Report on Pan-European database

Authors: Jeroen Content, Koen Frenken

Document Identifier

D3.2 Pan-European database with new indicators of related variety at national and regional (NUTS2) level; related variety indicators based on sectors, products and tasks

Version

1.0

Date Due

M12

Submission date

31st May 2016

WorkPackage

3

Lead Beneficiary

UU

Grant Agreement Number 649378

Content

1. Executive summary

This paper presents datasets on regional variables, in particular on related and unrelated variety on the one hand and opportunity-driven entrepreneurship on the other hand, for EU regions. We discuss the sources, nature and quality of the data, construction of variables as well as the descriptive statistics and spatial patterns. We end with an outlook, discussing the future research questions that can be addressed empirically using this dataset.

2. Dataset

In this document we present a dataset on variety and entrepreneurship at the regional level within EU countries. The data collected and presented here, help us understand how regional variety may enhance entrepreneurship which in turn can contribute to regional economic development. The core idea is that variety provides opportunities for recombination of ideas, skills and technologies leading to new products and services, and hereby, to new employment growth. One would expect that this creative and risky process is often, though not exclusively, carried out by entrepreneurs. The joint presence of variety and entrepreneurs, then, may lead to higher rates of employment growth. This reasoning is related to the Knowledge Spillover theory of Entrepreneurship, where entrepreneurs act as a conduit for knowledge spillovers in the regional economy (Acs et al. 2009, Fritsch and Kublina 2016), in this case, knowledge spillovers arising from recombining knowledge across industries.

In this report we present the data that has been collected to empirically investigate the interplay between variety, entrepreneurship and regional economic development.

Entrepreneurship

The key variable in our work package is the regional rate of entrepreneurship across European regions. We use the entrepreneurship variable both as a *dependent* variable when we analyse the regional determinants of entrepreneurship (including variety measures as key determinants), and as an *independent* variable when we analyse the determinants of regional economic growth as well as national diversification patterns in export portfolios.

We choose not to measure entrepreneurship simply by new firm formation, because many of the new firms created are not characterised by opportunity-driven entrepreneurship, but rather by necessity-driven entrepreneurship or for legal reasons (Shane 2008). To capture entrepreneurship that contributes to economic development, one should use a measure of opportunity-driven entrepreneurship instead. This is why we rely on the Global Entrepreneurship Monitor (GEM) data, which is survey-based data at the individual level and explicitly measures opportunity-driven entrepreneurship. Each year, the GEM conducts an

adult population survey on a representative sample of a minimum 2000 individuals per country, who are different each year. Using this data, total entrepreneurial activity is measured as the share of the working age population (from 18 until 64) that is involved in the creation of a business at the time the survey was conducted. Someone classifies as an entrepreneur when he or she engaged in any activity to start and those running a new business less than 3.5 years old. Since we break down the country numbers into regional numbers at the NUTS2 level, the annual survey waves are not representative at the regional level, as these still are based on the 2000+ individuals who are sampled at the national level. For this reason, we pool regional data over multiple waves, as to get a more reliable number for total entrepreneurial activity in the region. Of course this comes at the cost of some time variation.

The main advantage of this data is that the GEM distinguishes between necessity-driven and opportunity-driven entrepreneurs. The former are pushed into entrepreneurship due to the fact that they need an income but have no other options. The latter are individuals that get involved in the process of starting a firm to pursue an opportunity despite other options that they have. Opportunity driven entrepreneurs are generally associated with innovative new firms that have the potential and ambition to grow and create jobs. This distinction is made on the motives of an individual to get involved in the process of starting a firm, which he or she has indicated in the survey. Our data covers 27 EU Member States on the NUTS2 level for the years 2007 until 2015 and are available at Utrecht School of Economics.

Regional Data

For regional control variables this dataset relies on Eurostat (2016). The annual growth rate of gross domestic product and value added at current market prices, population, and population density on the NUTS2 level for the years 2006 - 2014 are drawn from this source.

Variety

To understand how variety affects regional development, and the role of entrepreneurship in this, we have to construct a variety variable. Variety refers to the sectoral composition of a regional economy. Following Frenken, Van Oort, & Verburg (2007) we want to distinguish between related and unrelated variety, as variety in related sectors, all else equal, is expected to yield more knowledge spillovers. To be able to distinguish between related and unrelated variety, a detailed sectoral classification is needed to capture relatedness between sectors (preferably 4-digit or 5-digit hierarchical sectoral classification). However, Pan-European data available for NUTS2 regions like Cambridge Econometrics only distinguishes between employment in 2-digit sectors. Hence, these are ill-suited to use. While some have used these data to distinguish between related (2-digit) and unrelated (1-digit) variety (De Groot et al., 2015), one should wonder whether sectors assigned to the same 2-digit level are sufficiently related to capture the spillovers we are after.

Following van Oort et al. (2015) and Cortinovis & Van Oort (2015), we therefore use the ORBIS dataset provided by Bureau van Dijk (2015). This dataset contains pan-European firm data that can be aggregated to the appropriate sectoral levels for EU NUTS-2 regions to compute regional variety measures. We can exploit these data to analyse the effect of related and unrelated variety on entrepreneurship.

The advantage of using the ORBIS dataset is that it includes employment data for about 80 million firms in Europe for the period 2006 until 2014. It also contains information on the sector in which the firms operate based on the 4-digit NACE classification scheme. The geographic location of firms is included at the NUTS2-level (which in turn can be aggregated to national statistics). There are, however, some disadvantages in using this dataset as well. A point of concern is that the distribution of firms in terms of their size is not representative.

Only those firms that are obligated to report annually are included. This means that smaller firms are not included in the data. In order to correct for this bias, Cortinovis and Van Oort (2015) opted to rescale the employment variables to match the Eurostat employment rates. We need to investigate if such rescaling is appropriate for our purposes. In calculating the related and unrelated variety at the regional level we have to weight the importance of this firm size bias against the error we would introduce by rescaling. Our measure should proxy for the relatedness of variety in the region, which is not likely to be much affected by the unobserved, smallest firms.

The variety can be calculated with an entropy measure, and related and unrelated variety can then be measured by decomposing the entropy measure in an unrelated and related part. In this, we follow Frenken et al. (2007) who were the first to apply these entropy measures at the regional economic level, building on Jacquemin and Berry (1979) who applied these entropy measures to measure related and unrelated variety at the firm-level.

For the calculation of unrelated variety we make the assumption that sectors that belong to different 2-digit sectors are unrelated. Additionally, 4-digit sectors within each of these 2-digit sectors are assumed to be related because they belong to the same sector and therefore are likely to use similar technologies. The approach used by Frenken et al. (2007) computes the entropy of the employment shares at a given digit level of an industry classification scheme. The 4-digit shares p_i are summed to compute the 2-digit shares P_g:

$$(1)\quad P_g = \sum_{i \in S_g} p_i$$

Unrelated Variety (UV), the entropy *between* 2-digit sectors, is given by:

$$(2)\quad UV = \sum_{g=1}^{G} P_g \log_2 \left(\frac{1}{P_g}\right)$$

Entropy, H_g, *within* each 2-digit sector, is given by:

$$(3)\quad H_g = \sum_{i \in S_g} \frac{p_i}{P_g} \log_2 \left(\frac{1}{p_i / P_g} \right)$$

And Related Variety is the weighted sum of this Entropy, where the weights are given by employment shares, such that:

$$(4)\quad R = \sum_{g=1}^{G} P_g H_g$$

Equation (4) thus sums the entropy values within each 2-digit sector, weighted by the respective employment shares to reflect the relative importance of each 2-digit sector in the regional economy.

The tables 1, 2, and 3 below contain metadata and descriptives for the dataset. Table 1 describes the variables, table 2 shows the summary statistics, and table 3 contains the correlation matrix.

Table 1: Variable description

Variable	Description
Year	Years 2006-2014
Region	NUTS2 region identifier
Related variety	Weighted sum of entropy at the 4-digit level within each 2-digit sector.
Unrelated variety	Entropy at the 2-digit level.
TEA	Percentage of the working age population that is involved in Total Early-stage Entrepreneurship
TEA (Opportunity)	Percentage of the working age population that is involved in Total Early-stage opportunity-driven Entrepreneurship
GDP	Gross Domestic Product in millions of euros at PPP
Population	Population on January 1st
Population density	Population density of the average population per square kilometer
Employment	Rate of employment of the working age population (15-65 years)
Unemployment	Rate of unemployment of the working age population (15-65 years)

Table 2: Summary statistics

Variable	Mean	Std. Dev.	Min	Max	Observations
Year	2010	2.583	2006	2014	2,520
Region	140.5	80.973	1	280	2,520
Related variety (RV)	3.589	0.799	0.053	4.785	2,520
Unrelated variety (UV)	3.148	0.352	0.545	3.753	2,520
TEA	6.167	2.983	0.915	25.826	1,570
TEA (Opportunity)	4.451	2.080	0.287	11.769	1,574
GDP	50,140	58,557	1,075	649,101	2,302
Population	1,912,625	1,551,480	27,000	12,000,000	2,265
Population density	379.730	950.739	3	10,438.2	2,495
Employment	66.162	8.014	38.9	85.1	2,447
Unemployment	8.285	4.774	0	37	2,445

It is clear from the descriptives that the dataset contains a wide variety of regions, ranging from sparsely populated rural regions with low economic activity to densely populated urban regions with high levels of economic activity. In our analysis of the data such variation is important to identify the effects we are after. None of the values listed above, however, seem to be out of reasonable range.

Table 3: Correlation matrix

	RV	UV	TEA	TEA (Opp.)	GDP	Pop.	Pop. den.	Emp.	Unemp.
Related variety (RV)	1.000								
Unrelated variety (UV)	-0.190	1.000							
TEA	-0.113	0.045	1.000						
TEA (Opportunity)	-0.145	0.093	0.825	1.000					
GDP	0.076	0.243	-0.158	-0.071	1.000				
Population	0.222	0.183	-0.049	-0.075	0.833	1.000			
Population density	-0.241	0.225	0.040	0.097	0.146	0.015	1.000		
Employment	0.021	0.045	-0.139	0.012	0.155	-0.150	-0.073	1.000	
Unemployment	-0.085	0.130	0.081	0.007	-0.031	0.161	0.121	-0.486	1.000

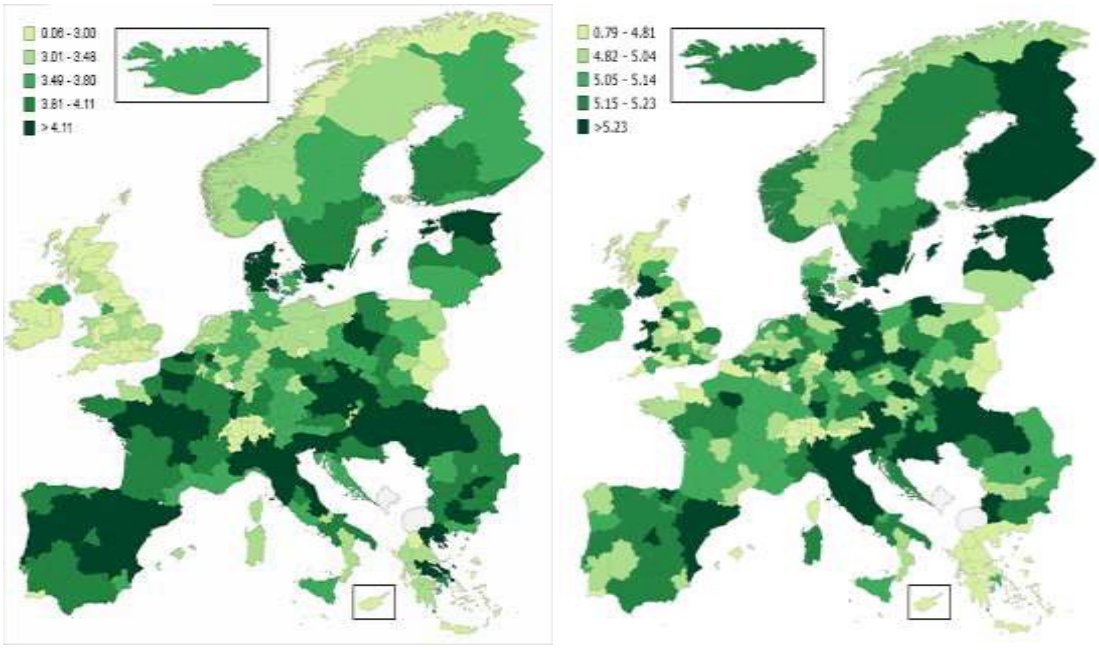
The data contains 280 NUTS2 regions divided over 32 European countries (table 4). For each region there are 9 observations for the years 2006 until 2014.

Table 4: Countries in the dataset (number of regions between brackets)

Austria (9)	Greece (13)	Norway (7)
Belgium (11)	Hungary (7)	Poland (16)
Bulgaria (6)	Iceland (1)	Portugal (5)
Croatia (2)	Ireland (2)	Romania (8)
Cyprus (1)	Italy (21)	Slovakia (4)
Czech Republic (8)	Latvia (1)	Slovenia (2)
Denmark (5)	Liechtenstein (1)	Spain (18)
Estonia (1)	Lithuania (1)	Sweden (8)
Finland (5)	Luxemburg (1)	Switzerland (7)
France (22)	Malta (1)	United Kingdom (37)
Germany (37)	Netherlands (12)	

The maps in figure 1 depict the related- and unrelated variety measures for the most recent year in the dataset, which is 2014. The figure on the left shows related variety, whereas the figure on the right shows unrelated variety.

Figure 1. Related variety (left) and unrelated variety (right) in 2014.

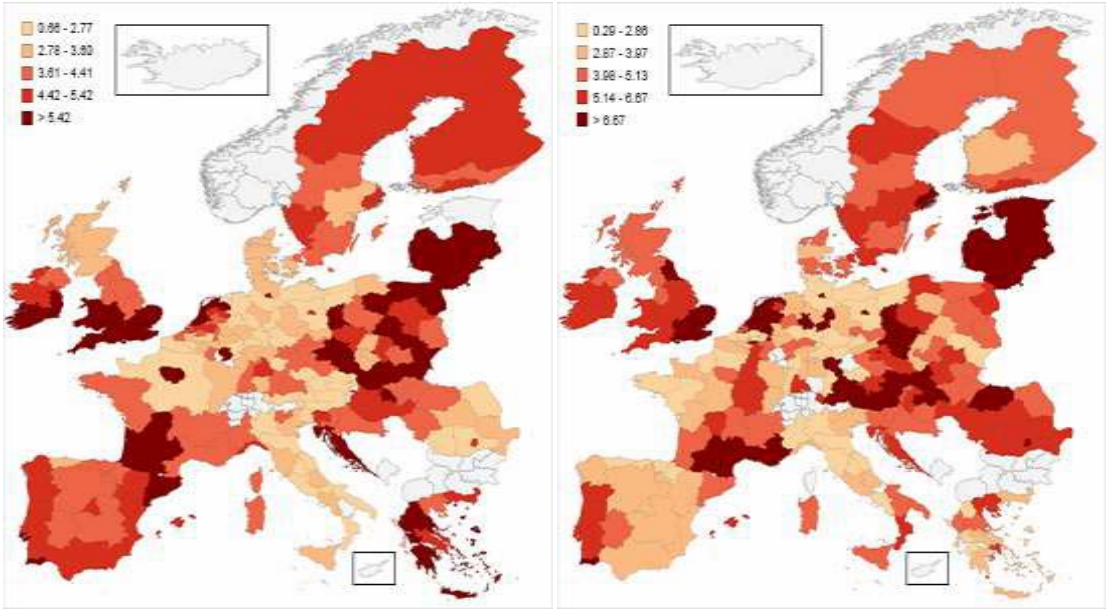


What these maps show is that the regions with high levels of related variety exhibit low levels of unrelated variety and *vice versa*. The correlation between the variety measures is even negative. Part of this negative correlation results from the decomposition, which splits total variety into related and unrelated variety. However, there are cases of regions with

both high levels of related variety and high levels of unrelated variety (e.g., the region of Catalunya in Spain) and regions with both low levels of related variety and low levels of unrelated variety (e.g., the region of Crete in Greece). This suggests that the level of aggregation chosen as cut-off point between unrelated and related variety is a relevant one.

Figure 2 depicts the average rate of opportunity-driven entrepreneurship for two periods of time. The left hand side shows the average of 2007 until 2011, whereas the right side shows the average of 2011 until 2014.

Figure 2. Average rate of opportunity-driven entrepreneurship between 2007-2011 (left) and 2012-2014 (right).



What is clear in this picture is that higher rates of opportunity-driven entrepreneurship appear to be more frequent in urban regions, including the "usual suspects" of London and Paris, but also other urban regions hosting the cities of Amsterdam, Bucharest, Hamburg, Rotterdam and Vienna. The regions with the lowest amount of entrepreneurship during this period can be found in more rural areas of France and Germany, as well as many Italian regions.

The correlation matrix provides a first look at the relationship between variety and entrepreneurship. The correlations between the two variety measures and the rate of opportunity-driven entrepreneurship is very low, and for related variety even negative. Whether variety enhances entrepreneurship, however, can only be determined if we control for other factors, which remains to be done.

Some South-East European countries are excluded in the maps of figure 1 (i.e. Bosnia and Herzegovina, Serbia, Albania, Macedonia, and Montenegro). This is because the ORBIS dataset does not contain enough information about these specific countries to construct the variety measures. Countries that are included in the most recent NUTS classification of 2013 but are not included in the ORBIS dataset are in the maps of figure 1 but are not shaded.

As some countries do not participate in the GEM, there are no data on the rate of entrepreneurship in these countries. Countries that are included in the most recent NUTS classification of 2013 but did not participate in the GEM survey are in the maps of figure 1 but are not shaded. Data on entrepreneurship is not available in the period of 2007-2011 for countries Austria, Bulgaria, Switzerland, Cyprus, Estonia, Iceland, Liechtenstein, Luxembourg, Malta, and Norway. For the period 2012-2014 the entrepreneurship data is not available for the countries Bulgaria, Switzerland, Cyprus, Iceland, Liechtenstein, Luxembourg, Malta, and Norway.

Products

Related variety measures based on entropy measures of sector data as presented above have the disadvantage that relatedness is discrete (two sectors are either related or unrelated) and pre-defined by the hierarchical sector classification. Hidalgo et al. (2007) instead introduced the concept of product space, using a proximity indicator based on how often two products co-occur in countries' export portfolios. In this way, one obtains a continuous variable of relatedness with each two products being more or less related to each other, and also a measure that evolves over time. Hidalgo et al. (2007) then argue that if a country has a comparative advantage in producing a certain product, chances are high it will also have a comparative advantage in products that are related to it in terms of, for

instance what kind of institutions, infrastructure, physical factors, or technology is needed. A region or country may then be expected to develop new products, which are related to products it already is producing (see also Content and Frenken, 2016).

A downside of using export data, however, is that these data are collected at the country level. Still, one could argue that what is related at the national level is also related at the regional level. Hence, we will use the updated dataset used in the original study by Hidalgo et al. (2007). The number of observations, however, will be many as the unit of analysis will not be a country as such, but each product-country pair. European countries can be analyzed more specifically, for example, by interacting variables with dummies of particular EU countries, or the EU as a whole. One may expect that EU countries tend to diversity in more related sectors than for example the U.S., given that EU "coordinated-market" institutions are less flexible in reallocating capital and labour from one sector to an unrelated sector compared to US "liberal-market institutions" (Boschma and Capone 2015).

The dependent variable will be the emergence of new export specializations in each country and each year, and the relatedness of existing specializations and the main independent variable, with the national level of entrepreneurship as moderating variable. One hypothesis may be that entrepreneurship allows one to diversify into more unrelated products, as entrepreneurs are less bound to existing knowledge and institutions than existing firms and better able to perceive new connections and combinations between previously rather disconnected knowledge bases.

In order to measure proximity Hidalgo et al. (2007) start by using export data to measure whether a country has a Revealed Comparative Advantage (RCA) in a certain product, which is given by

$$(5)\quad R\hat{A}_{c,\hat{}} = \frac{\frac{\hat{}_{c,\hat{}}}{\sum_c \hat{}_{c,\hat{}}}}{\frac{\sum_{\hat{}} \hat{}_{c,\hat{}}}{\sum_{\hat{},c} \hat{}_{c,\hat{}}}}$$

where $x_{c,i}$ is the export value of country c in product i. When RCA is larger than one, the share of exports of country c in product i is larger than the share of that same product in the world trade. Hidalgo et al. (2007) then define that when a country's RCA for a certain product is greater than or equal to one, the country is a specialized exporter of that product. Using this measure, the proximity between two products is then given by

$$(6)\quad \varphi_{\hat{},\hat{}} = \hat{}\{P(R\hat{A}_{\hat{}}|R\hat{A}_{\hat{}}), P(R\hat{A}_{\hat{}}|R\hat{A}_{\hat{}})\}$$

where $P(RCA_i|RCA_j)$ is the conditional probability that a country exports product i when it already is exporting product j. Equation (6) is interpreted as the proximity between two products i and j and defines that proximity as the minimum of the conditional probability of a country exporting a product given that it exports the other. As mentioned above this way of measuring proximity between two products has the advantage that it creates a continuous variable of relatedness. The motivation for measuring proximity in this manner is that if two products at a certain time require for instance similar institutions, infrastructure, physical factors, or technologies they will be more likely to be manufactured by the same country at that time. To test the hypothesis that a country is more likely to become a specialized exporter in a product that is related to the products it already is exporting, Hidalgo et al. (2007) developed a measure of how close a country is to each of the products it is not already exporting with a comparative advantage. This is measured by product density, and is given by

$$(7)\quad \hat{e}\hat{}\hat{}\hat{}\hat{}_{\hat{},c} = \frac{\sum_{\hat{}} \varphi_{\hat{},\hat{}} R\hat{A}_{c,\hat{}}}{\sum \varphi}$$

$$=$$

$$\hat{}_{\hat{},\hat{}}$$

where proximity between product *i* and *k* is given by $\phi_{i,k}$ and $RCA_{c,k}$ is one if country *c* has a comparative advantage in exporting product *k*, or zero if it has not. The density around a certain product will be high if a country is already exporting most of the related products with a comparative advantage and can be one at the maximum, in which it exports all of the related products. At the minimum it will be zero as it exports none of the related products with a comparative advantage. We aim to cross-reference our measure of related variety with the measures by Hidalgo et al. (2007) for robustness and further scrutiny of our results.

Tasks

Variety is generally depicted at a feature of the sectoral composition of export mix of a region. Alternatively, one may also analyse the variety is tasks executed in a region. Indeed, one can expect knowledge to flow and to be recombined between tasks as much as between sectors or products. The focus on tasks in understanding economic development has become increasingly relevant (Hanson 1994, Baldwin and Robert-Nicoud 2014). As a result, regions may focus on excelling in specific tasks involved in multiple value chains rather than on excelling in specific products. A region specialized in software development, for example, can serve many different sectors and products.

One way global value chains are mapped is using world Input-Output tables (Timmer et al. 2014). Indeed, using such data one can reconstruct input-output relations between sectors and countries. However, such data only lends itself to national level analysis given that more fine-grained data are not available at a pan-European regional level.

As an alternative, we propose to explore the regional-level sector data at the 4-digit level derived from the ORBIS data for analyses reasoning from tasks rather than sectors. To do so, we will have to proxy particular tasks by particular 1-digit or 2-digit sectors and then compute the related variety within such a sector. We propose to do this by selecting the NACE sub-sectors that are commonly associated with generic tasks present in many value chains. Here, we propose the following subset:

C28 - Manufacture of machinery and equipment n.e.c.

H - Transporting and storage

J62 - Computer programming, consultancy and related activities

J63 - Information service activities

K64 - Financial service activities, except insurance and pension funding

M - Professional, scientific and technical activities

As we already have computed the related variety within these categories, zooming in on more generic task sectors we might be able to assess whether regions with a related variety in tasks realise higher regional growth rates.

3. Outlook

Using the GEM and ORBIS databases, we constructed pan-European regional measures of variety and entrepreneurship as to explore and analyse the relationship between variety, entrepreneurship and regional development. Variety is further decomposed by related and unrelated variety. Spillovers arising from the recombination of ideas, skills and technologies by entrepreneurs are expected to arise among related sectors, creating regional growth.

The specific relations we are interested to explore further is in first instance the relationship between related and unrelated variety on the one hand, and opportunity-driven entrepreneurship on the other. Here, we expect both types of variety to enhance entrepreneurship, but more so for related variety than for unrelated variety given that one expects the former measure to pick up most of the spillovers among sectors.

In a second stage, we want to analyse the joint effect of related variety and entrepreneurship on regional growth. Here, two lines of argument can be followed. One can understand entrepreneurship as a moderating variable and analyse whether the entrepreneurial activity enhances the extent to which regional economic growth is fuelled by the variety of the regional economy, be it in terms of sectors or tasks. This would follow the study set-up of the work by Fritsch and Kublina (2016) on German regions, and would extend

this work to the European level. One can also see variety as a mediating variable between entrepreneurship and regional growth. In such a set-up, one can test whether related and unrelated variety enhances regional growth directly and/or indirectly via entrepreneurship. Following this reasoning, one understands variety as enhancing the opportunities to become an entrepreneur, which in turns leads to regional growth. Combining both approaches we would specify a model in which both entrepreneurship and variety have a direct effect on regional growth and investigate the interaction. Given the limited time dimension of our data it will prove challenging to make strong causal inferences, but we can still test some of the more straightforward hypotheses.

Finally, we will analyse the process leading to the addition of new economic activities as measured by the emergence of a new export specialization. This exercise will be done at the country level given the lack of export data at the regional level. Still, given that every European country consists of one or more regions in our dataset, we can link the two analyses. Following Hidalgo et al. (2007), we will take the emergence of one new export specializations as dependent variable, and the proximity of existing export specializations to this new product as a "related-variety" variable. Entrepreneurship, then, can be hypothesized to moderate this relation, that is, to compensate for a relative lack of related variety. Put differently, if a country has many entrepreneurs, one would expect that the new specializations are less related to the existing base of export products in a country than if for countries with few entrepreneurs.

References

Baldwin, R., & Robert-Nicoud, F. (2014). Trade-in-goods and trade-in-tasks: An integrating framework. *Journal of International Economics*, *92*(1), 51-62.

Boschma, R., Capone, G. (2015) Institutions and diversification: Related versus unrelated diversification in a Varieties of Capitalism framework, *Research Policy*, 44(10), 1902-1914.

Bureau van Dijk (2015). *ORBIS*. Amsterdam: Bureau van Dijk.

Caragliu, A., de Dominicis, L., de Groot, H.L.F. (2016) Both Marshall and Jacobs were right! *Economic Geography* 92(1), 87–111.

Colombelli, A. (2016). The impact of local knowledge bases on the creation of innovative start-ups in Italy. *Small Business Economics*, in press.

Content, J., Frenken, K. (2016). Related variety and regional growth: a review, FIRES deliverable D3.1, 26 February, Utrecht University.

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.

Eurostat (2016). *Regional statistics by NUTS classification (reg)*. Eurostat: Luxembourg.

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

Fritsch, M., Kublina, S., (2016) *The Role of Absorptive Capacity and Entrepreneurship*, Jena Economic Research Papers No 2016-009

Groot, H. L. F. de, Poot, J., & Smit, M. J. (2015). Which Agglomeration Externalities Matter Most and Why? *Journal of Economic Surveys*, in press.

Hanson, G. H. (1994). *Localization economies, vertical organization and trade* (No. w4744). National Bureau of Economic Research.

Hartog, M., Boschma, R., & Sotarauta, M. (2012). The Impact of Related Variety on Regional Employment Growth in Finland 1993–2006: High-Tech versus Medium/Low-Tech. *Industry and Innovation*, 19(6), 459–476.

Jacquemin, A.P., and C.H. Berry (1979), Entropy measure of diversification and corporate growth. *Journal of Industrial Economics* 27(4), 359–369.

Oort, F. van, Geus, S. de, Dogaru, T. (2015). Related variety and regional economic growth in a cross-section of European urban regions. *European Planning Studies* 23(6), 1110–1127.

Shane S. (2008). *The Illusions of Entrepreneurship: The Costly Myths that Entrepreneurs, Investors, and Policy Makers Live By*. New Haven, CT: Yale Univ Press.

Timmer, M.P., Aziz Erumban, A., Los, B., Stehrer, R., de Vries, G.J. (2014). Slicing up global value chains, *Journal of Economic Perspectives* 28(2) 99–118.