\hat{P}_{i,t}) + \phi_i + \eta_t + \epsilon_{i,t} \tag{5}$$

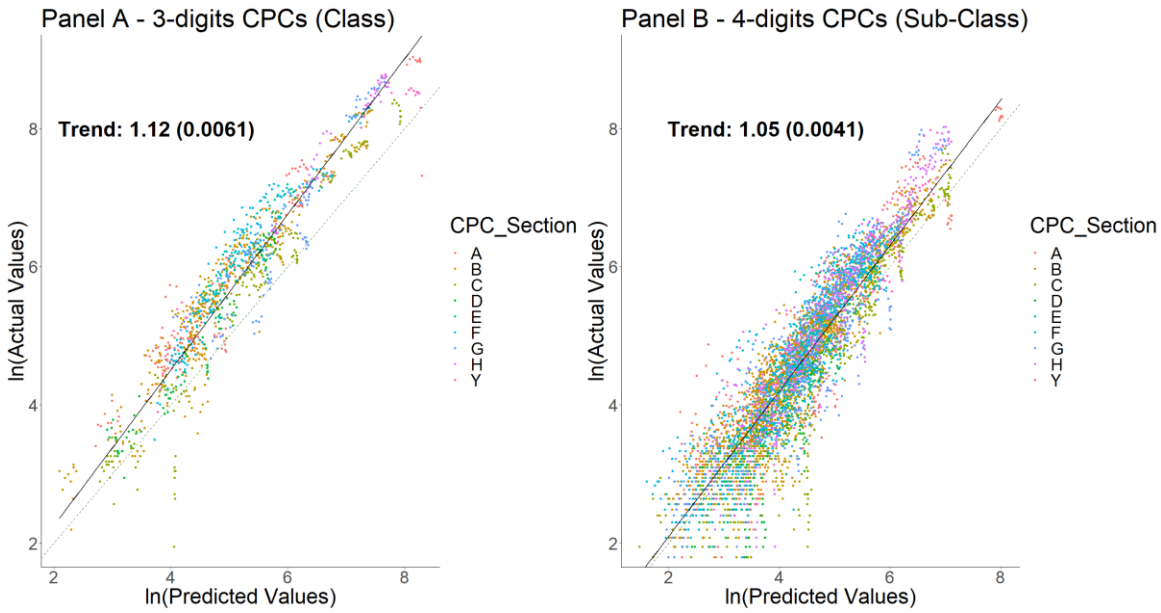
where $P_{i,t}$ and $\hat{P}_{i,t}$ are the actual and predicted patent volumes, respectively. $\phi_i$ is a fixed effect for each patent class, and $\eta_t$ represents a time fixed effect. $\epsilon_{i,t}$ is a disturbance term. The core idea behind the linear regression is to compare how well the model predicts actual rates of innovation by CPC. In other words, $\beta$ informs how actual patenting moves with our predictions. Or, as Acemoglu et al. (2016)

explain: "$\beta$ captures whether the actual patenting in technology $j$ is abnormally high relative to its long-term rate when it is predicted to be so based upon past upstream innovation rates. A $\beta$ estimate of one would indicate a one-to-one relationship between predicted and actual patenting" (p.11486).

# Results

The first step into our analysis is to reproduce the findings by Acemoglu et al. (2016). Along these lines, the plot below shows the estimations obtained from a simple linear regression between the actual patent volumes and its predicted amount for the years 2005-2014. The first column describes the results obtained using the 3-digits CPCs, while the column to the right shows the 4-digits CPCs. The graph plots the relationship between the two variables for all classes with at least five documents per annum. All results include clustered standard errors at the CPC level. The colors of each observation follow its section symbol – viz, one-digit CPC – and we display a 45-degree line by a dotted green line.

FIGURE 3 – Predicted Vs. Actual Patent Volumes, 2005-2014




Our estimations are in line with the earlier outputs from Acemoglu et al. (2016). Like them, we find a strong and statistically significant correlation between our estimated patenting rate for the years 2005 to 2014 and the actual CPC volumes in the period. If we focus on the 3-digits CPCs, for example, we estimate that raising the predicted value by 1% translates into a 1.12% growth of actual patenting. The results for the 4-digits CPCs are even more potent with a nearly one-to-one relationship between the predicted and actual value – we estimated a trend parameter equal to 1.05.

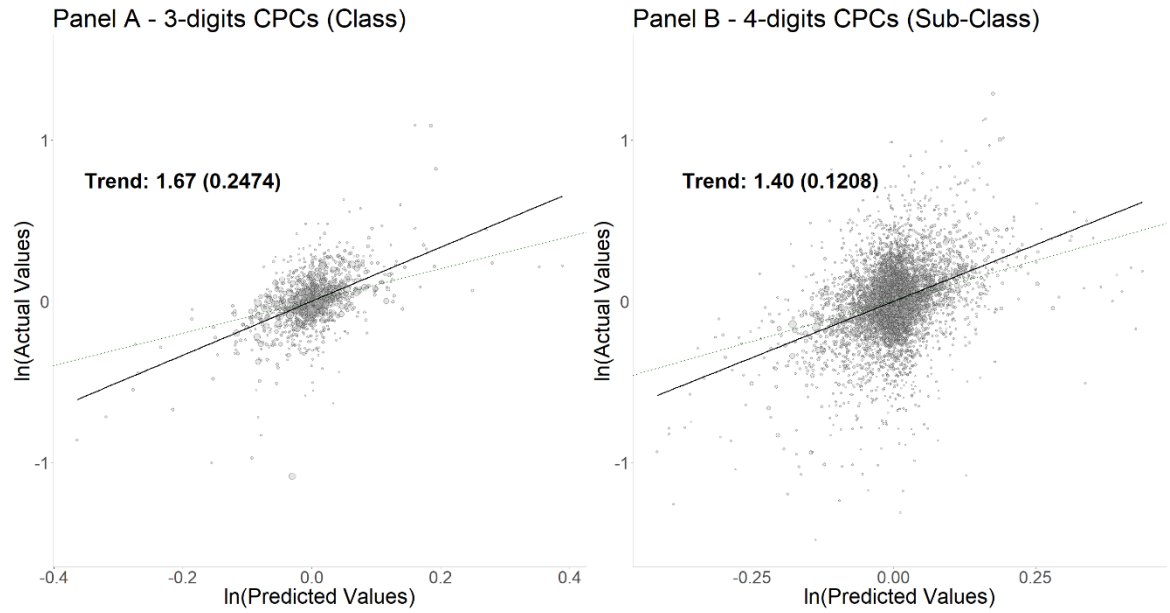
Interestingly, compared to Acemoglu et al. (2016), we seem to underestimate the patenting rates for the European sample. Indeed, they report a relationship between actual and predicted patenting below one (0.85), which suggests their predictions were higher than the real number of patents in the period. In turn, we find that our trend parameter is larger than one – i.e., our predictions were smaller than the actual values. What is more, the absolute difference to one for both our estimates and theirs is nearly identical – we find a 0.12 departure whilst they report a 0.15.

There are likely a few reasons behind this data. First and foremost, we are using different patent classification systems – CPC versus USPC. Thus, we also use distinct hierarchical levels with an unequal number of classes[2]. By the same token, the European Patent Office applies a different set of citation laws that its US counterpart (Bryan et al., 2020). And finally, the two regions have different patenting patterns, practices, and habits. So, we ought to expect systematic differences across the innovation networks for the two places.

Another possible explanation for the contrasting outputs is the time frame under consideration for each analysis. Acemoglu et al. (2016) used the years 1995-2004 to make their predictions. On the other hand, we are focusing on the following decade – the years between 2005-2014. And there was notable growth in the patent application rates post-2005. Hence, because we are using patenting trends before 2004 to calculate the expected number of patents, the change in application rates might bias – underestimate – our results. The same holds for both the number of CPCs and citations for each patent application (Appendix II). Along these lines, if we estimate the same regression but making our predictions for the period 1995-2004, we might find lower than one estimates too[3].

To confirm the results from the earlier regressions, we follow Acemoglu et al. (2016) and estimate a fixed-effect model. Viz, we plot below the results from a weighted linear regression where we first removed the patent codes and year averages from both the predicted and actual values. We employ weights based on patenting per CPC during 1994-2005. As before, we only considered CPCs with at least five patents per year and used clustered standard errors. For reference, we also display a 45-degree dotted line in green, and the size of each observation is proportional to its weight.

FIGURE 4 – Residual Approach, Weighted, 2005-2014




[2] They used both the USPC's subcategories and classes, which hold 36 and 484 codes, respectively. In turn, we used the CPC's classes (127 codes) and subclasses (654).

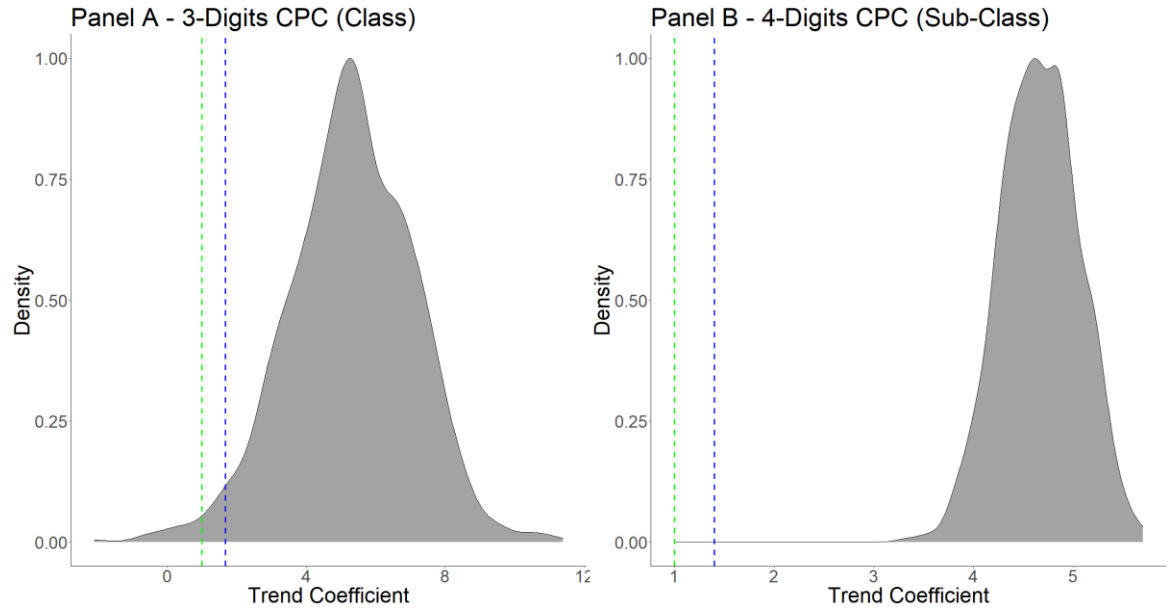
[3] When we estimate a weighted fixed effect regression for the 4-digits CPCs between 1995-2004 – perhaps the closest variation to Acemoglu et al. (2016) preferred model specification – we find a trend coefficient equal to 0.87. And this result is very similar to the one observed by the earlier authors.

Of course, there are other robustness checks one can apply to examine how sound are the results from our regressions – and Acemoglu et al. (2016) offers a few options. Most notably, to measure the importance of the knowledge flows across the CPCs, we could rebuild the estimations while ignoring self-citations[4]. However, though valuable, we feel these tests do not support the narrative we are building. Instead, we decided to follow Pichler et al. (2020) and performed a permutation exercise.

To achieve so, we reproduced our analysis using a randomized version of the innovation networks. For each lag period, we reshuffled the $H_{p,q}$ adjacency matrix linking patents through their citations. We then changed the direction of the reference – we randomly point the edge to another target patent. Then, we recalculate the knowledge flows between the CPCs while using these random copies of the $H_{p,q}$ matrix. Doing so guarantees the overall number of citations stays the same and that larger CPCs will obtain more references due to their sizes. Thus, we ensure that "in the randomized control networks the nodes still have the same weighted outgoing links, but now randomly pointing to other nodes" (Pichler et al., 2020, p.2).

In total, we made 1,000 random copies of the innovation networks using the method elaborated in the above paragraph. And, each time, we used these random graphs to predict patenting volumes for the 2005-2014 period. We then recalculated the weighted fixed-effect regressions and saved the coefficients for each permutation[5]. Below we plot the density distribution for the trend parameters we collected from each random version of the linear regression. We also include a vertical line showing the perfect prediction benchmark – i.e., one. And, in blue, we highlight our empirical estimations.

FIGURE 5 – Density Distribution of the Random Trend Coefficients




[4] We did remake the analysis while ignoring self-citations among the CPCs. And the results – displayed in Appendix III – are in line with the earlier observations. That is, our outcome is robust even when we ignore the citation matrix diagonal.

[5] We also saved the estimations for simple linear regressions. And the results are equally valid. Using the 3-digits CPCs, for example, we find an average random trend parameter equal to 1.2 – with a 0.0003 standard deviation. So, we calculate the random estimations are statistically higher than our real output. But, because these random trends are highly skewed and distant for the actual value, it is hard to visualize the results in a graph. Plus, we consider the weighted fixed effects offer a more robust analysis. Thus, we opted to exclude the outputs from the linear regression randomization test.

Therefore, the estimates collected through the randomization exercise are statistically higher than the real trend parameter we calculated using the innovation networks. Indeed, on average, we find that increasing the random predictions by 1% will represent a near 5% growth in actual patenting. It seems; thus, the random parameters are significantly underestimating the volume of patents in the 2005-2014 period. So, we may conclude the actual results obtained for the time are not a product of chance. That is, indeed, the innovation networks help to anticipate future patenting for Europe.

To conclude, we used the Acemoglu et al. (2016) method and showed it performs just as well for the EPO patents. Further, using the randomization exercise provided by Picher et al. (2020), we demonstrate that the estimations are not a product of chance. In the next section, we wish to develop the current methodology one step further to account for a more complex innovation network.

## Combined CPCs

The methodology developed by Acemoglu et al. (2016) and adopted thus far in this analysis looks exclusively at knowledge flows across individual technological classes. It employs citations between CPC units to capture how related these classes are, and how one draws inspiration from the other. However, it does not account for the complex combinations of technologies within a patent. Each document usually contains multiple CPCs. Likewise, it often cites more than one prior invention and CPCs too. These unique mixtures of classes and citations are the fundamental building blocks that contributed to the development of the patent.

Arthur (2009) describes how technology progresses, evolving by combining existing elements of prior inventions. Along these lines, studying patents and their listed CPCs, empirical data tells that the combination of classes and, in particular, the atypical mix of CPCs are the main drivers of innovation and knowledge complexity (Uzzi et al., 2013; Youn et al., 2014).

Moreover, theoretical work around the NK-model illustrates the importance of accounting for the relationship between each invention's building blocks. Indeed, supporting these models is the core assumption that "technologies can be decomposed into components." And that "components interact with other components" to ultimately determine the technologies costs and productivity (Auerswald et al., 2000; McNerney et al., 2011). In other words, under these premises, the interactions between the patent codes are vital for understanding the selection forces acting upon them. Therefore, ignoring the CPCs combination within a patent might be detrimental to our methodology. And, to improve our model, we ought to account for the CPCs sequences and their citation patterns.

Furthermore, we must equally acknowledge the strong relationship between the classes' combinations and citations. As discussed in the methodology, one can use the CPCs co-occurrence, or the reference flows across them to build innovation networks. And a correlation exercise shows how the two approaches are heavily related. For instance, if we plot both matrices using the patents from 1995-2004, we find a 0.9 correlation coefficient between the two.

In summation, we need to understand the complex interactions between the CPCs and their knowledge flows. Otherwise, we could produce a biased measure of cognitive proximity, or a weak estimation of future patent volumes.

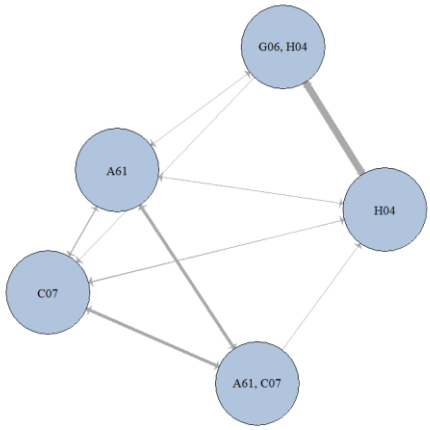
Accounting for the combinations of CPCs within patents may also affect how we look at the knowledge spillovers across codes. First, consider a patent type AB – i.e., a patent containing both codes A and B – which cites an earlier document type A. How should we interpret such citation? Do we make them as a knowledge flow from A to B? Or perhaps we should assume the new patent adds component B to a prior knowledge created by A? Alternatively, say a patent type AC cites a document

AB. Then, how do we understand this scenario? Is the relationship between A and C equivalent to the one observed between B and C? Or, drawing from biology, should we represent the new file as a mutant offspring whereby the B gene was replaced by a C type?

Along these lines, perhaps a better approach is to build the knowledge spillover networks considering not the codes alone but the combined CPCs. That is, we can change the "level of selection" and look for the citation patterns across patent's "genotypes." Here, we graph a directed network where each unique CPC sequence is a node. And whenever one patent "genotype" cites another, we draw an edge between them. We illustrate these combined CPC networks in Figure 6. On the right, we draw an example edge list for the graph containing a few code sequences; while in the right, we plot the network associated with this sample edge list.

FIGURE 6 – The Combined CPCs Network

| Source | Target | Weight |
|---|---|---|
| H04 | G06, H04 | 0.3827 |
| C07 | A61, C07 | 0.1816 |
| A61 | A61, C07 | 0.1673 |
| A61 | C07 | 0.0911 |
| A61, C07 | C07 | 0.0897 |
| G06, H04 | H04 | 0.0325 |
| A61, C07 | A61 | 0.0222 |
| C07 | A61 | 0.0197 |
| A61 | H04 | 0.0009 |
| A61 | G06, H04 | 0.0006 |
| C07 | H04 | 0.0003 |
| H04 | A61 | 0.0003 |
| A61, C07 | H04 | 0.0002 |
| H04 | C07 | 0.0001 |
| G06, H04 | C07 | 0.0000 |
| G06, H04 | A61 | 0.0000 |




*Note:* The table on the right shows an example edge list for the combined CPCs approach. The table shows the real knowledge flows across five CPCs we hand-selected for the period 1995-2004. For simplicity, we are ignoring self-citations. We calculated the weights according to Equation 2 – discussed in the methodology session. The graph, on the right, plots the network generated by this edge list. Both the colours and the size of the arrows depend upon the edge's weight.

We must remark that using the CPC sequences significantly increases the computational power required to perform our analysis. Focusing on the 3-digits CPCs, we find about 46,000 different combinations of codes throughout the period in consideration. Therefore, if we were to construct our analysis around these sequences, we would require a 46,000 x 46,000 citation matrix. And this network would likely be very sparse – most combinations are rare and seldom receive citations.

Still, we proceed to make our predictions using the CPC sequences – instead of looking exclusively to the codes alone. To achieve this, we calculate the $CiteFlow_{i \to j,a}$ variable shown in Equation 3 just like we did before. And we do the same for the predicted and actual patent volumes. Indeed, the unique difference regarding the methodology is that now the subscripts $i$ and $j$ refer to a sequence of codes – e.g., G06-H04 or A61-C07.

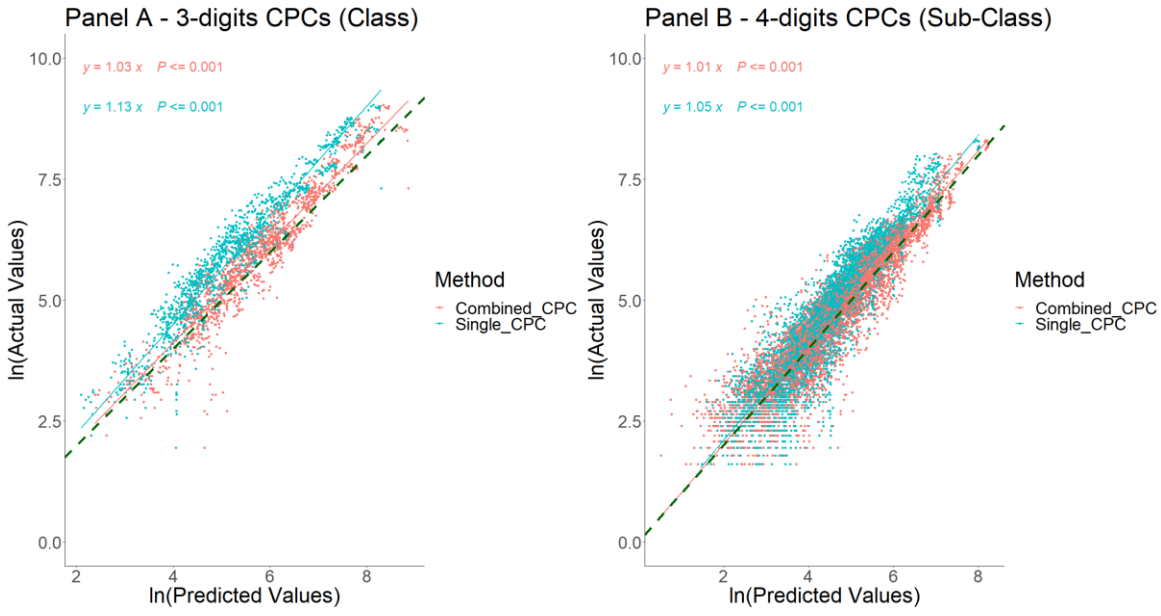
The methodology used for the single codes can perform just as well for the combined CPCs. Indeed, looking at the 3-digits CPCs, for example, a simple linear regression shows that a 10% growth

in the number of predicted patents for each CPC sequence translates into a 12.2% rise in the actual number of documents. So, the innovation networks method developed by Acemoglu et al. (2016) also predicts the combined CPCs. It can anticipate the growth of particular combinations. Although, because it only judges existing sequences, it does not estimate the chances of new CPC mix.

More importantly, we wish to test how the combined method compares to the discrete CPC version. It is hard to compare the two regressions on different samples, so we need first to disaggregate the predictions for the combined approach. Namely, after we estimate the number of patents for each independent CPC sequence, we divide them into solo codes and compute their shares. So, we split all patents type "G06-H04" into their two components and attribute a 0.5 share for both CPCs – i.e., G06 and H04. Then, we add the shares of each CPC and reproduce the initial analysis performed on the individual patent classes.

The graph below displays the comparison between the two methods. We show, in blue, the correlation between the natural logarithm of the actual values and predicted ones using the single CPC method. And, in red, we highlight the same correlation using the novel combined-CPC approach. The figure plots on the left a simple linear regression between the two variables using the 3-digits CPCs, whereas the one to the right shows the outputs from the 4-digits CPCs. Like before, we included in the analysis only those CPCs with at least five documents per annum. We also added a 45-degree dashed green line for reference.

FIGURE 7 – Combined CPC vs. Single CPC Approach, 2005-2014




As one may conclude from Figure 7, the combined CPC method outperformed the original regression. Indeed, it provides a nearly perfect prediction for both the 3-digits and 4-digits CPCs – with an estimated trend parameter equal to 1.03 and 1.01, respectively.

The results, in turn, highlight the importance of accounting for the complex combination of CPCs within patents. It shows that we can improve the predictions of the model when we account for their interactions – as the theoretical NK-models already suggested.

## Regions

Another step forward from Acemoglu et al. (2016) is to examine how the innovation network predicts the technological development of different regions. Indeed, the authors conclude their examination by commenting on how they "believe that this approach can be pushed to consider regional variation and firm-level variation, which can further help us understand the causal impact of patenting on economic and business outcomes" (p.11487).

Evolutionary Economic Geography studies the dynamics of knowledge creation across places. A chief concern of this literature is to examine patterns of regional diversification and specialization. And, to our knowledge, no previous model has sought to use citation flows to anticipate the paths of knowledge production within regions.

Kogler et al. (2017) used the co-occurrence of CPCs within patents to map the evolution of specialization in the EU15 regions. They observed changes in time and space regarding the "average knowledge relatedness" – how places become more or less specialized in comparison to others. Thus, we decided to employ the innovation networks method proposed by Acemoglu et al. (2016) to estimate the future growth in patenting activity for each CPC at different NUTS2 regions of Europe. That is, we sought to test if the same method adopted thus far can also anticipate the dynamics of local knowledge creation demonstrated by Kogler et al. (2017).

We wish to adapt the methodology proposed by Acemoglu et al. (2016) to estimate regional patent development between 2005-2014. The most straightforward alternative is to assume the innovation networks – which maps the citation patterns across the CPCs – are universal. The graphs capture a fundamental and global relationship between the classes, which does not depend on regional capabilities. Nonetheless, the local patent portfolio affects the ability of each NUTS2 to engage and to implement the knowledge spillovers among the patent classes. That is, though the association between downstream and upstream CPCs is constant, a region lacking the upstream potential will unlikely develop many products on either code.

Although preserving the innovation network constant might seem like a strong assumption, we highlight it is in line with previous research done on regional specialization. Indeed, when measuring local average knowledge relatedness, one usually computes the relatedness degree between CPCs using the global patenting portfolio. Then, they look at the CPCs at the regional patents to find the average relatedness classes (Whittle & Kogler, 2020).

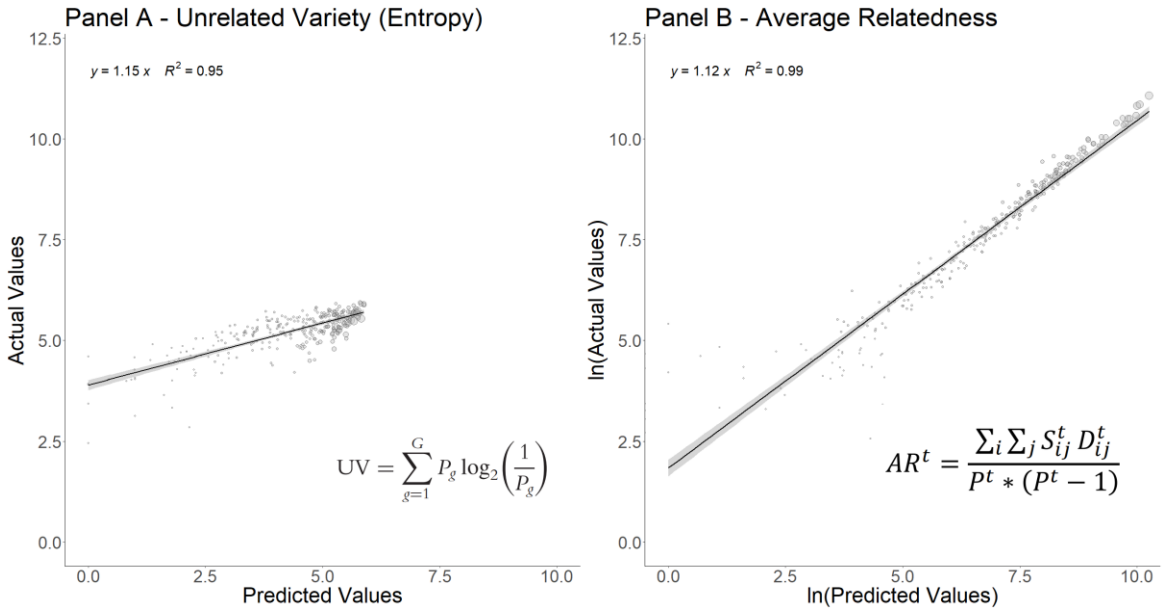
In mathematical terms, we express the assumption above by keeping the $CiteFlow_{i\rightarrow j,a}$ matrices in Equation 4 constant while allowing the patenting vector $P_{j,t-a}$ to change over the different NUTS2. That is, we can readily adjust Equation 4 to estimate the future number of patents in a given region and CPC. All it takes is to multiply the constant $CiteFlow_{i\rightarrow j,a}$ variable by a local patenting vector like $P_{j,r,t-a}$ – where the subscript $r$ represents the NUTS2.

After estimating the patent volumes between 2005-2014 for each CPC and NUTS2, we test how well the model describes regional patenting. In the appendix, we show the estimations from a weighted fixed-effect regression between the logarithms of predicted and actual CPC amounts using the complete European sample – i.e., we include all NUTS2 observations in the regression model. Once more, the results are significant and robust. That is, the method developed by Acemoglu et al. (2016) also explains regional patenting trends.

Besides, we can test whether the innovation networks model can predict local diversification and specialization. The EEG literature offers numerous measures for these attributes and – using the available tools from the relevant literature – we calculated two estimates of the regional knowledge expertise. First, we employed the Shannon Entropy formula to the real and predicted incidence of CPCs in a region between 2005-2014 to measure how diverse they are – or to estimate their Unrelated Variety (Frenken et al., 2010). Next, we calculated the degree of relatedness across all CPCs for the period and gathered the NUTS2 average relatedness score[6]. Balland (2017) offers a concise review of these two methods and how to compute them.

Figure 8 demonstrates the correlation between the predicted and actual scores gathered for the two metrics. In these graphs, each point represents a NUTS2 region, and its size is proportional to the total number of patents produced between 2005-2014. We also included the formulas used to compute these variables – as described in the relevant literature. Again, the innovation networks model proves accurate at forecasting both regional specialization and diversification – at least, as measured by these descriptive statistics from the EEG literature[7].

FIGURE 8 – Regional Diversity and Specialization, 3-Digits CPC, 2005-2014




*Note:* The equations, displayed in the bottom, describe how we calculated each of these variables. We collected them from Franken et al. (2007) and Kogler et al. (2015) - respectively.

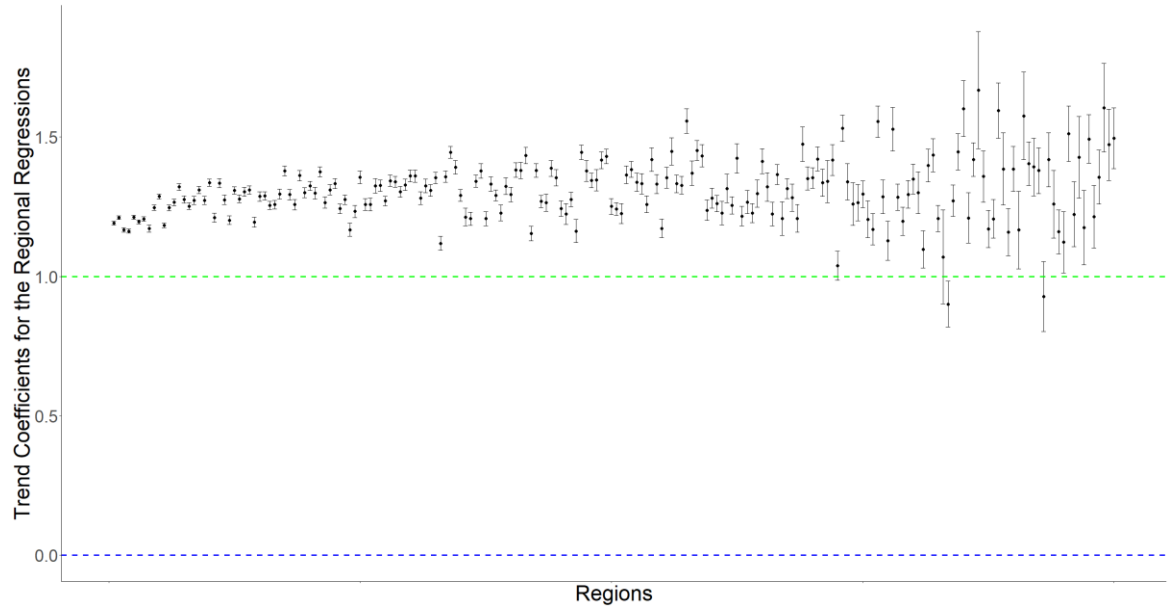
Still, we wish to understand how well the model predicts the patenting development for each NUTS2 region separately. To this end, we estimate a regression model for each NUTS2 individually. We sampled our final dataset to include only the predicted and actual values for one region; then we reproduced our linear regression using these subsamples. Given the smaller number of patents per place, the NUTS2 regressions include only those codes with more than one document per annum.

[6] To calculate average relatedness, first we estimate how related the CPCs are. To do so, we used the complete European sample data for the period 2005-2014. And we calculated relatedness based on the real cooccurrence of CPCs in all patents. Then, we used the predicted and real incidence of the CPCs in each region to find the NUTS2 average relatedness scores.

[7] We picked these two metrics because we find them to be the most prevalent in the literature. However, we could reproduce this analysis using other alternative variables from the EEG literature. Indeed, we made a similar regression for the regional Diversity scores – a variable capturing the number of 'industries' where a NUTS2 has a relative comparative advantage (RCA).

Figure 9 plots the outputs from the OLS regression using 3-digits CPCs and the regional subsamples. The line segments represent the confidence intervals, and we ordered them according to total patenting in the region between 1995-2004. As such, starting to the left, we have the place with most patents in the decade – Ile de France. We only show the results for those regions with more than 150 patents filed in that period. Thus, we plot the results for 200 NUTS2 only. We include a blue horizontal line to represent the null hypothesis suggesting the model has no predictive power over the NUTS2 patenting. For reference, we also added a green line across the value of one.

FIGURE 9 – Regional Coefficients, Linear Regression, 3-Digits CPCs, 2005-2014




Hence, the innovation networks approach does rather well when predicting future patenting for most regions. Nevertheless, the results differ significantly across places. While for some NUTS2 (e.g., those in the left-hand side of Figure 9) the model does as good as using the entire European patent population, for others, the methodology has no significant economic value. Furthermore, it seems the model is particularly efficient for predicting the development of those regions with most patents in the 1995-2004 period. Looking at Figure 9, it becomes clear how, as we move towards those places with fewer patents, our estimations worsens significantly.

These results are perhaps not surprising given those regions have more CPCs and patents to serve as the upstream source of innovation. Since we keep the innovation matrices constant, the difference between our regional estimates comes from their patenting volumes in the period before 2005-2014. And needless to say, those places with more patents per class will by definition observe larger effects when predicting patenting at the period in hand. Thus, we ought to have better predictions for those places with more upstream resources.

Another possible explanation is that those laggard NUTS – those with fewer patents to begin with – are dependent on knowledge developed elsewhere. Therefore, one might propose that smaller places like Dublin (IE02) are importing knowledge from other NUTS2 at a higher rate. So, to model the dynamics of CPCs in those places, we must account as well for the informational flows across NUTS2. In other words, our current regional model is not adequate to represent these laggard areas. And we need to implement a more realistic model.

So, we propose another method to predict regional patenting where we account for spillover across places. We still hold the innovation matrix constant, but we add an effect – a second matrix accounting for the citation flows in a given CPC across space. We do so by modifying Equation 4 as:

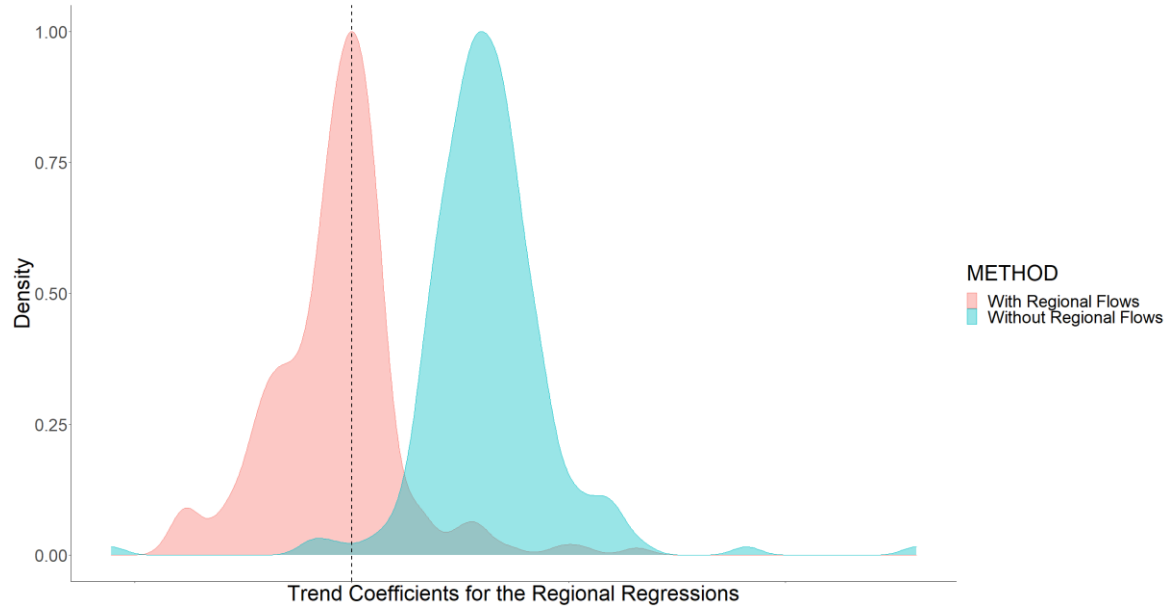
$$\hat{P}_{i,r,t} = \sum_{j=1}\sum_{a=1} CiteFlow_{i\to j,a}\, P_{j,t-a} + \sum_{r\prime\neq r}\sum_{a=1} CiteFlow_{i,r\to r\prime,a}\, P_{ir\prime,t-a} \qquad (5)$$

where $\hat{P}_{i,r,t}$ is a vector containing the predicted patenting values for each CPC in the region *r*. And $CiteFlow_{i,r\to r\prime,a}$ compute, for each CPC, the rate of citation from region *r* to *r'* at the lag period *a*[8]. Here, we ignore self-citations. That is, we do not account for the citations within the same regions – as these are already represented in the first summation term. As before, $P_{i,r\prime_{,},t-a}$ is a vector holding the number of patents in class *i* at region *r'* and time *t-a*. The rest of the equation is identical to the one used to make our regional predictions.

Adding the regional flows improves the quality of our predictions considerably. Indeed, for 199 NUTS2 in our sample, we collect better predictions once we include knowledge flows across regions – measured as the absolute distance to the perfect prediction benchmark.

To visualize how these two approaches to predict future regional patenting compare, in Figure 10, we plot the density distribution for their slope coefficients. In blue, we display the distribution from the original method – without citation flows across NUTS2. And in red, we have the same distribution for the new approach. As one can notice, the predictions made with the regional knowledge flows are more skewed towards one – shown as a dashed line.

FIGURE 10 – Regional Coefficients Distribution, 3-Digits CPCs, 2005-2014




[8] We build the knowledge flows matrix across NUTS2 like the original innovation networks. That is, we first create a graph of citing patents and another bi-partite network linking the documents to places. Because each patent can have more than one inventor residing in different NUTS2, we had to row-normalize the patent-NUTS graph too. At last, we obtain the regional knowledge flow matrix by using the same algebraical projection described in Equation 1.

As a conclusion, the innovation-networks model performs well for the NUTS2 regions of Europe. However, there is room for improvement. For example, we show that accounting for regional spillovers could lead to better estimations for most places. By the same token, there are other reasonable means to improve how we predict the local patenting dynamics. For example, one could formulate a regional selection model, in which the citation flow matrix varies according to the place. Likewise, using the combined CPC method described in the previous section, one could improve our regional estimates too. Nonetheless, these are beyond the scope of the current report.

# Discussion

In this essay, we tested the innovation networks method created by Acemoglu et al. (2016). We applied it to a new data sample consisting of patents filed with the EPO between 2005-2014. And we showed how the model indeed performs rather well in this new context. Perhaps surprisingly so, given the different citation laws across Europe and the US, as well as general differences in patenting across the two regions and the distinct classification systems.

Moreover, we sought to strengthen the original method devised by the authors and provided two alternative approaches for measuring and forecasting patent progress. First, we considered how the CPCs interact when they belong to the same document. We showed that accounting for the CPCs combinations improves the quality of our predictions.

We also tried to develop a method to estimate patenting development for the regions of Europe. And, most importantly, we demonstrated that incorporating citation flows across places to the original model improves our estimations for the NUTS2. Altogether, we find the innovation networks model and its variants performed well. Further, it provided a viable path to study the dynamics of knowledge creation, how some technologies become more popular with time and regional specialization.

Nevertheless, the current method is not perfect. And the quality of its predictions changes significantly depending on time, region, and CPC. Hence, we expect that future research will help us understand what makes the model better suited for some NUTS2 and classes. Perhaps there are intrinsic features such as how diverse a region is at the starting period that would improve our analysis. Or the model is possibly particularly suited to forecasting more developed technologies – those with more patents and established spillover channels. Investigating these questions will enhance how we can model the selection of CPCs, provide better insights into patenting growth, and better predictions.

Our ultimate goal is to understand and accurately forecast how the CPCs become more/less popular with time. We seek, thus, to model the selection dynamics of these patent codes. As such, prior evolutionary models provide the ideal inspiration for us to build robust estimates for the changes in CPCs popularity. It can serve as a valuable source of knowledge spillover to enhance the already good literature developed after Acemoglu et al. (2016).

Evolutionary models on the economics of innovation most often focus on the creation of new goods. Seldom do they examine how the shares of a fixed class of products change with time. That is, while the prior literature emphasized the emergence of new technologies from the existing body of knowledge, it mostly overlooked the frequency-dependent dynamic of technology evolution – or why some technologies grow more popular than others.

So, to our knowledge, the literature still lacks a robust model which appraises how the frequency and relatedness between patent classes influence the future shares of these CPCs across regions. And we reckon that such a model would prove fundamental for us to better understand and predict patenting development for the European regions.

We defend, therefore, a dynamic model where technologies compete for attention. Based on an evolutionary model whereby those patent classes collecting most resource ought to output more patent variants. Along these lines, we suspect that evolutionary game theory (Nowak, 2006) proves the ideal means to embody said evolutionary dynamics. Indeed, the innovation networks put forward by Acemoglu et al. (2016) is remarkably similar to the replicator dynamics – often used by these game theory models. As such, it seems that both methodologies are highly compatible and complementary. It seems the replicator dynamics is a suitable candidate for modelling the CPCs selection.

However, we must also acknowledge the shortcomings of these models. The competition for survival which is their central focus will eventually lead to a stable equilibrium where regions specialize in a single CPC – the one with most patents at the initial step. That is, the replicator dynamics is perhaps not suited to examine the more complex relationship between patent classes – which compete for attention, but also combine and cooperate to produce new technologies. For this reason, in the past, authors used the replicator approach only to understand the competition between alternative standards of the same technology – when typically one method comes out as dominant (Arthur, 1989; Heinrich, 2018; Kim et al., 2018). Yet, we hope that recent advances on evolutionary models which include structured populations and network effects (Santos et al., 2008; Tarnita et al., 2009; Canter et al., 2019) could prove efficient means to model the competition, combination, and selection of CPCs.

Furthermore, we consider the outputs from this report as suggesting viable paths to develop such a robust model of technology selection. Take, for instance, our results using the combined CPCs. It highlights that one must account for the CPCs interactions when proposing a new model of selection. Therefore, we suggest using a "multi-level" selection model as a possible alternative that reflects this insight. We imagine a scenario where the CPCs must combine to produce new products and their fitness depends, at least partially, on how likely they can match with others. Still, the arranged CPC sequences also compete for resources. Perhaps, one might assume the least common combinations are the ones with the most potential for development, and they have a higher expected payoff – i.e., we can define the combined CPCs payoff as a decaying function of popularity. The fitness of each CPC would depend on how likely they can mix with others and how frequent these combinations are. Thus, we could employ the similarity-popularity plane developed by Tacchella et al. (2020) to estimate the CPCs expected fitness and simulate its selection.

Alternatively, one can choose to ignore the atomic representation of the CPCs and propose a selection model focused on its sequences. That is, embracing the patent "genotype" analogy, one can employ a variation of the mutator-replicator-dynamics model (Nowak, 2006) to study how the combined CPCs sequences become more or less popular. To do so, one could use the number of citations collected by each combination to define its fitness, its number of offspring. Further, one could use the citation flows across CPCs to estimate mutation rates – how likely one sequence type produces a mutant of a different kind. Or, one could employ the "genotype-phenotype" maps used by Woodard and Clemons (2014) to construct a fitness landscape for the CPC sequences and use this map along with the mutation rates to study the CPCs evolution.

The outputs for the NUTS2 regions also provide a path forward for modelling selection. Namely, we showed that accounting for knowledge spillovers across places can improve our analysis. Thus, one might use this insight to propose a metapopulations model to study the evolution of CPC populations

(Day & Possingham, 1995). In other words, we can structure the CPCs into spatially separated groups, which nonetheless interact through the permanent migration between them – representing here by the citation flows across the NUTS2. Again, we hope this model would provide valuable insights into how technological classes grow more or less popular with time. Foremost, it would highlight the role of regions on the evolutionary dynamics of patent classes.

In summary, this report shows the power of using selection dynamics to predict future patenting. Moreover, our modifications of the Acemoglu et al. (2016) model not only improved our estimations but also led us to model the selection process more rigorously. Along these lines, we reckon these models could provide us with valuable insights into knowledge creation and specialization. These insights may enable us to study the consequences of innovation, or to enhance how we appraise and formulate technology policies.

As Acemoglu et al. (2016) recognize, the selection dynamics model could be used to identify regional patenting trends. That is, one could employ it to spot variations in technological diffusion across time and space – which, in turn, we can use to test the impact of patenting on wellbeing. For example, prior literature questions whether knowledge relatedness matters for the "resilience of regional economies to exogenous shocks" (Rocchetta et al., 2019, p.1421). The innovation networks method could assist in finding patterns of specialization and thus supporting their argument.

We think the innovation networks can also help us to estimate the impact of government policies seeking to improve a region's core knowledge on a given technology class. Suppose the local government finds a CPC particularly important for its future and therefore promotes a policy seeking to strengthen the regional advantage in that CPC. Using the innovation networks, one can estimate how the local patenting portfolio would have developed in the absence of the policy. That is, we can measure a counterfactual similar to a synthetic control (Abadie et al., 2010). And we can compare the regional growth in that CPC to what we expected based on its anterior patenting. Thus, we can employ the innovation networks method to test the influence of the policy on local diversification.

These networks can also inform policymakers in the development of new plans. The EU focuses much attention and resources on the so-called Smart Specialization Strategies. Nonetheless, to achieve these goals, one needs first to estimate how well the region's core competencies suits. the development of a given industry. To produce smart specialization, one requires reliable measures of "technology embeddedness" (Buarque et al., 2020) – which are still sparse in the literature. Hence, a natural question is: can we use the selection dynamics approach to test which regions have the core competence to develop a particular industry? Or can we use the methods to examine which classes are most suited for investing in a given NUTS2? Can the innovation networks method also be used to inform and enhance smart-specialization policies? We hope our future research will be able to answer these questions as well.

# References


1. Abadie, A., Diamond, A., & Hainmueller, J. (2010). Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program. *Journal of the American statistical Association*, 105(490), 493–505.

2. Acemoglu, D., Akcigit, U., & Kerr, W. R. (2016). Innovation network. *Proceedings of the National Academy of Sciences*, 113(41), 11483–11488.

3. Alstott, J., Triulzi, G., Yan, B., & Luo, J. (2017). Mapping technology space by normalizing patent networks. *Scientometrics*, 110(1), 443–479.

4. Arthur, W. B. (1989). Competing technologies, increasing returns, and lock-in by historical events. *The Economic Journal*, 99(394), 116–131.

5. Arthur, W. B., & Polak, W. (2006). The evolution of technology within a simple computer model. *Complexity*, 11(5), 23–31.

6. Arthur, W. B. (2009). *The nature of technology: What it is and how it evolves.* Simon and Schuster.

7. Auerswald, P. Kauffman, S., Lobo, J. & Shell, K. (2000). The production recipes approach to modelling technological innovation: An application to learning by doing. *Journal of Economic Dynamics and Control,* 24(3), 389–450.

8. Balland, P. A. (2017). Economic Geography in R: Introduction to the EconGeo package. *Available at SSRN 2962146*.

9. Boschma, R., Balland, P. A., & Kogler, D. F. (2015). Relatedness and technological change in cities: the rise and fall of technological knowledge in US metropolitan areas from 1981 to 2010. *Industrial and Corporate Change*, 24(1), 223–250.

10. Bryan, K. A., Ozcan, Y., & Sampat, B. (2020). In-text patent citations: A user's guide. *Research Policy*, 49(4), 103946.

11. Buarque, B. S., Davies, R. B., Hynes, R. M., & Kogler, D. F. (2020). OK Computer: the creation and integration of AI in Europe. *Cambridge Journal of Regions, Economy and Society*, 13(1), 175–192.

12. Canter, U., Savin, I., & Vannuccini, S. (2019). Replicator dynamics in value chains. *Industrial and Corporate Change,* 28(3), 589–611

13. Day, J. R., & Possingham, H. P. (1995). A stochastic metapopulation model with variability in patch size and position. *Theoretical Population Biology*, 48(3), 333–360.

14. Frenken, K., Van Oort, F., & Verburg, T. (2007). Related variety, unrelated variety, and regional economic growth. *Regional Studies*, 41(5), 685–697.

15. Heinrich, T. (2018). Network externalities and compatibility among standards: A replicator dynamics analysis. *Computational Economics,* 52(3), 809–837.

16. Hidalgo, C. A., Klinger, B., Barabási, A. L., & Hausmann, R. (2007). The product space conditions the development of nations. *Science*, 317(5837), 482–487.

17. Hill, B. M. (1975). A simple general approach to inference about the tail of a distribution. *The Annals of Statistics*, 1163–1174.

18. Jain, S. and Krishna, S. (2005). *Graph theory and the evolution of autocatalytic networks*. In Handbook of Graphs and Networks (eds S. Bornholdt and H.G. Schuster).

19. Jara-Figueroa, C., Jun, B., Glaeser, E. L., & Hidalgo, C. A. (2018). The role of industry-specific, occupation-specific, and location-specific knowledge in the growth and survival of new firms. *Proceedings of the National Academy of Sciences*, 115(50), 12646–12653.

20. Jia, M., Taufer, E., & Dickson, M. M. (2018). Semi-parametric regression estimation of the tail index. *Electronic Journal of Statistics*, 12(1), 224–248.

21. Kauffman, S. A. (2000). *Investigations*. Oxford University Press.

22. Kim, K., Jung, S., Hwang, J. & Hong, A. (2018). A dynamic framework for analyzing technology standardization using network analysis and game theory. *Technology Analysis & Strategic Management,* 30(5), 540–555.

23. Knuepling, L., & Broekel, T. (2020). Does relatedness drive the diversification of countries' success in sports? *European Sport Management Quarterly*, 1–23.

24. Kogler, D. F., Rigby, D. L., & Tucker, I. (2013). Mapping knowledge space and technological relatedness in US cities. *European Planning Studies*, 21(9), 1374–1391.

25. Kogler, D. F., Essletzbichler, J., & Rigby, D. L. (2017). The evolution of specialization in the EU15 knowledge space*. Journal of Economic Geography*, 17(2), 345–373.

26. Korhonen, J., & Kasmire, J. (2013). *Adder: a simplified model for simulating the search for innovations*. Working paper presented at druid.

27. Leydesdorff, L., Kogler, D. F., & Yan, B. (2017). Mapping patent classifications: portfolio and statistical analysis, and the comparison of strengths and weaknesses. *Scientometrics*, 112(3), 1573–1591.

28. Lucas, R. (1988) On the mechanics of economic development, *Journal of Monetary Economics,* 22, 3–39.

29. McNerney, J., Farmer, J. D., Redner, S. & Trancik, J. E. (2011). Role of design complexity in technology improvement. *Proceedings of the National Academy of Sciences*, 108(22), 9008–9013.

30. Napolitano, L., Evangelou, E., Pugliese, E., Zeppini, P., & Room, G. (2018). Technology networks: the autocatalytic origins of innovation. *Royal Society Open Science,* 5(6), 172445.

31. Newman, M. E. (2005). Power laws, Pareto distributions and Zipf's law. *Contemporary Physics*, 46(5), 323–351.

32. Nooteboom, B. (2000). Learning by interaction: absorptive capacity, cognitive distance, and governance. *Journal of Management and Governance*, 4(1-2), 69–92.

33. Nowak, M. A. (2006). *Evolutionary dynamics: exploring the equations of life*. Harvard university press.

34. Pichler, A., Lafond, F., & Farmer, J. D. (2020). Technological interdependencies predict innovation dynamics. *Technological Interdependencies Predict Innovation Dynamics (March 2, 2020)*.

35. Rocchetta, S., & Mina, A. (2019). Technological coherence and the adaptive resilience of regional economies. *Regional Studies*, 53(10), 1421–1434.

36. Santos, F. C., Santos, M. D., & Pacheco, J. M. (2008). Social diversity promotes the emergence of cooperation in public goods games. *Nature*, 454(7201), 213–216.

37. Silverberg, G. (2002). The discrete charm of the bourgeoisie: quantum and continuous perspectives on innovation and growth. *Research Policy*, 31(8-9), 1275–1289.

38. Silverberg, G. & Verspagen, B. (2005). A percolation model of innovation in complex technology spaces. *Journal of Economic Dynamics and Control,* 29(1-2), 225–244.

39. Silverberg, G., & Verspagen, B. (2007). Self-organization of R&D search in complex technology spaces. *Journal of economic interaction and coordination*, 2(2), 211–229.

40. Tacchella, A., Napoletano, A., & Pietronero, L. (2020). The Language of Innovation. *PloS one*, 15(4), e0230107.

41. Tarnita, C. E., Antal, T., Ohtsuki, H., & Nowak, M. A. (2009). Evolutionary dynamics in set structured populations. *Proceedings of the National Academy of Sciences*, 106(21), 8601–8604.

42. Uzzi, B., Mukherjee, S., Stringer, M., & Jones, B. (2013). Atypical combinations and scientific impact. *Science*, 342(6157), 468–472.

43. Van Dam, A. & Frenken, K. (2020). Variety, complexity, and economic development. *Research Policy,* 103949*.*

44. Whittle, A., & Kogler, D. F. (2020). Related to what? Reviewing the literature on technological relatedness: Where we are now and where can we go? *Papers in Regional Science*, *99*(1), 97–113.

45. Woodard, C. J. & Clemons, E. K. (2012, July). Modelling technology evolution using generalized genotype-phenotype maps. In: *Proceedings of the 14th annual conference companion on Genetic and Evolutionary Computation* (p.p.323–330)*.*

46. Yoon, B., & Magee, C. L. (2018). Exploring technology opportunities by visualizing patent information based on generative topographic mapping and link prediction. *Technological Forecasting and Social Change*, 132, 105–117.

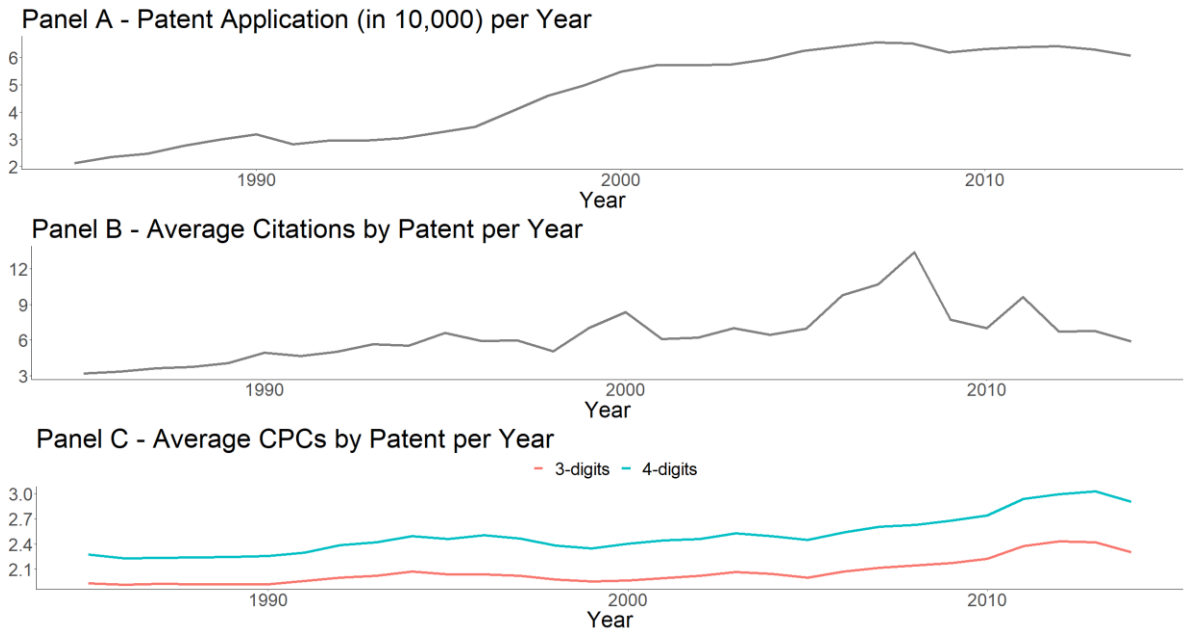
47. Youn, H., Strumsky, D., Bettencourt, L. M., & Lobo, J. (2015). Invention as a combinatorial process: evidence from US patents. *Journal of the Royal Society interface*, 12(106), 20150272.

# Appendix

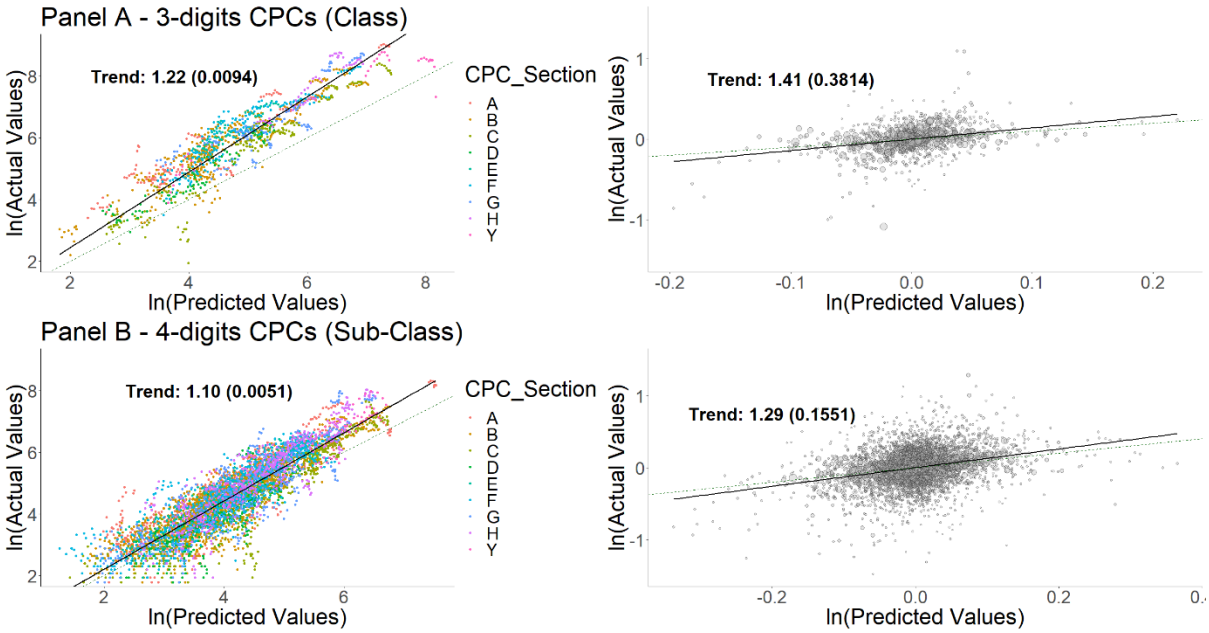
## Appendix I – Time Correlation of the Innovation Matrices


*Note:* The graphs display the correlation scores between the three decades under consideration. Period 1 contains the years between 1985 and 1994. Period 2 covers the decade between 1995 and 2004. And Period 3 refers to the 2005-2014 period.

## Appendix II – Patent Application, Citation, and CPCs over Time

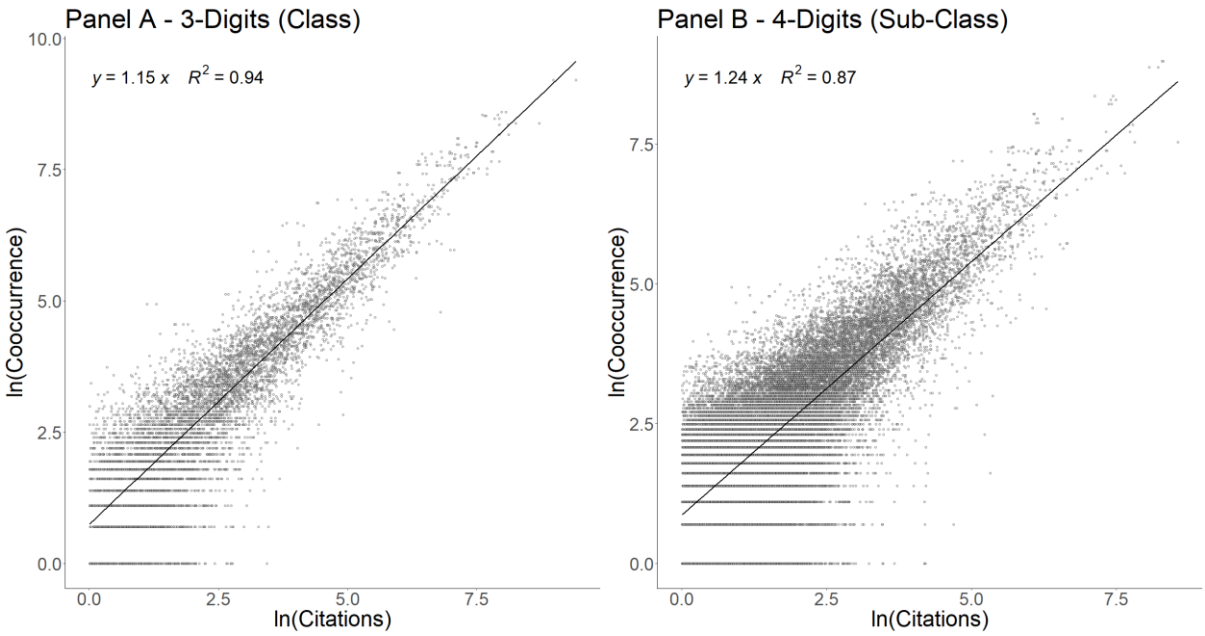


## Appendix III – Predicted Vs. Actual Patent Volumes, External Network, 2005-2014




*Note:* The figure displays outputs from linear regressions. We calculated predicted patenting for the period using the external network only. That is, for this exercise, we ignored self-citations when estimating the volume of patents for the 2005-2014 period. The first column shows the results from a simple linear regression. Whereas the second column shows the outputs from a weighted fixed-effect model - we removed both year and CPCs averages. The weights represent the total patenting per class in the 1995-2004 period. All graphs include clustered standard errors at the CPC level. And a 45-degree dotted line for reference. We only include those classes with at least five patents per annum.

## Appendix IV – Citation and Cooccurrence Between CPCs, 1995-2004



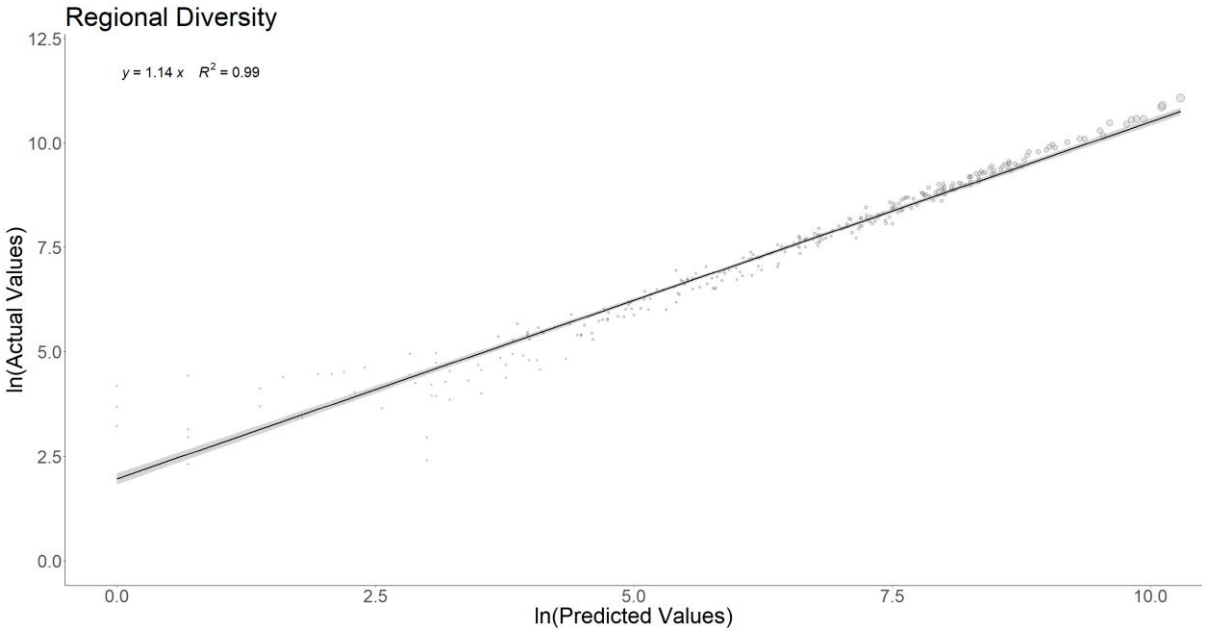

*Note:* The figure shows the correlation between the number of citations and the co-occurrence of CPCs during the 1995-2004 period. Each point represents a CPC pair. The y-axis measures how often these pairs show in the same patent file during the period, while the x-axis measures the citation flows between the two - as measured by the $C_{i,j}$ matrix. The graph highlights how these two measures are strongly correlated.

## Appendix V – Predicted Vs. Actual Patent Volumes, Regions, 2005-2014

| | **Without Regional Flows** | | **With Regional Flows** | |
|---|---|---|---|---|
| Trend Parameter | 1.20*** | 1.46*** | 1.02*** | 1.37*** |
| | (0.0234) | (0.0413) | (0.0105) | (0.0235) |
| Year Fixed Effect | NO | YES | NO | YES |
| CPC Fixed Effect | NO | YES | NO | YES |
| NUTS2 Fixed Effect | NO | YES | NO | YES |
| Adj. R-Squared | 0.97 | 0.99 | 0.98 | 0.99 |
| No. OBS | 66,182 | 66,182 | 66,182 | 66,182 |

*Note:* The table shows the outputs from weighted linear regressions where the independent variable is the natural logarithm of predicted patent volumes per CPC and NUTS2. The dependent variable is the actual number of patents observed in each region and code. The table includes data on all NUTS2-CPC pairs with at least one patent per annum. The weights are the number of patents in 1994-2005. "Without Regional Flows" means that we did not include citation flows across regions when predicting the patent volumes per CPC in each NUTS2. "With Regional Flows," in turn, accounts for the citation patterns across places. The first column of each group shows the results from simple regressions, whereas the second column includes fixed effects for year, patent code, and regions. All the estimations used clustered standard errors at the CPC level. *** $p < 0.01$.

## Appendix VI – Regional Diversity, 3-Digits CPC, 2005-2014




*Note:* The graph shows the correlation between actual and predicted regional diversity for the 2005-2014 period. "Regional Diversity" is a count variable which measures the number of patent classes where the NUTS2 has a relative comparative advantage (Balland, 2017).