ERASMUS UNIVERSITY ROTTERDAM
Erasmus School of Economics

Master thesis Industrial Dynamics & Strategy

Regional related and unrelated variety and the innovative performance of European NUTS-2 regions

Abstract:
This Master thesis presents a scientific replication of Castaldi, Frenken and Los (2015). The hypotheses that related variety is positively associated with innovative output, and that unrelated variety is positively associated with breakthrough innovations are confirmed using data on European NUTS-2 regions. Furthermore, the thesis adds to the related-unrelated variety literature by explicitly considering the difference between technological and industrial variety.

Name student: H.H.E. (Hanne) Oelen
Student ID number: 483691ho

Supervisor: N. Cortinovis
Second assessor: F. van Oort

Date final version: October 31, 2018

1. Introduction

Smart specialisation is the magic word in contemporary European regional policies. It refers to a regional innovation strategy that prioritizes research and innovation investment in the areas for which the region possesses a competitive advantage. As part of the European Commission's Cohesion Policy, regions must develop a Research and Innovation strategy for Smart Specialisation in order to receive structural funding. The aim of the European Commission's Cohesion Policy is to reduce differences between regions and to ensure growth across Europe (European Commission, 2018).

Whether regional specialisation is indeed more beneficial for regional economic growth, innovation and stability compared to regional diversification, is the central question in the well-known *Marshall vs. Jacobs* debate. In so far as innovation is a process of recombining existing knowledge into new ideas, knowledge spillovers constitute an important input for innovative activity. The question is whether spillovers between firms in the same industry (specialisation) or between firms in different industries (diversification) are a more important driver of regional innovative performance.

Castaldi, Frenken and Los (2015) examines how the variety of economic activities in a region affects its inventive performance. The authors expect that knowledge spillovers between firms operating in closely related sectors result in a high number of innovations, most of which will be incremental or process innovations. On the other hand, radical innovations that represent important technological breakthroughs are expected to arise from the recombination of knowledge from unrelated industries. Castaldi et al. (2015) confirms the hypotheses using patent data for U.S. states in the period 1977 to 1999.

The study presented in this paper is a scientific replication of Castaldi et al. (2015, hereafter: CF&L). The hypotheses of CF&L are tested using patent data for 266 European NUTS-2 regions between 2007 and 2013. The results support both hypotheses, confirming that for European regions, related variety is most strongly associated with a high number of innovations, while unrelated variety is associated with the ability to produce radical innovations. Furthermore, this study adds to the related-unrelated variety literature by explicitly considering the difference between technological and industrial variety, which stems from choices in variable definition.

The remainder of this paper is structured as follows: section 2 briefly discusses the related literature; section 3 reviews the measurement and methodology used in this study; section 4 discusses the descriptive statistics; section 5 presents the results of the main analysis; section 6 considers the difference between technological and industrial variety and introduces two additional hypotheses; section 7 provides the robustness checks; finally, the paper is concluded in section 8 with a short discussion and some final remarks.

2. Related literature

Throughout history, economic activities have clustered geographically and by now it is well-accepted in economics that firms benefit from locating near one another, although the mechanisms through which they do so remain at least partly unknown. Agglomeration economics deals with the spatial co-location patterns of firms and industries. As summarized by Frenken, Van Oort and Verburg (2007), firms may benefit from a large shared market and availability of suppliers (*localization economies*), from locating in large and dense urban areas (*urbanization economies*) or from locating in regions with a variety of sectors (*Jacobs externalities*). Localization economies are also known as Marshall or Marshall-Arrow-Romer (MAR) externalities, after Marshall (1890).

Jacobs externalities are named after Jane Jacobs, who introduced the concept in her influential book *The Economy of Cities* (Jacobs, 1969). She argued that urban economies, which are more diverse than average, are particularly supportive of competition and knowledge spillovers between industries (Desrochers & Hospers, 2007). Jacob's ideas were picked up by Lucas (1988) and gained popularity following a seminal paper by Glaeser, Kallal, Scheinkman and Shleif (1992) in which Jacobs externalities are contrasted to Marshall-Arrow-Romer and Porter externalities. The findings suggest that local competition and urban variety drive employment growth, thus supporting the concept of Jacobs externalities (Glaeser et al., 1992).

The impact of different types of agglomeration economies has been a much-debated topic in the years following Glaeser et al. (1992), and so far, the evidence in the literature has been unable to settle the Marshall vs. Jacobs debate (Beaudry & Schiffauerova, 2009; Melo, Graham & Noland, 2009). In a recent meta-analysis, De Groot, Poot and Smit (2016) find that the evidence points most strongly towards Jacob's theory of agglomeration externalities, thus confirming the initial findings of Glaeser et al. (1992). However, in Glaeser et al. (1992) a considerable number of the studies under investigation did report either insignificant results or results in favour of Marshall externalities.

Frenken et al. (2007) proposes an extension to the concept of Jacobs externalities by introducing the related variety hypothesis. A differentiation is made between related and unrelated variety and different effects on regional economies are expected for both types. Frenken et al. (2007) argues that knowledge can only spill over from one sector to the next when there are at least some complementarities in the capabilities of those sectors. Therefore, regions home to firms that operate primarily in closely related sectors (related variety) are expected to be especially conductive of Jacobs externalities, promoting knowledge spillovers and innovation. On the other hand, a broad variety of firms operating in unrelated sectors (unrelated variety) is expected to act as a portfolio that shields regions from sector-specific unemployment shocks. Finally, regional specialisation is expected to promote incremental product and process innovations, resulting in productivity growth. Frenken et al. (2007) assumes that related variety promotes knowledge spillovers and innovation, which result in employment growth. It is hypothesized that related variety is positively related to employment growth, that unrelated variety is negatively related to unemployment growth, and that regional specialisation is positively related to productivity growth. Using a data set on Dutch NUTS-3 regions spanning the years 1998-2006, Frenken et al. (2009) finds evidence to support the hypotheses.

Since its introduction, the related variety hypothesis has been tested by different authors, most of whom focus on the effects on employment and productivity growth (Content & Frenken, 2016). Summarizing 16 papers that follow the original related variety hypothesis, Content and Frenken (2016) concludes that most empirical evidence confirms the theory that related variety is positively related to employment growth, although the effects may be sector specific. The papers considered in Content and Frenken (2016) focus on some measure of economic growth as dependent variable. A surprisingly small number of papers examines the effect of related variety on innovation, the hypothesized linking mechanism between variety and economic performance.

To the best of my knowledge, there are only three studies that directly test how variety affects the innovative performance of regions. First, Tavassoli and Carbonara (2014) uses data on 81 Swedish functional regions to test the effect of internal and external knowledge intensity and variety on regional innovation output, as measured by patent application counts. The results suggest that related variety is positively related to the innovativeness of regions, while such an effect is not found for unrelated variety. Furthermore, the innovation output of regions increases with the intensity of external knowledge flowing to the region (Tavassoli & Carbonara, 2014).

Second, CF&L expands the related variety hypothesis further by arguing that both related and unrelated variety foster innovation, though the type of innovation created is expected to differ. Knowledge from related sectors is expected to be recombined easily and it is hypothesized that related variety increases the number of innovations in a region. Furthermore, it is hypothesized that while knowledge spillovers between unrelated sectors is less likely, successful recombination of unrelated knowledge results in more radical, breakthrough innovations. The hypotheses are tested and confirmed using patent data for U.S. states between 1977 and 1999.

Third, Miguelez and Moreno (2018) follows CF&L and examines the effect of related and unrelated variety on the number of patents and the quality of patents, respectively. As an extension, the paper examines how the inflow of related knowledge into the region affects innovative output. Different to CF&L, Miguelez and Moreno (2018) focusses on patent-intensive sectors only. Using data on 255 European NUTS-2 regions, the results confirm the hypotheses concerning related and unrelated variety, also in accordance with both Tavasolli and Carbonara (2013) and CF&L. Moreover, Miguelez and Moreno (2018) finds that the inflow of similar external knowledge and not the inflow of related external knowledge in to the region affects regional innovativeness.

Scientific replications of non-experimental studies are useful in economics because the results found for a particular time period or region do not necessarily transfer to other time periods or regions (Hamermesh, 2007). Until now, the linking mechanism of the related variety hypothesis, the effect of variety on innovative performance which in turn affects economic growth, has only been tested for Sweden (Tavassoli & Carbonara, 2014), the United States (CF&L) and a small number of patent-intensive industries in Europe (Miguelez & Moreno, 2018). This paper presents a replication of CF&L, testing the same hypotheses using a similar research design, but for a different time period and region. CF&L test their hypotheses for 51 U.S. states for the years 1977 to 1999, while this study uses data on 266 European NUTS-2 regions between 2007 and 2013. The patent data used in this study includes all sectors that produced at least one patent in the time period, and represents a much broader variety of industries compared to the five patent-intensive industries in Miguelez and Moreno (2018).

The hypotheses tested are identical to that of CF&L (p. 770):

- Hypothesis 1: Regional related variety is positively associated with regional inventive performance.
- Hypothesis 2: Regional unrelated variety is positively associated with the regional ability to produce breakthrough innovations.

3. Measurement and methodology

Patent data is used to capture both innovative output and variety in regions. Patents are a useful proxy for innovative output because inventions have to meet certain standards in terms of novelty, originality and usefulness before a patent is granted; an important drawback from using patent data is that not all inventions are patented (Bottazzi & Peri, 2003).

The original patent data used in this study was kindly provided by dr. N. Cortinovis from Erasmus University Rotterdam and is appended with data on regional R&D expenses and employment of researchers, taken from the Eurostat regional database. The patent data is collected from the OECD REGPAT database and spans the years 2007 to 2013. This dataset provides information on individual patents, including the regions and inventors to which they are assigned, the year in which they were applied and the technological sections, classes and subclasses to which they are assigned. Patents are classified according to the International Patent Classification of the WIPO and assigned to one or more of 8 broad sections, 122 classes and 629 subclasses. Patents are not uniquely assigned to regions, inventors and classifications and are therefore weighted across regions and technological classifications when calculating the dependent and independent variables. The study is carried out for 266 European NUTS-2 regions, a list of which is included in Appendix I.

The first dependent variable is a measure of the inventive performance of regions, approximated by the number of innovations. Following CF&L, this is measured by a count of patents assigned to the region. Patents are weighted across regions according to the number of researchers to which a patent is assigned in each region. The resulting continuous variable is rounded to integer values to arrive at a count variable that is similar to that of CF&L.

The second dependent variable measures the capability of regions to produce breakthrough innovations. Breakthrough innovations result in high-value patents which are sometimes referred to as *superstar patents*. The value of patents can be approximated by measures based on forward citations. The most-cited patents within a cohort are usually considered to represent breakthrough innovations, with a cut-off point defined at 95% or 99% (see in example Ahuja & Lampert, 2001 and Zheng & Yang, 2014). CF&L uses a more refined methodology and endogenously derives the share of superstar patents by exploiting the statistical properties of the frequency distribution of forward citation numbers (CF&L, p. 770). Taking in to account that superstar patents continue to receive significant amounts of citations many years after filing while regular patents do not, the authors are able to identify superstar patents and predict which recent patents will likely be superstar patents. Due to the short time period of the present dataset, this measure is not feasible here and a simpler method is used. Comparing patents applied in the same year and assigned to the same technological class, the top 5% most-cited patents are identified as superstar patents. As before, patents are weighted according to the number of regions and inventors to which they are assigned before calculating the share of superstar patents for each region.

Next, the main independent variables are measures of regional variety. Information on the classification of patents is exploited to calculate the related and unrelated variety of regions using entropy statistics. Over the past decades, entropy measures have been applied in the context of regional variety (see in example Frenken et al., 2007; Miguelez & Moreno, 2018). It is a measure of uncertainty and when applied to a regional context, it provides information on the diversification or specialisation of regions, for example in terms of employment patterns across sectors or innovative activity across technological classes. An important advantage of the entropy measure is that it can be decomposed at different levels of aggregation, so that variety at different levels can be included in a single model without necessarily causing collinearity (Frenken et al., 2007). The following paragraphs provide a brief summary of the entropy measure, more detailed accounts are found in Frenken et al. (2007) and CF&L. Since entropy statistics require units to be uniquely assigned to classifications at different levels of hierarchy (Frenken et al., 2007), patents are weighted according to the number of regions, sections and subclasses before aggregating to regional level and calculating the measures of variety.

The weighted patents are uniquely assigned to four-digit subclasses, which fall exclusively within two-digit classes, which in turn are assigned to one of eight one-digit sectors. As patents in a region are more equally distributed among broadly defined one-digit sections, the stock of patents in this region is more diversified across a highly diverse portfolio of industries. This is unrelated variety (UV), which is measured by entropy at the one-digit level. Formally,

$$UV_{it} = \sum_{k=1}^{8} s_{k,it} \ln \left(\frac{1}{s_{k,it}}\right) \tag{1}$$

where $s_{k,it}$ is the probability that a weighted patent, assigned to region i and applied in year t, falls in one-digit section k. As in CF&L, an intermediate level of variety is introduced here. This intermediate level of variety is called *semi-related variety* and is the weighted average of variety between two-digit sectors within each broadly defined one-digit section. Formally, semi-related variety is defined as entropy at the two-digit class level minus entropy at one-digit section level:

$$SRV_{it} = \sum_{l=1}^{122} s_{l,it} \ln\left(\frac{1}{s_{l,it}}\right) - \sum_{k=1}^{8} s_{k,it} \ln\left(\frac{1}{s_{k,it}}\right) \tag{2}$$

where $s_{l,it}$ is the probability of a weighted patent being assigned to one of 122 classes. Related variety is constructed in a similar manner, where $s_{m,it}$ is the probability that a weighted patent is assigned to one of 629 subclasses. Formally,

$$UV_{it} = \sum_{m=1}^{629} s_{m,it} \ln\left(\frac{1}{s_{m,it}}\right) - \sum_{l=1}^{122} s_{l,it} \ln\left(\frac{1}{s_{l,it}}\right) \tag{3}$$

Finally, as in CF&L, one control variable is included in the base model. R&D expenditures are a proxy of investment in innovation. R&D expenditures are measured in million purchasing power standards (PPS) at 2005 prices, to facilitate comparison between European regions. Missing values are filled in using linear interpolation to limit the loss of observations.

Following CF&L, the regression equations are estimated using *generalized linear models* (GLM). Observations are pooled across states and years; all variables except the number of patents (a count) are standardized; one-year lags of independent variables are included because innovative output is related primarily to efforts made in the past; and a year variable is included to account for time trends. The hypotheses are tested using the following equations, which mirror equations (8) and (9) in CF&L:

$$\begin{aligned} NUMPATENTS_{it} = \; & \alpha^N + \beta_1^N UV_{i,t-1} + \beta_2^N SRV_{i,t-1} + \beta_3^N RV_{i,t-1} + \gamma^N RD_{i,t-1} \\ & + \delta^N \boldsymbol{d} + v_{it} \end{aligned} \tag{4}$$

$$\begin{aligned} SHARESUPER_{it} = \; & \alpha^S + \beta_1^S UV_{i,t-1} + \beta_2^S SRV_{i,t-1} + \beta_3^S RV_{i,t-1} + \gamma^S RD_{i,t-1} \\ & + \delta^S \boldsymbol{d} + v_{it} \end{aligned} \tag{5}$$

in which **d** is a vector containing country dummies and a time variable. Instead of country dummies, CF&L include individual dummy variables for each region to account for region-specific time-invariant effects. Therefore 51 additional parameters, one for each state, have to be estimated. To apply the same method in this study would increase the number of dummy variables to 266, one for each NUTS-2 region. This greatly reduces the efficiency of the model, causing what is known as the *incidental parameters problem* (Lancaster, 2000; Greene, 2002). To mitigate this problem, country dummies are included in place of individual region dummies. Regions within a single country are part of the same institutional context and laws and will generally be characterised by a similar culture and attitude towards innovation.

A negative binomial model is estimated for (4) since the number of patents is a count variable, and a linear model is estimated for equation (5). Finally, the models are estimated using robust standard errors.

4. Descriptive statistics

The following paragraphs discuss the descriptive statistics and how they compare to CF&L. Descriptions of the variables and the main statistics are presented in Table 1 on the next page, which compares to Table 1 in CF&L (p. 773). The analysis is carried out for 266 European NUTS-2 regions for the period of 2007 to 2013. Isolated overseas regions of Spain, France and Portugal are treated as outliers and are consequently dropped. In some cases, shares of superstar patents of 50% or higher are observed, resulting from regions producing only one or two patents in a given year, one of which is highly successful. These cases are Åland (Finland) in 2013, Podlaskie (Poland) in 2009 and Sud-Muntenia (Romania) in 2007, and these observations are dropped. Finally, all observations with missing values for the number of patents were also dropped.

The first dependent variable is the number of patents, which differs considerably between regions. There are regions which produce very few patents, such as Molise (Italy) and Yugoiztochen (Bulgaria). In other regions, many patents are applied each year. Top-patenting regions include the Parisian Region (France), Upper Bavaria (Germany) and Stuttgart (Germany). The number of patent applications varies from 0 in several years and regions to 3278 in the Parisian Region in 2008. In CF&L, the difference between regions is larger, ranging from 12 to 15 404 patent applications. The top five patenting states in CF&L produce around 45% of all patents in both 1977 and 1999. The top five patenting NUTS-2 regions in this study account for around 19% of all patents applied each year. Patent applications are more evenly spread across the 266 NUTS-2 regions compared to the 51 U.S. states in CF&L. Figure 1 on the next page maps the distribution of patent applications across the NUTS-2 regions in 2007.

Variable		Min	Max	Mean	SD
Number of patents	Total number of patents applied in year t assigned to inventors located in the region	0	3278	229.60	390.31
Share of superstar patents	Share (%) of superstar patents in total patents in the region	0	33.33	3.31	3.51
UV	Entropy at the section (1-digit) level of the International Patent Classification (IPC)	0	2.06	1.68	0.33
SRV	Entropy at the class (3-digit) level of the IPC minus entropy at the section (1-digit) level	0	2.09	1.28	0.50
RV	Entropy at the subclass (4-digit) level of the IPC minus entropy at the class (3-digit) level	0	1.50	0.76	0.38
RD	Total R&D expenditures (in million PPS at 2005 prices)	1444	15 568	883	1435

Table 1: Variables (NUTS-2)

A clear pattern emerges, with the highest number of patents concentrating in the Netherlands, Germany, Austria, Switzerland and Northern Italy.

The share of superstar patents in 2007 is mapped in Figure 2 below. Comparing Figure 1 and 2, regions with a high share of superstar patents do not correspond one-to-one with those with a high number of patents. Regions in Eastern Europe and Southern Italy that do not produce many patents, are among the regions that produce relatively most superstar patents. There are also regions, in example in the south east of Great Britain and in Belgium, that produce both a high number of patents and relatively many superstar patens. With the share of superstar patents ranging between 0% and 33.33%, differences between regions are larger compared to CF&L, where the share of superstar patents ranges from 2% to 7%.

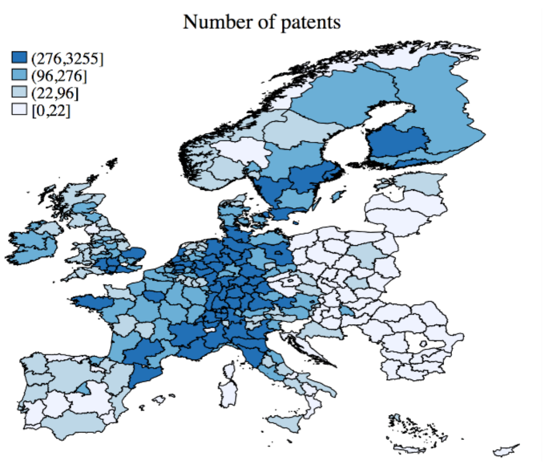

Figure 1: Number of patents (NUTS-2, 2007)

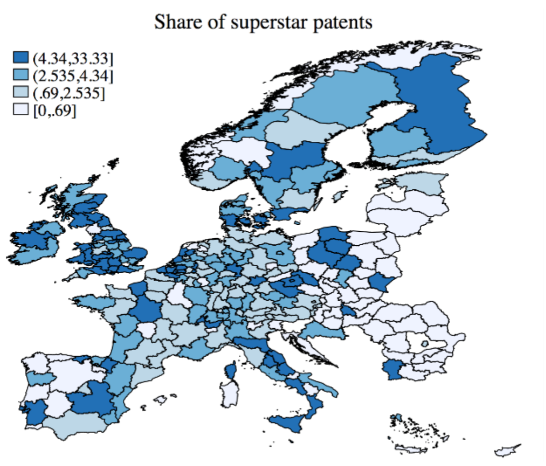

Figure 2: Share of superstar patens (NUTS-2, 2007)

For all three measures of variety, the minimum value of 0 is found for regions where few patents are applied, in which case all patents belong to the same section, class or subclass. Examples are Sud-Est (Romania), Extremadura (Spain) and Algarve (Portugal). The upper panel of Figure 3 on the next page maps unrelated variety in 2007. The maximum attainable value of UV is *ln(8)* = 2.08, so with a mean of 1.68, regions' portfolio of patents were quite diversified across the broadly defined sections. The average unrelated variety shows a small increase over time, from 1.64 in 2007 to 1.68 in 2007. With respect to unrelated variety, the patterns identified here are similar to that of CF&L, with most regions being quite diversified across broadly defined categories.

A map of semi-related variety is given in the middle panel of Figure 3. The pattern looks similar to that of unrelated variety, with the highest values in North West and Central Europe, intermediate values in North and South Europe and the lowest values in Eastern Europe. Between 2007 and 2013, the average semi-related variety increased slightly from 1.24 to 1.26. The maximum attainable value for semi-related variety is *ln(122)-ln(8)* = 2.72. With an average value of 1.28, variety at the class level is well below the maximum value. Patents seem to be concentrated in a number of technological classes within each section.

Related variety decreased slightly from 0.77 in 2007 to 0.73 in 2013. The maximum attainable value for related variety is *ln(629)-ln(122)* = 1.64, and with an average of 0.76, patens are fairly unevenly spread across the 629 subclasses. A map is given in the lower panel Figure 3, and the pattern is similar to that of both unrelated and semi-related variety.

As in CF&L, the coefficient of variation (the ratio of the standard deviation to the mean) increases with the level of detail at which variety is measured, indicating that there is more variation in the level of related variety between regions, than in the level of unrelated variety.

The average R&D expenditures rose from 818 million PPS in 2007 to 941 million PPS in 2013. Regions that spent most on R&D are located in North West and Central Europe, and along the Mediterranean coastlines of Italy, France and Spain. Regions in Eastern Europe have the lowest R&D expenditures. The Parisian Region in France invested most in R&D, with an average of 14 891 million PPS per year, more than 6000 million PPS more than the runner-up, the German region Stuttgart (on average 8704 million PPS per year). Regions with lowest R&D expenditures include Åland (Finland, 2.48 million PPS per year) and Severen Tsentralen (Bulgaria, 7.50 million PPS per year). As in CF&L, large differences exist between the regions and average R&D spending grew over time.

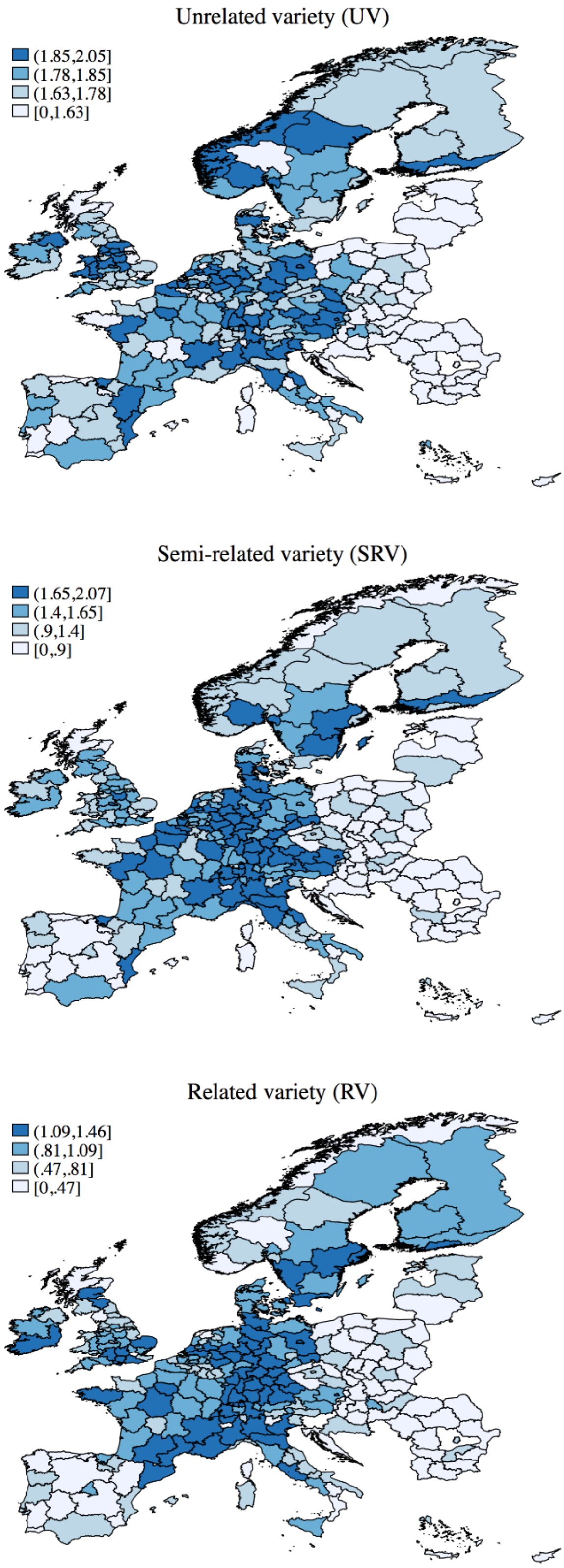

Figure 3: Unrelated, semi-related and related variety (NUTS-2, 2007)

	Number of patents	Share of superstar patents	RD_{t-1}	UV_{t-1}	SRV_{t-1}
Share of superstar patents	-0.001				
RD_{t-1}	0.877**	0.038			
UV_{t-1}	0.262**	0.130**	0.254**		
SRV_{t-1}	0.439**	0.070**	0.371**	0.716**	
RV_{t-1}	0.606**	0.087**	0.0567**	0.558**	0.764**

Table 2: Correlation matrix (NUTS-2)
*** significant at 5%*

Table 2 above presents the correlation matrix (compare to Table 2 in CF&L, p. 775). In Figure 1 and 2 the observed patterns for the number of patents and the measures of variety were similar. This is confirmed in the correlation matrix: the number of patents is positively and significantly related to R&D expenditures and all measures of variety. The number of patents and the share of superstar patents are not correlated, confirming the observation from Figures 1 and 2 that they do not follow a similar pattern. The share of superstar patents in positively and significantly related to all three measures of variety, but not to R&D expenditures. The different measures of variety are positively and significantly correlated, in line with the observed patterns in Figure 3.

Finally, the scatterplot in Figure 4 on the right shows how region averages of related and unrelated variety are positively related. This is line with CF&L, who find that an increase of 0.1 in unrelated variety results in an average increase of 0.22 in related variety (see Figure 1 in CF&L, p. 774). The relationship does not appear to be linear here. High related variety is only observed in conjunction with high unrelated variety. Apparently, there are no regions which are specialized in one or few broadly defined sections and highly diversified within those sections. On the other hand, low related variety is found for both high and low values of unrelated variety, indicating that specialisation at subclass level is common for both regions highly diversified at the level of broadly defined sections, and regions which are specialized in one or few sections.

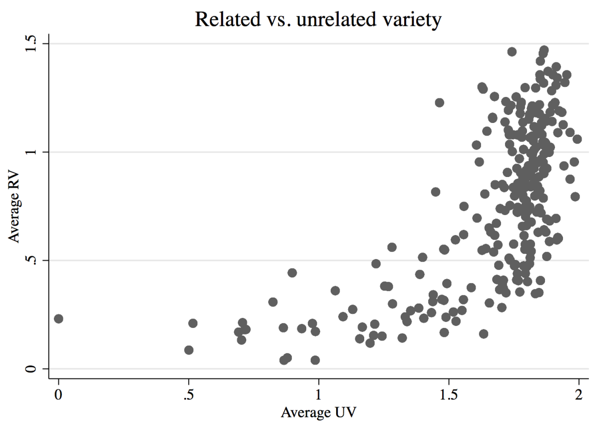

Figure 4: Related vs. unrelated variety
Dots represent year averages for each NUTS-2 region between 2007 and 2013.

5. Results

The regression results for (4) and (5) are given in Table 3 below. Following CF&L, four nested models are estimated: the first includes only R&D expenditures; the second also incorporates the cluster dummies and time trend variable; the third is the full model and the fourth excludes semi-related variety. For ease of comparison, the results of CF&L are repeated in *italics*. Wald tests indicate that including the measures of variety in the third model significantly improves the fit of the model for both dependent variables.

5.1 Number of patents

The results for equation (4) are presented in the upper panel of Table 3. R&D expenditures have a positive and significant effect on the number of patents. Different to CF&L, the effect of R&D spending is robust to adding more explanatory variables to the model, although its magnitude decreases. At least between 2007 and 2013, higher spending on R&D increased the number of patents in the NUTS-2 regions. The coefficient for the time trend is negative and significant, indicating that in general, the number of patents decreased over time.

The full model is represented by Model 3. The effects of the different types of variety are

	Model 1		Model 2		Model 3		Model 4	
	CF&L	**NUTS-2**	*CF&L*	**NUTS-2**	*CF&L*	**NUTS-2**	*CF&L*	**NUTS-2**
Dependent variable: number of patents								
RD_{t-1}	*0.910****	**1.334*****	*0.068*	**0.976*****	*0.087*	**0.496*****	*0.093*	**0.478*****
Country dummies			*Yes*	**Yes**	*Yes*	**Yes**	*Yes*	**Yes**
Time trend			*0.301****	**-0.028*****	*0.298****	**-0.130*****	*0.303****	**-0.117*****
UV_{t-1}					*-0.084*	**0.173*****	*-0.086*	**0.291*****
SRV_{t-1}					*-0.046*	**0.338*****		
RV_{t-1}					*0.325***	**0.480*****	*0.322***	**0.624*****
N	*877*	**1734**	*877*	**1734**	*877*	**1716**	*877*	**1716**
Dependent variable: share of superstar patents								
RD_{t-1}	*0.216****	**0.024****	*0.167****	**0.051*****	*0.197****	**0.018***	*0.210****	**0.019***
Country dummies			*Yes*	**Yes**	*Yes*	**Yes**	*Yes*	**Yes**
Time trend			*0.378****	**0.035*****	*0.334****	**0.018****	*0.345****	**0.018****
UV_{t-1}					*0.118****	**0.090*****	*0.117****	**0.088*****
SRV_{t-1}					*-0.103****	**-0.006**		
RV_{t-1}					*0.085*	**0.060***	*0.078*	**0.057****
N	*877*	**1723**	*877*	**1723**	*877*	**1707**	*877*	**1707**

Table 3: Regression results (NUTS-2)
** significant at 10%, ** significant at 5%, *** significant at 1%*

all positive and significant at 1%. The hypothesis that related variety is positively associated with innovative output is therefore confirmed. Additionally, semi-related and unrelated variety are also positively associated with the number of patents. Since all variables are standardized, the magnitude of the effects can be compared easily. The effect of related variety is strongest, and the effect of semi-related variety is almost equally strong. The effect of unrelated variety is smaller: a coefficient of 0.173 compared to 0.480 for related variety. The results indicate that variety at all levels is related to a higher number of patents but that the effect is strongest for related and semi-related variety. If low variety at all levels is equated to strong specialisation (Aarstad, Kvitastein, & Jakobsen, 2016), then specialisation is negatively associated with innovative output. These results differ from CF&L, which only finds a positive and significant result for related variety. CF&L estimates a fourth model excluding semi-related variety to check the robustness of the results for related and unrelated variety. The same is done here and the results are presented in the last column of Table 3. Excluding semi-related variety does not affect the results for the other variables.

5.2 Share of superstar patents

The lower panel of Table 3 presents the results for equation (5). The effect of R&D expenditures on the share of superstar patents is positive and significant in all specifications, suggesting that R&D expenditures have a positive effect on the ability to produce radical innovations. The estimated effect of the time trend is also positive and significant in all specifications: the average share of superstar patents increased between 2007 and 2013. The coefficient for unrelated variety is positively related to the share of superstar patents, and highly significant at 1%, thereby confirming hypothesis 2. Additionally, an increase in related variety is associated with an increase in the share of superstar patents, while no significant effect is found for semi-related variety. The fourth model shows that when semi-related variety is left out of the equation, the results for the other variables are not affected. The results for the share of superstar patents are different from those of CF&L, which finds a negative and significant effect for semi-related variety, but none for related variety. The results suggest that in Europe, both related and unrelated variety contribute to the creation of breakthrough innovations, though the effect seems to be stronger for unrelated variety.

6. Industrial versus technological variety

So far, the measures of variety have been calculated using patent data. However, patent data only provides information on patentable, economically valuable inventions. Any variety measure derived from patent data actually measures the variety of technologies. An alternative way to measure variety is to calculate it using the industrial classification of employment data. If variety based on patent data represents *technological variety*, then variety based on employment is the *industrial variety.* Intuitively, variety based on employment provides a more complete representation of the economic variety within regions, as it incorporates all economic activities present in the region and not just those industries that produce patentable technologies. Industrial variety measures not only the diversity of technologies, but also the variety of applied knowledge and skills in the region.

The distinction between technological and industrial variety is not made explicitly in the related and unrelated variety literature. Most studies that examine the effects of variety on regional economies use industrial variety as explanatory variable, including Frenken et al. (2007), Bishop and Gripaios (2010) and Hartog, Boschma and Sotarauta (2012). In contrast, Tavassoli and Carbonara (2014), CF&L and Miguelez and Moreno (2018) use technological variety to study the effect of variety on regional innovative performance.

To consider only technological variety as determinant of the innovative performance of regions underestimates the importance of applied skills and knowledge. By focussing solely on technological variety, implicitly it is assumed that only knowledge spillovers between patenting industries are relevant for innovation. The definition of Jacobs externalities, however, does not limit knowledge spillovers to high-patenting sectors. As defined in Frenken et al. (2007, p. 687), Jacobs externalities are external economies which are available to all firms in the region and arise from a diversity of sectors within the region. Indeed, as quoted in CF&L, Jacobs (1969, p. 59) notes that:

> "the greater the sheer numbers and varieties of divisions of labor already achieved in an economy, the greater the economy's inherent capacity for adding still more kinds of goods and services. Also the possibilities increase for combining the existing divisions of labor in new ways."

Jacobs does not make a distinction between high-patenting, technology-intense sectors and other sectors. Instead, she views the variety of employment as an important driver of innovation. As innovation is conceptualized here as the recombination of existing technologies, the availability of technologies from related and unrelated sectors can be considered a prerequisite for innovation. This availability is captured by technological variety. However, to test whether Jacobs externalities are indeed positively associated with the number of patents and the ability to produce superstar patents, it is more appropriate to use variety measures based on employment. In the remainder of this section, two additional hypotheses are tested:

- Hypothesis 3: Regional related industrial variety is positively associated with regional inventive performance.
- Hypothesis 4: Regional unrelated industrial variety is positively associated with the regional ability to produce breakthrough innovations.

The measures of industrial variety used to test hypotheses 3 and 4 are provided by dr. N. Cortinovis. The measures are calculated using entropy statistics and the data is retrieved from the ORBIS database (for details, see Cortinovis & Van Oort, 2015). Unrelated industrial variety is defined as the variety between 21 broadly defined sections of the NACE industrial classification, while related industrial variety is defined as variety within those sections. This provides a strict definition of unrelated variety, as only sectors belonging to different broadly defined sections are considered unrelated. The definition of industrial related variety makes no distinction between semi-related and related variety, as was the case with technological variety. Both types of related variety are captured by the measure of industrial related variety. Correlation analysis reveals that the measures of technological and industrial variety are not strongly correlated, which confirms that they measure different underlying regional structures. The highest correlation (0.061) is found for technological and industrial unrelated variety.

That technological and industrial variety are indeed different phenomena, can also be observed from the scatterplots presented in Figure 5 and 6 below. In both figures, the average technological variety is plotted on the y-axis and the industrial variety on the x-axis. In Figure 5, high technological UV is matched by high industrial UV in a majority of the observations.

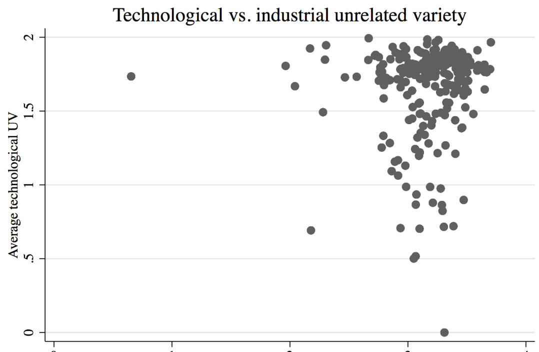

Figure 5: Technological vs. industrial UV (NUTS-2)

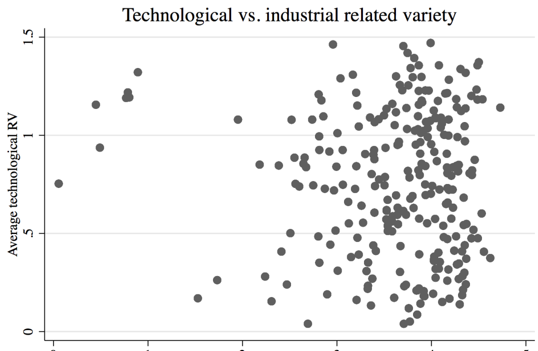

Figure 6: Technological vs. industrial RV (NUTS-2)

However, there are also observations with low technological UV and high industrial UV, or vice versa. The correlation between the two is 0.061 and significant at 5%. Figure 6 shows no clear pattern in the relationship between technological and industrial related variety, for which the correlation is -0.056 and significant at 5%.

The results for hypothesis 3 and 4 are presented in Model 5 in Table 4 below. For the number of patents, the estimated effects for R&D expenditures, the time trend and technological UV, SRV and RV do not differ considerably from those reported in Table 3. Additionally, the estimated coefficient for industrial related variety is positive and significant at 1%, thus confirming hypothesis 3. The magnitude of the effect is comparable to that of technological variety, indicating that both measures of variety are equally relevant for the number of patents. The results for hypothesis 4 are reported in the right column of Model 5. Again, the estimated coefficients for the time trend and technological UV, SRV and RV do not change when industrial unrelated variety is included in the model. However, the effect of R&D expenditures on the ability to produce breakthrough innovations is no longer significant when industrial unrelated variety is included in the model. The estimated effect of industrial unrelated variety is positive and significant at 5%, thus providing evidence to support hypothesis 4. The results suggest that not only a variety of related and unrelated technologies is relevant for innovation, but also a variety of applied skills and knowledge.

	Model 5	
Dependent variable:	**Number of patents**	**Share of superstar patents**
$RD_{t\text{-}1}$	0.550***	0.012
Country dummies	Yes	Yes
Time trend	-0.109***	0.016*
$UV_{t\text{-}1}$	0.170***	0.086***
$SRV_{t\text{-}1}$	0.251***	0.000
$RV_{t\text{-}1}$	0.430***	0.060*
Industrial $UV_{t\text{-}1}$		0.056**
Industrial $RV_{t\text{-}1}$	0.410***	
N	*1716*	*1707*

Table 4: Technological VS. industrial variety (NUTS-2)
** significant at 10%, ** significant at 5%, *** significant at 1%*

7. Robustness checks

6.1 Spatial effects

One concern with the analysis so far, and one that is also addressed in CF&L, is that the analysis does not account for spillover effects between regions. However, as with other economic activities, the innovation climate in one region is affected by that of its neighbours. As a robustness check, CF&L accounts for R&D expenditure spillovers. For each state, the sum the R&D efforts of neighbouring states is calculated and included as an additional explanatory variable in the model. A similar procedure is followed here. The R&D efforts of nearby regions are weighted using an inverse distance matrix, with a cut-off point at 300 km from the centre of the region. In previous research, R&D knowledge spillovers were found to occur within a 300km radius (Bottazzi & Peri, 2003). Therefore, only regions within a 300km radius should be considered as neighbours (Cortinovis & Van Oort, 2015). The inverse of the distance between regions is taken to allow the strength of spillovers to decrease with distance. Finally, the matrix is row-standardized. The use of an inverse distance matrix is preferred over the contingency-based spatial weighting method used in CF&L, because European regions can differ substantially in terms of geographic size. At some places, knowledge may spill over several regional borders, while at others, it will not even be able to cross one region border. Furthermore, the presence of islands within the dataset limits the feasibility of contingency-based spatial weighting.

Model 6 in Table 5 on the next page presents the results for this first robustness check. The estimated coefficients for the number of patents do not differ strongly from those estimated in Models 3 and 5. R&D efforts in nearby regions is positively related to the number of patents, and the effect is highly significant at 1%. In terms of magnitude, the effect of own-region R&D expenditure is stronger compared to that of R&D expenditures of neighbouring regions: a coefficient of 0.551 compared to 0.107 for RD(neighbours). No significant effect was found for the R&D expenditures of neighbouring regions on the share of superstar patents, while the effects for the other variables remain comparable to those of Models 3 and 5. Apparently, investments in innovation made in neighbouring regions do not affect the own region's ability to produce radical innovations.

	Model 6	Model 7	Model 8
Dependent variable: number of patents			
RD_{t-1}	0.551***	0.536***	
$RD(neighbours)_{t-1}$	**0.107*****	**0.114*****	
$Researchers_{t-1}$			**0.579*****
$Res(neighbours)_{t-1}$			**0.109*****
$Pop.\ Density_{t-1}$		**0.083*****	**0.058*****
Country dummies	Yes	Yes	Yes
Time trend	-0.116***	-0.114***	-0.110***
UV_{t-1}	0.177***	0.175***	0.124***
SRV_{t-1}	0.249***	0.246***	0.249***
RV_{t-1}	0.425***	0.421***	0.408***
Industrial RV_{t-1}	0.395***	0.406***	0.484***
N	*1716*	*1710*	*1686*
Dependent variable: share of superstar patents			
RD_{t-1}	0.012	0.011	
$RD(neighbours)_{t-1}$	**0.005**	**0.005**	
$Researchers_{t-1}$			0.027**
$Res(neighbours)_{t-1}$			0.014
$Pop.\ Density_{t-1}$		**0.009**	0.003
Country dummies	Yes	Yes	Yes
Time trend	0.016*	0.017*	0.016*
UV_{t-1}	0.086***	0.086***	0.085***
SRV_{t-1}	0.000	-0.001	0.008
RV_{t-1}	0.059*	0.059**	0.049
Industrial UV_{t-1}	0.057**	0.057*	0.056*
N	*1707*	*1701*	*1677*

Table 5: Robustness checks (NUTS-2): spatial effects, urbanization economies, inputs for innovation.
** significant at 10%, ** significant at 5%, *** significant at 1%*

6.2 Urbanization economies

A second potential issue concerns the effect of urbanization economies, which has thus far been omitted from the analysis. Densely populated urban areas provide additional benefits to firms because cities are home to universities, research laboratories, trade associations and other knowledge generating institutions, which in turn give rise to a social, political and cultural environment that facilitates knowledge spillovers (Frenken et al., 2007). Moreover, the cultural and creative sectors which are present primarily in urban areas, might accelerate the generation of breakthrough innovations, as the creative atmosphere facilitates the recombination of previously unrelated technologies. Therefore, urbanization economies likely affect the innovation output of regions, in terms of the number of patents but especially concerning the share of superstar patents. As a second robustness check, a measure of urbanization economies is included in the model. Urbanization economies are approximated by population density, measured as inhabitants per square kilometre. The data is collected from the Eurostat regional database.

The results are presented in Model 7 of Table 5. Population density is positively and significantly associated with the number of patents but not to the share of superstar patents. The results for the other variables do not change in sign, significance or relative magnitude. The conclusions with respect to hypotheses 1-4 drawn earlier are robust to including urbanization effects and additionally it is found that urbanization economies are conductive of incremental innovation.

6.3 Inputs for innovation

A third way to test the robustness of the results is to define the main input for innovation differently. In the models estimated above, this input is defined as R&D expenditures in million PPS at 2005 prices. An alternative is to define it in terms of human capital. The employment of researchers in the region, measured in full-time equivalents (FTEs) is used as alternative measure in Model 8 in Table 5. The data is also from the Eurostat regional database. The employment of researchers in the own region has a positive and significant effect on both the number of patents and the share of superstar patents, though the effect is stronger for the first. The employment of researchers in neighbouring regions also positively affects the number of patents in the own region, though own-region employment is more important. In comparison to earlier models, the effect of related variety on the share of superstar patents turns insignificant when replacing R&D expenditures by the employment of researchers. This somewhat weakens the evidence that related variety is also relevant for breakthrough innovations, which was concluded earlier. The main conclusions regarding the hypotheses do remain valid.

6.4 NUTS-1 regions

As a fourth and final robustness check, the analysis is repeated almost completely at a larger geographical unit of analysis. So far, the analysis in this paper was carried out at NUTS-2 level. The NUTS classification is a hierarchical system that divides the economic territory of the

European Union in comparable regions, with the purpose of facilitating socio-economic analysis. At NUTS-1 level, the population size of regions is between 3 and 7 million and at NUTS-2 between 800 000 and 3 million. For the analysis of regions, NUTS-1 correspond to major socio-economic regions, while NUTS-2 regions are the basic regions for the application of regional policies (Eurostat, sd). This provides the rationale for using NUTS-2 regions for the main analysis in this paper. However, which level is most relevant for regional innovation strategies differs per country. In any case, even if regional policies are applied at NUTS-2 level, harmonising strategies at over a larger area is likely to improve their effectiveness. The analysis is repeated at NUTS-1 level to check whether the conclusions also hold at a larger geographical scale.

The results for the full model at NUTS-1 are presented in Model 9 in Table 6 on the next page, while Models 10-12 present several robustness checks. Due to data limitations, industrial variety is not included at NUTS-1 level. Descriptive statistics are presented in Appendix II. The results of the analysis at NUTS-1 level do not change the conclusions regarding hypotheses 1 and 2, while hypotheses 3 and 4 are not tested here. For larger geographical regions, related variety is also positively and significantly associated with a higher number of patents and unrelated variety to the share of superstar patents. As before, R&D expenditures in the own region is positively related to the number of patents, but not to breakthrough innovations. In contrast to NUTS-2 regions, the R&D expenditures in neighbouring NUTS-1 regions do not affect the number of patents in the own region. Perhaps this can be explained by the geographical size of NUTS-1 regions, which may generally be too large to facilitate meaningful knowledge spillovers across borders. Urbanization economies, as approximated by population density, have a significant positive effect on the share of superstar patents, but not on the number of patents. When the main input for innovation is defined in terms of the employment of researchers in Model 12, the estimated effect for population density on the number of patents turns significant, therefore no definitive conclusions can be drawn with respect to population density and the number of patents.

	Model 9	Model 10	Model 11	Model 12
Dependent variable: number of patents				
$RD_{t\text{-}1}$	0.589***	0.592***	0.591***	
RD(neighbours)$_{t\text{-}1}$		-0.015	-0.016	
Researchers$_{t\text{-}1}$				0.657***
Res(neighbours)$_{t\text{-}1}$				-0.020
Pop. Density$_{t\text{-}1}$			-0.016	-0.052***
Country dummies	Yes	Yes	Yes	Yes
Time trend	-0.132***	-0.131***	-0.131***	-0.128***
$UV_{t\text{-}1}$	0.118**	0.118**	0.119**	0.090***
$SRV_{t\text{-}1}$	0.344***	0.345***	0.338***	0.339***
$RV_{t\text{-}1}$	0.566***	0.566***	0.571***	0.513***
N	*660*	*660*	*660*	*660*
Dependent variable: share of superstar patents				
$RD_{t\text{-}1}$	-0.009	-0.013	-0.012	
RD(neighbours)$_{t\text{-}1}$		0.017	0.019	
Researchers$_{t\text{-}1}$				0.017
Res(neighbours)$_{t\text{-}1}$				0.009
Pop. Density$_{t\text{-}1}$			0.050***	0.049***
Country dummies	Yes	Yes	Yes	Yes
Time trend	0.037*	0.036*	0.035*	0.036*
$UV_{t\text{-}1}$	0.137***	0.138***	0.135***	0.137***
$SRV_{t\text{-}1}$	-0.199	-0.201	-0.181	-0.179
$RV_{t\text{-}1}$	0.144*	0.144**	0.130**	0.105
N	*657*	*657*	*657*	*657*

Table 6: Full model and robustness checks (NUTS-1)
** significant at 10%, ** significant at 5%, *** significant at 1%*

6. Final remarks

The analysis in this paper confirms the findings of CF&L for a different geographical area and time period. For European regions, related variety is found to be positively associated with innovative output and unrelated variety with the ability to produce breakthrough innovations. While the focus of CF&L and the first part of this paper is on technological variety, the second part of this paper highlights the difference between technological and industrial variety. Not only a variety of patentable technologies, but also variety in a broader sense – including applied skills and knowledge – is found to be relevant for the innovativeness of regions. Future research might try to disentangle the effects of technological and industrial variety and examine more closely how variety and knowledge spillovers affect innovation.

While the previous section addresses several concerns with the analysis presented in this paper, other issues are not tackled. First, even though some spatial effects were considered in section 7, the analysis does not constitute a true spatial model. Important spatial interdependencies might therefore be overlooked, while these may be highly relevant for regions in a globalised world, perhaps even more so for smaller European regions than for large U.S. states. Second, endogeneity issues likely occur because the main dependent and independent variables are calculated from the same patent data. As a related issue, a selection bias occurs since only regions with at least one patent application are included in the analysis. Consequently, the estimated coefficients are conditional upon regions producing at least one patent in a given time period. Third, the measurement of superstar patents in this paper differs from that used in CF&L. While the two methods of identifying superstar patents are not likely to result in highly divergent sets of superstar patents, this does somewhat limit the comparability of the results. Finally, reverse causality may occur because patent-intensive regions might be more attractive for locating new businesses and institutions in related fields, so that a high number of patents may increase the relayed variety in a region. The same holds for regions with a high number of superstar patents, which may increase the attractiveness of regions to such an extent that even completely unrelated industries settle in the region, thus increasing unrelated variety.

As for regional innovation strategies, the results suggest that not specialisation, but diversification is key to improving innovative performance. Under mild specialisation policies, innovative activity might be sustained in the short term as related variety stimulates incremental and process innovations. However, in the long term, breakthrough innovations are necessary to prevent regional lock-in in old industries, which tend to be replaced by new industries over time. These breakthrough innovations are strongly associated with unrelated variety. The results provide clear policy recommendations. A national strategy of diversification with regional specialisation might be feasible for smaller countries such as the Netherlands, as knowledge spills over only within a 300km radius. On the other hand, larger countries should guard a minimum level of diversification within regions to ensure sufficient levels of knowledge spillovers between related and unrelated industries.

Bibliography

Aarstad, J., Kvitastein, O. A., & Jakobsen, S.-E. (2016). Related and unrelated variety as regional drivers of enterpriseproductivity and innovation: A multilevel study. *Research Policy, 45*(4), 844–856.

Ahuja, G., & Lampert, C. M. (2001). Entrepreneurship in the Large Corporation: A Longitudinal Study of How Established FirmsCreate Breakthrough Inventions. *Strategic Management Journal, 22*(6/7), 521-543.

Beaudry, C., & Schiffauerova, A. (2009). Who's right, Marshall or Jacobs? The localization versus urbanization debate. *Research Policy, 38*(2), 318-337 .

Bishop, P., & Gripaios, P. (2010). Spatial externalities, relatedness and sector employment growth in Great Britain. *Regional Studies, 44*(4), 443–454.

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

Castaldi, C., Frenken, K., & Los, B. (2015). Related Variety, Unrelated Variety and Technological Breakthroughs: An analysis of US State-Level Patenting. *Regional Studies, 49*(5), 767-781.

Content, J., & Frenken, K. (2016). Related variety and economic development: a literature review. *European Planning Studies, 24*(12), 2097-2112.

Cortinovis, N., & Van Oort, F. (2015). Variety, economic growth and knowledge intensity of European regions: a spatial panel analysis. *The Annals of Regional Science, 55*(1), 7-32.

De Groot, H. L., Poot, J., & Smit, M. J. (2016). Which agglomeration externalities matter most and why? *Journal of Economic Surveys, 30*(4), 756–782.

Desrochers, P., & Hospers, G.-J. (2007). Cities and the economic development of nations: an essay on Jane Jacobs' contribution to economic theory. *The Canadian journal of regional science*, 119-134.

European Commission. (2018). *What is Smart Specialisation?* . Retrieved from Smart Specialisation Platform: http://s3platform.jrc.ec.europa.eu/what-is-smart-specialisation-

Eurostat. (n.d.). *Background*. Retrieved from Eurostat: https://ec.europa.eu/eurostat/web/nuts/background

Frenken, K., Van Oort, F., & Verburg, T. (2007). Related Variety, Unrelated Variety and Regional Economic Growth. *Regional Studies, 41*(5), 685-697.

Glaeser, E. L., Kallal, H. D., Scheinkman, J. A., & Shleif, A. (1992). Growth in Cities. *Journal of Political Economy, 100*(6), 1126-1152.

Hamermesh, D. S. (2007). Viewpoint: Replication in economics. *Canadian Journal of Economics, 40*(3), 715-733.

Hartog, M., Boschma, R., & Sotarauta, M. (2012). The impact of related variety on regional employment growth in Finland 1993–2006: High-tech versus medium/lowtech. *Industry and Innovation, 19*(6), 459–476.

Jacobs, J. (1969). *The Economy of Cities.* New York, NY: Vintage.

Lancaster, T. (2000). The incidental parameter problem since 1948. *Journal of Econometrics, 95*(2), 391-413.

Lucas, R. (1988). On the mechanics of economic development. *Journal of Monetary Economics, 22*(1), 3-42.

Makles, A. (2012). Stata tip 110: How to get the optimal k-means cluster solution. *The Stata Journal, 12*(2), 347–351.

Marshall, A. (1890). *Principles of Economics.* London: Macmillan.

Melo, P. C., Graham, D. J., & Noland, R. B. (2009). A meta-analysis of estimates of urban agglomeration economies. *Regional Science and Urban Economics, 39*(3), 332-342.
Miguelez, E., & Moreno, R. (2018). Relatedness, external linkages and regional innovation in Europe. *Regional Studies, 52*(5), 688-701.
Tavassoli, S., & Carbonara, N. (2014). The role of knowledge variety and intensity for regional innovation. *Small Business Economics, 43*(2), 493–509.
Zheng, Y., & Yang, H. (2015). Does Familiarity Foster Innovation? The Impact of Alliance Partner Repeatedness on Breakthrough Innovations. *Journal of Management Studies, 52*(2), 213-230.

Appendix I: List of NUST-2 regions

Regions are classified according to the 2013 NUTS classification. As described in the main body of the text, outlier observations were dropped, as well as observations with missing values for the dependent variables, and the isolated overseas regions of Portugal, Spain and France.

Country	**NUTS-2 regions**
Austria	AT11, AT12, AT13, AT21, AT22, AT31, AT32, AT33, AT34
Belgium	BE10, BE21, BE22, BE23, BE24, BE25, BE31, BE32, BE33, BE34, BE35
Bulgaria	BG31, BG32, BG33, BG34, BG41, BG42
Croatia	HR03, HR04
Cyprus	CY00, CY01
Czech Republic	CZ01, CZ02, CZ03, CZ04, CZ05, CZ06, CZ07, CZ08
Denmark	DK01, DK02, DK03, DK04, DK05
Estonia	EE00
Finland	FI19, FI1B, FI1C, FI1D, FI20
France	FR10, FR21, FR22, FR23, FR24, FR25, FR26, FR30, FR41, FR42, FR43, FR51, FR52, FR53, FR61, FR62, FR63, FR71, FR72, FR81, FR82, FR83
Germany	DE11, DE12, DE13, DE14, DE21, DE22, DE23, DE24, DE25, DE26, DE27, DE30, DE40, DE50, DE60, DE71, DE72, DE73, DE80, DE91, DE92, DE93, DE94, DEA1, DEA2, DEA3, DEA4, DEA5, DEB1, DEB2, DEB3, DEC0, DED2, DED4, DED5, DEE0, DEF0, DEG0
Greece	EL30, EL42, EL43
Hungary	HU10, HU21, HU22, HU23, HU31, HU32, HU33
Ireland	IE01, IE02
Italy	ITC1, ITC2, ITC3, ITC4, ITF1, ITF2, ITF3, ITF4, ITF5, ITF6, ITG1, ITG2, ITH1, ITH2, ITH3, ITH4, ITH5, ITI1, ITI2, ITI3, ITI4
Latvia	LV00
Lithuania	LT00
Luxembourg	LU00
Malta	MT00

Netherlands	NL11, NL12, NL13, NL21, NL22, NL23, NL31, NL32, NL33, NL34, NL41, NL42
Norway	NO01, NO02, NO03, NO04, NO05, NO06, NO07
Poland	PL11, PL12, PL21, PL22, PL31, PL32, PL33, PL34, PL41, PL42, PL43, PL51, PL52, PL61, PL62, PL63
Portugal	PT11, PT15, PT16, PT17, PT18
Romania	RO11, RO12, RO21, RO22, RO31, RO32, RO41, RO42
Slovakia	SK01, SK02, SK03, SK04
Spain	ES11, ES12, ES13, ES21, ES22, ES23, ES24, ES30, ES41, ES42, ES43, ES51, ES52, ES53, ES61, ES62
Sweden	SE11, SE12, SE21, SE22, SE23, SE31, SE32, SE33
Switzerland	CH01, CH02, CH03, CH04, CH05, H06, CH07
United Kingdom	UKC1, UKC2, UKD1, UKD2, UKD3, UKD4, UKD5, UKD6, UKD7, UKE1, UKE2, UKE3, UKE4, UKF1, UKF2, UKF3, UKG1, UKG2, UKG3, UKH1, UKH2, UKH3, UKJ1, UKJ2, UKJ3, UKJ4, UKK1, UKK2, UKK3, UKK4, UKL1, UKL2, UKM2, UKM3, UKM5, UKM6, UKN0

Appendix II: Descriptive statistics at NUTS-1

The main analysis in this paper is carried out for NUTS-2 regions. According to Eurostat, NUTS-2 regions are the basic regions for the application of regional policies, while NUTS-1 regions constitute major socio-economic regions (Eurostat). Given the wide variety of institutional structures in Europe, the relevant region size for regional policies might vary per country. Therefore, the analysis is repeated in total at NUTS-1 level. The results of are discussed in the main body of the text. Table II-1 summarize the descriptive statistics; Table II-2 presents the correlation matrix; Figures II-1 to II-5 present maps of the dependent and main independent variables; Figure II-6 presents a scatterplot of region averages of related and unrelated variety.

Variable		**Min**	**Max**	**Mean**	**SD**
Number of patents	Total number of patents applied in year t assigned to inventors located in the region	0	6548	630.25	1052.47
Share of superstar patents	Share (%) of superstar patents in total patents in the region	0	100	3.57	5.03
UV	Entropy at the section (1-digit) level of the International Patent Classification (IPC)	0	2.01	1.77	0.27
SRV	Entropy at the class (3-digit) level of the IPC minus entropy at the section (1-digit) level	0	2.10	1.44	0.46
RV	Entropy at the subclass (4-digit) level of the IPC minus entropy at the class (3-digit) level	0	1.57	0.94	0.39
RD	Total R&D expenditures (per million euros)	1	17 813	2314	2913

Table II-1: Descriptive statistics at NUTS-1 (N=669)

	Number of patents	Share of superstar patents	RD_{t-1}	UV_{t-1}	SRV_{t-1}
Share of superstar patents	-0.026				
RD_{t-1}	0.900**	0.000			
UV_{t-1}	0.237**	0.083**	0.299**		
SRV_{t-1}	0.445**	0.015	0.476**	0.748**	
RV_{t-1}	0.585**	0.092	0.656**	0.609**	0.809**

Table II-2: Correlation matrix at NUTS-1
** significant at 5%*

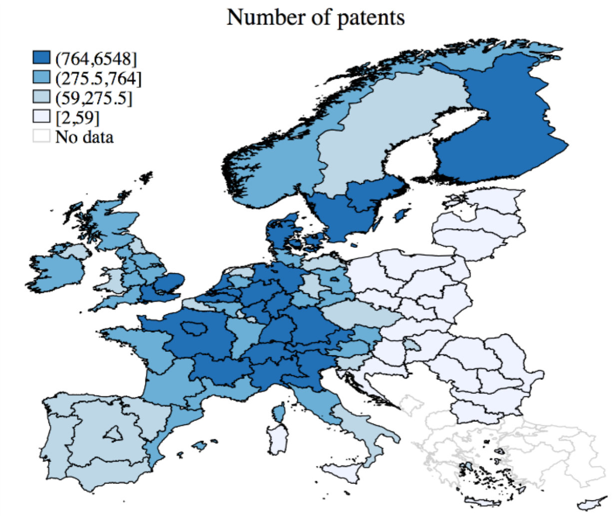

Figure II-1: Number of patents (NUTS-1, 2007)

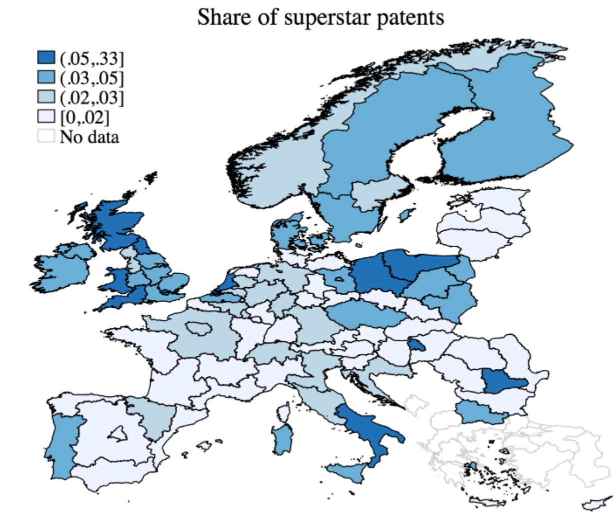

Figure II-2: Share of superstar patents (NUTS-1, 2007)

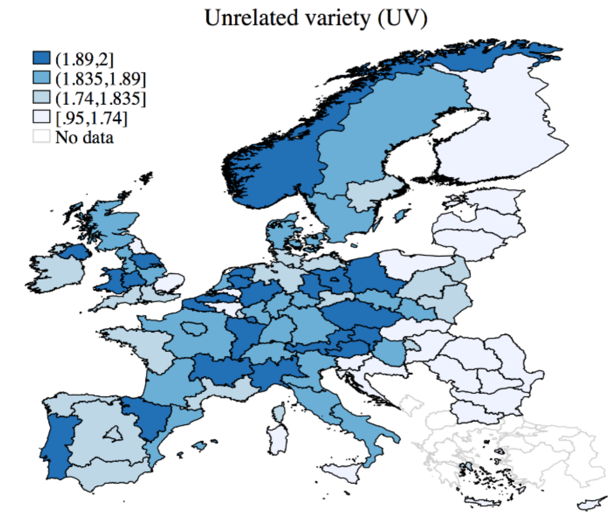

Figure II-3: Unrelated variety (NUTS-1, 2007)

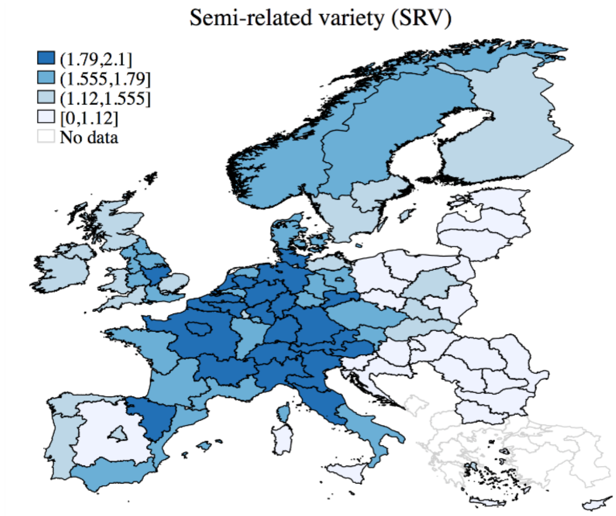

Figure II-4: Semi-related variety (NUTS-1, 2007)

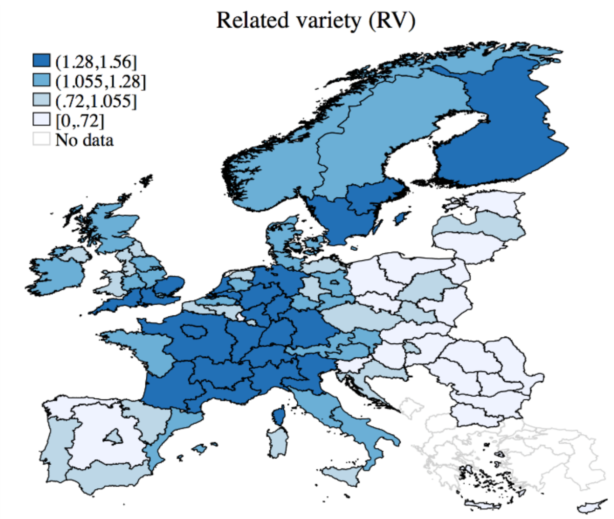

Figure II-5: Related variety (NUTS-1, 2007)

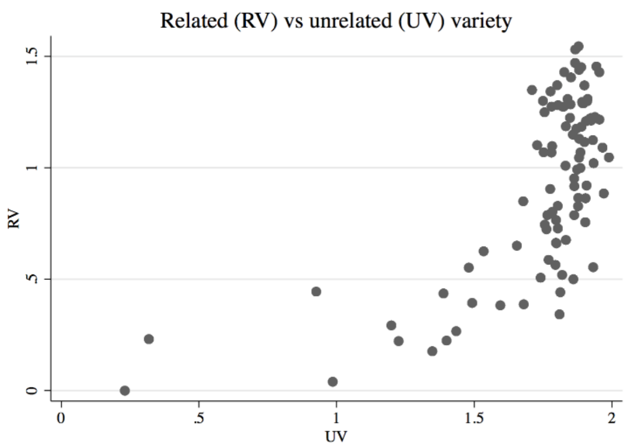

Figure II-6: Related (RV) versus unrelated (UV) variety. *Dots represent year averages for each NUTS-1 region between 2007 and 2013.*