Papers in Evolutionary Economic Geography

13.02

Related Variety, Unrelated Variety and Technological Breakthroughs: An analysis of U.S. state-level patenting

Carolina Castaldi, Koen Frenken, Bart Los

Utrecht University
Urban & Regional research centre Utrecht

Related Variety, Unrelated Variety and Technological Breakthroughs: An analysis of U.S. state-level patenting

Carolina Castaldi [a]
Koen Frenken [a]
Bart Los [b]

[a] School of Innovation Sciences, Eindhoven University of Technology,
PoBox 513, 5600MB, Eindhoven, The Netherlands

[b] Groningen Growth and Development Centre,
PoBox 800, 9700 AV Groningen, The Netherlands; b.los@rug.nl

Corresponding author: c.castaldi@tue.nl

Abstract: We investigate how variety affects the innovation output of a region. Borrowing arguments from theories of recombinant innovation, we expect that related variety will enhance innovation as related technologies are more easily recombined into a new technology. However, we also expect that unrelated variety enhances technological breakthroughs, since radical innovation often stems from connecting previously unrelated technologies opening up whole new functionalities and applications. Using patent data for US states in the period 1977-1999 and associated citation data, we find evidence for both hypotheses. Our study thus sheds a new and critical light on the related-variety hypothesis in economic geography.

Keywords: recombinant innovation, regional innovation, superstar patents, technological variety, evolutionary economic geography

JEL-codes: O31; R11

1. Introduction

Innovation is commonly held to be the key factor in regional development, underlying short-run productivity gains and long-run employment growth through new industry creation. Since innovation processes draw on knowledge that is often sourced locally (ALMEIDA and KOGUT, 1999; STUART and SORENSON, 2003; BRESCHI and LISSONI, 2009), regional development is essentially an endogenous process with strong path dependencies (IAMMARINO, 2005; RIGBY and ESSLETZBICHLER, 2006) akin to an evolutionary branching process (FRENKEN and BOSCHMA, 2007; NEFFKE et al., 2011).

In so far as knowledge is drawn from a variety of sectors, as in "recombinant innovation" (WEITZMAN, 1998), the sectoral composition of a region will affect the rate and direction of technical change in regions (EJERMO, 2005). In this context, it has been argued that the more sectors are related, the more easily knowledge created in one sectoral context can be transferred to other sectoral contexts. Hence, variety *per se* may not support innovation; rather it is "related variety" (NOOTEBOOM, 2000; FRENKEN et al., 2007) that provides the basis for knowledge spillovers and recombinant innovation, spurring productivity and employment growth. The related-variety hypothesis has motivated a large number of other empirical studies on the effect of related variety in sectoral composition on regional productivity and employment growth (ESSLETZBICHLER, 2007; FRENKEN et al., 2007; BOSCHMA and IAMMARINO, 2009; BISHOP and GRIPAIOS, 2010; QUATRARO, 2010; ANTONIETTI and CAINELLI, 2011; BRACHERT et al., 2011; QUATRARO, 2011; BOSCHMA et al., 2012; HARTOG et al., 2012; MAMELI et al., 2012). Results tend to show that related variety indeed supports productivity and employment growth at the regional level, though some studies suggest that the effects are sector-specific (BISHOP and GRIPAIOS, 2010; MAMELI et al., 2012).

In putting forward their hypothesis on related variety, FRENKEN et al. (2007) associated related variety as being supportive of knowledge spillovers and recombinant innovation, which in turn would support regional growth. In their analysis of the impact of related variety on productivity and employment growth, however, they did not provide direct evidence on the relationship between related variety and innovation processes as such. Hence, the question remains open whether related variety supports innovation[1] (TAVASSOLI and CARBONARA, 2012). In this paper, we aim to further develop the notion of related variety and its effect on innovation. We do so within a theoretical framework that explicitly distinguishes between related and unrelated variety and predicts differential effects of the two

[1] Actually, we focus on invention, since we do not address issues of successful commercialization, but solely focus on technological attainments. Throughout the paper, we will use the terms innovation and invention interchangeably, since the theory of recombinant innovation has been framed in terms of innovation rather than invention.

types of variety on innovation processes. We take issue with the notion that related variety supports all the kinds of innovation. Instead, we argue that related variety is supportive of the bulk of innovations which incrementally builds on established cognitive structures across 'related' technologies, while unrelated variety provides the building blocks for technological 'breakthroughs' stemming from combinations across unrelated knowledge domains. Since such radical innovations often stem from connecting previously unrelated technologies, these innovations lead to whole new functionalities and applications, and span new technological trajectories for their further improvement (DOSI 1982). As a result, the unrelated technologies lying at the root of the breakthrough innovations, become more related over time.

Within this new theoretical framework, we test two hypotheses. The first hypothesis contends that related variety of the existing knowledge stock in a region enhances its overall innovation rate, while a high degree of unrelated variety does not have effects. The second hypothesis states that unrelated variety of the regional knowledge base supports the rare breakthrough innovations, while related variety does not have such an effect.

We use a criterion based on the numbers of citations to a patent as included in subsequent patent documents (so-called forward citations) to operationalize the concepts of incremental innovation and breakthrough innovations (SILVERBERG and VERSPAGEN, 2007; CASTALDI and LOS, 2012). The dataset contains all utility patents granted by the US Patent and Trademark Office between 1977 and 1999, for which the first inventor resided in the United States. Information on the locations of first inventors is used to assign patents to U.S. states, which we use as regional units. To construct variables regarding various types of variety of the regional knowledge base, we used technological classification schemes at different levels of aggregation, as designed by the US Patent and Trademark Office. The actual construction of related-variety and unrelated-variety variables is rooted in entropy statistics (FRENKEN et al., 2007).

Our results show a positive effect of related variety on regional innovation in general, and a positive effect of unrelated variety when looking at regions' capability to forge breakthrough innovations. This finding is shown to be robust for the inclusion of spatially lagged R&D variable, that is, the sum of R&D investments in neighbouring states.

The rest of this paper is structured as follows. Section 2 gives a brief overview of theoretical concepts regarding the interplay of existing pieces of knowledge in recombinant innovation processes. We introduce our methods in Section 3, which includes a discussion of the procedure adopted to make a distinction between incremental innovations and breakthrough innovations. In Section 4, we show how the numbers of produced breakthrough innovations varies across states and provide indications of differences in the variety of their knowledge

bases, before testing the hypotheses using econometric estimation techniques. Section 5 concludes.

2. Variety, recombination and innovation

Technological innovation is commonly understood to be a cumulative process, in which most new artefacts are being invented by re-combining existing technologies in a new manner (BASALLA, 1988; ARTHUR, 2007). The recombination is a novelty in itself, but could only emerge given the pre-existence of the technologies being recombined. As a recent and telling example, one can think of smart phones, which combine technologies related to batteries, chips, antennas, audio, video, display, and Internet. In this context, Schumpeter famously spoke of innovation as the bringing about of new combinations ("Neue Kombinationen"), an idea which continues to inspire evolutionary theorising in economics (BECKER et al., 2012). A more recent and very similar concept is that of "recombinant innovation" defined as "the way that old ideas can be reconfigured in new ways to make new ideas" (WEITZMAN, 1998, p. 333). This concept motivated new formal models of innovation within the evolutionary economics literature, including one on optimal variety in recombinant innovation (VAN DEN BERGH, 2008) and another on the role of recombinant innovation in technological transitions (FRENKEN et al., 2012).

In a regional context, it follows from the notion of recombinant innovation that, to the extent that innovation processes draw on geographically localised knowledge, regions with a more diverse stock of knowledge would have a greater potential for innovation. This is in line with Jacobs' argument that cities hosting many different industries would experience more innovation as the exchange of knowledge by people with different backgrounds would lead to more new products and processes. As JACOBS (1969, p. 59) observed, "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." This mechanism was later labelled as Jacobs externalities, which refer to positive externalities arising from the co-location of different sectors (GLAESER et al., 1992).

FRENKEN et al. (2007) added to Jacobs' argument that regions hosting related industries can more easily engage in recombinant innovation. Such related industries draw from different but not completely disconnected knowledge bases. In the words of FRENKEN et al. (2007, p. 687), related variety "improves the opportunities to interact, copy, modify, and recombine ideas, practices and technologies across industries giving rise to Jacobs externalities". One expects the related-variety hypothesis to hold for innovation in general. However, it should be recognized that unrelated varieties can sometimes be combined successfully as well. Such innovations render pieces of knowledge that were previously unrelated to become related, in

the form of an artifact or service exemplar that paves the way for future innovations to follow suit. Indeed, while recombinant innovation among previously unrelated domains is more likely to fail, such innovations, when successful, are also more likely to be of a radical nature as recombination across unrelated technologies can lead to complete new operational principles, functionalities and applications (FLEMING, 2001; SAVIOTTI and FRENKEN, 2008).

Hence, the opposition between related and unrelated variety can be misleading, and both types of variety can lead to innovation. Related variety would raise the likelihood of innovations in general, but unrelated variety would raise the likelihood of breakthrough innovations, which in itself are rare. It is precisely in this context that DESROCHERS and LEPPÄLÄ (2011, p. 859) proposed "to consider the essence of innovation to be about making connections between previously unrelated things." Following this reasoning, one can understand the relatedness structure among technologies are evolving, albeit slowly, in a way that is driven by radical innovation that render previously unrelated technologies to become related (Figure 1).

The famous example of the car can help to illustrate the idea. In car technology various extant technologies were being recombined, notably engine technology, bicycle technology and carriage technology. These technologies were largely unrelated at the time the car technology was still in its infancy, but gradually became related through the development of the car. The reason why unrelated technologies can become related is that the new, recombinant technology provides a new context for extant technologies to be related, that is, to be recombined.

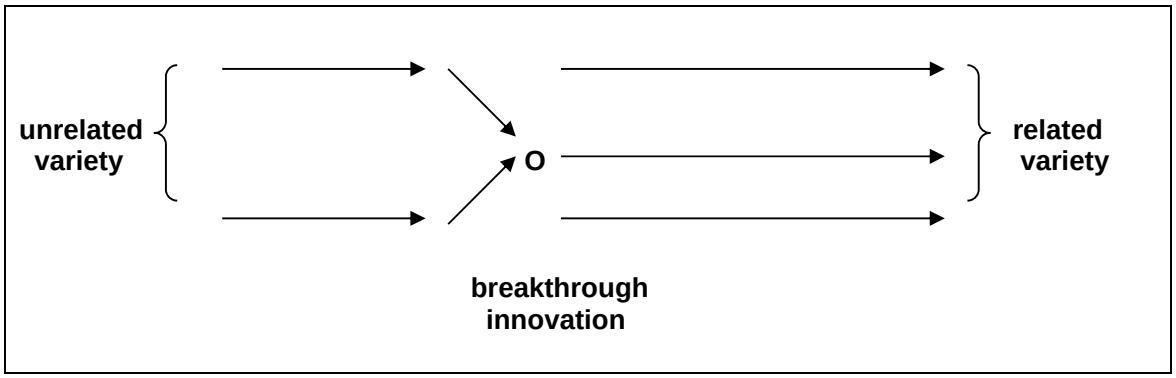


Figure 1. Breakthrough innovation turning unrelated into related variety

Turning to the regional level, one can expect regions with high levels of related variety to outperform regions with low levels of related variety in terms of the sheer number of inventions they produce. However, when it comes to breakthrough inventions, regions with

high levels of unrelated variety are expected to outperform regions with low levels of unrelated variety. These are the two hypotheses that will guide the remainder of our study.

3. Research design

We test our hypotheses using patent data. Their use to trace innovation is widespread and by now reasonably accepted. Patents have a number of attractive features with regard to the measurement and classification of inventive output. These particularly include the facts that formal novelty requirements have to be met to have a patent granted and that all patents are assigned to technological classes by independent and knowledgeable experts (SMITH, 2005). A more debated issue is how to quantify success in producing breakthrough innovations in a systematic way. Recently, empirical research on innovation has offered a number of alternatives, all basically aimed at capturing the value of patents (VAN ZEEBROECK, 2011). Citations received by patents (forward citation numbers) are a common indicator for patent value, as suggested already by TRAJTENBERG (1990). Many researchers have measured breakthrough inventions by considering the top-cited patents in a given subpopulation (e.g., AHUJA and LAMPERT, 2001; SINGH and FLEMING, 2010). These subpopulations are often chosen as cohorts of patents in a technological field or subfield, to provide a fair comparison between patents of different age ("young" patents did not have much time to receive citations) and technological field (in our period of analysis, many more patents were granted in a category like Chemical than in Computers and Communications, as a consequence of which Chemical patents generally receive more citations than Computers and Communications patents, see HALL et al., 2002). For this study, we use a refined methodology proposed by CASTALDI and LOS (2012) to identify what they term 'superstar patents'. The basic idea behind this methodology is to endogenously derive the share of superstars in a subpopulation of patents by exploiting statistical properties of the frequency distribution of forward citation numbers, which are characterized by a fat tail. This approach is original, as most studies use exogenously fixed (identical across years and technologies) criteria to distinguish between breakthrough and regular innovations instead, by defining breakthroughs as the patents belonging to the top 5% or top 1% quantiles of the citations distributions.

The statistical properties that spurred the initial application of the method were highlighted by SILVERBERG and VERSPAGEN (2007). They showed that a log-normal distribution fits most of the forward citations distribution for patents quite well, except for the tail: the numbers of received citations of highly cited patents rather follow a Pareto distribution. This implies that there are a few patents for which the "citations-generating" process is different. The technologies underlying such patents act as focusing devices for technological developments within new technological paradigms (DOSI, 1982). By estimating the number

of citations needed by a patent to fall into the Pareto tail of the forward citations distribution, CASTALDI and LOS (2012) classify US patents registered at USPTO as either superstars or not.[2] This estimation relies on a modified version of the estimation routine in SILVERBERG and VERSPAGEN (2007), based upon the so-called Hill estimator. The procedure also ensures that only patents with the same application year and belonging to an identical technological subcategory are compared. USPTO patents have been classified by HALL et al. (2002) in 6 broad technological categories and 36 technological subcategories, each corresponding to 417 even more disaggregate patent classes (HALL et al. 2002, pp. 41-42). The classification is part of the NBER Patent Citation database and its updates and allows assigning each registered patent to one single category, one single subcategory and one single patent class.

For our purposes here, we wish to count patents and superstar patents across regions. US patents included in the NBER database can be assigned to the US state of first inventor. The state will be our definition of a region in this study.[3] For each state and each year from 1976 to 1999, we have the number of total granted patents applied for in that year at the USPTO by inventors in that state and we also have estimates of how many of the total patents are superstar patents.[4] As our hypotheses relate to explaining regional innovative output, we work with two dependent variables for each state *i*:

a) the total number of granted patents with application year *t*, as a proxy for the general innovation performance of a state ($NUMPATENTS_{it}$); and
b) the share of superstar patents in all patents of the state with application year *t*, as a proxy for the ability to produce breakthrough innovations ($SHARESUPER_{it}$).

We choose to consider shares of superstars rather than absolute numbers, since shares tell us something about the type of innovative activity: shares indicate revealed comparative (dis)advantages in breakthrough innovation.

CASTALDI and LOS (2012) analyse the geographical concentration of superstar patents across US states and find that the regional clustering of superstar patents is much higher than for non-superstar patents. Apparently, companies locate their search for breakthrough

[2] For recent patents, the variation in the numbers of received citations is often insufficient to obtain accurate estimates of the number of citations required to fall into the Pareto tail. The fact that superstar patents tend to gather citations over a much longer period of time than regular patents is the main culprit for this. CASTALDI and LOS (2012) proposed an estimation framework to predict the odds of a "young" patent becoming superstar at later age, based on characteristics of the citations received already.

[3] With state-level data, one can control for state-specific fixed effects such as institutions, including state regulations concerning products and the labour market. Compared to smaller spatial units of analysis, state-level analysis also has the advantage of having a substantial number of breakthrough innovations per state.

[4] The original NBER Patent Citation database covers all patents granted at USPTO in the years 1975-1999. Bronwyn Hall updated the NBER database in 2002 and the NBER itself has published a new version with data until 2006. Since the latest update does not contain information about the location of inventors, we use the 2002 database.

innovations in very specific places, while the production of regular innovations happens in many more places. Their descriptive results already indicate that explaining regional performance in terms of breakthrough innovation requires different hypotheses than explaining regional innovative performance in more general terms.

We now turn to our explanatory variables. The key independent variables in our model will be measures of regional variety in innovative activity. Again, we use patent data, as patents tell us something about the technological fields in which states contribute innovations. In line with previous work, we measure variety with entropy indicators (GRUPP, 1990; FRENKEN, 2007). Entropy captures variety by measuring the "uncertainty" of probability distributions. Let E_i stand for the event that a region is patenting in a given technological field i and let p_i for the probability of event E_i to occur, with i=1, ..., n. The entropy level H is given by

$$H = \sum_{i=1}^{n} p_i \ln \left(1/p_i\right) \qquad (1)$$

with

$$p_i \ln \left(1/p_i\right) = 0 \qquad \text{if } p_i = 0$$

The value of H is bounded from below by zero and has a maximum of ln(n). H is zero if p_i = 1 for a single value of i and p_i = 0 for all other i. In the context of this study, such a situation would occur if a state would have all its patents in a single patent class. If a patent would be drawn from this state's patent portfolio, uncertainty about the patent class to which it belongs would be non-existent. The maximum value of ln(n) is attained if all p_i values are identical. In terms of our application, such a situation emerges if the shares of all patent classes in a state's patent portfolio are the same. If a patent were drawn at random from such a portfolio, the uncertainty about the patent class to which it belongs would be the largest.

Apart from its roots in information theory (see Theil 1972), a very appealing feature of entropy statistics is that overall entropy can be decomposed in entropy measures at different levels of aggregation (Frenken 2007). This allows us to construct variables that represent different levels of relatedness of variety in technological capabilities of states, as reflected in patent statistics. Assume that all events E_i (i=1,..,n) can be aggregated into a smaller number of sets of events S_1 , …, S_G in such a way that each event exclusively falls in a single set S_g, where g=1,…,G. For our data, this corresponds to the situation that all 417 patent classes can be grouped into one of the 36 more aggregated technological subcategories constructed by HALL et al. (2002), or at an even higher level of aggregation to one of their 6 technological categories. The probability that event E_i in S_g occurs is obtained by summation:

$$P_g = \sum_{i \in S_g} p_i \tag{2}$$

The entropy at the level of sets of events is:

$$H_0 = \sum_{g=1}^{G} P_g \ln \left(\frac{1}{P_g} \right) \tag{3}$$

H_0 is called the "between-group entropy". Within the present context, it would give an indication of the extent to which a state has patents that are evenly distributed over broadly defined technological categories. The entropy decomposition theorem specifies the relationship between the between-group entropy H_0 at the level of sets and the entropy H at the level of events as defined in (1). As shown by THEIL (1972), one obtains:

$$H = H_0 + \sum_{g=1}^{G} P_g H_g \tag{4}$$

The entropy at the level of events is thus equal to the entropy at the level of sets plus a weighted average of within-group entropy levels within the sets. For our purposes, (4) implies that we can consider technological variety at the lowest level of aggregation as the sum of technological variety within classes at a higher level of aggregation and variety between these classes..

As mentioned above, we rely on the technological classification by HALL et al. (2002). Because CASTALDI and LOS (2012) focused on 31 subcategories (leaving out all patents in Hall et al.'s "Miscellaneous" subcategories) in identifying superstar patents, we can only consider patents in 6 categories, 31 subcategories and 296 classes. We measure *unrelated variety* (*UV*) as the entropy of the distribution of patents over 1-digit categories, which tells us how diversified each state is across the 6 broad unrelated technological categories:

$$UV_{it} = \sum_{k=1}^{6} s_{k,it} \ln\left(1/s_{k,it}\right) \tag{5}$$

where $s_{k,it}$ represents the share of patents in technological category k in all patents granted with the first inventor in state i and applied for in year t. Next, we define *semi-related variety* (*SRV*) as the weighted sum of 2-digit entropies in each 1-digit category. The decomposition theorem (4) implies that this is the difference between the entropy measure at the level of 2-digit technological subcategories and *UV* itself:

$$SRV_{it} = \sum_{l=1}^{31} s_{l,it} \ln\left(1/s_{l,it}\right) - \sum_{k=1}^{6} s_{k,it} \ln\left(1/s_{k,it}\right) \quad (6)$$

in which l indexes the technological subcategories. Finally, *related variety* (*RV*) is the diversity of a state's patent portfolio at the most fine-grained classification. We compute it in a similar vein as SRV, but taking the difference between total entropy at the level of narrowly defined 3-digit patent classes and 2-digit technological subcategories:

$$RV_{it} = \sum_{m=1}^{296} s_{m,it} \ln\left(1/s_{m,it}\right) - \sum_{l=1}^{31} s_{l,it} \ln\left(1/s_{l,it}\right) \quad (7)$$

The related-variety and semi-related-variety indicators measure the within group variety components and indicate how diversified a state is within the higher level categories.

We should stress that (semi-)related and unrelated variety are not opposites, but orthogonal in their meaning (FRENKEN et al., 2007). In principle, a state can be characterized by both high related and unrelated variety. These would be states that are diversified into unrelated technological categories while being diversified into many specific classes in each of these categories as well. Any other combination of above-average and below-average levels of *UV*, *SRV* and *RV* is possible as well, at least theoretically, even if empirically related and unrelated variety tend to correlate positively (FRENKEN et al., 2007; QUATRARO, 2010; QUATRARO, 2011; BOSCHMA et al., 2012; HARTOG et al., 2012).

Next to our entropy measures, we also take into account each state's R&D expenditures (*RD*) as their key innovation input variable. R&D expenses give us a measure of the scale of inventive efforts in each state. We collect historical R&D data at the state level from NSF (2012). The figures cover total (company, federal, and other) funds for industrial R&D performance by US state for the years 1963-1998. Until 1995, data are available only for odd years since the R&D survey was administered every other year. We estimate the values for even years using linear interpolation. Next, the figures are expressed in constant 2005 dollars using GDP deflators.

We pool observations across states and years together and model each of the two dependent variables as a function of 1-year lag independent variables, namely our three entropy measures and R&D. The lag is there to account for the fact that inventive output is related to prior efforts, rather than happening simultaneously. These considerations are reflected in our two regression equations:

$$NUMPATENTS_{it} = \alpha^N + \beta_1^N UV_{i,t-1} + \beta_2^N SRV_{it-1} + \beta_3^N RV_{it-1} + \gamma^N RD_{it-1} + \delta^N \mathbf{d} + \varepsilon_{it} \quad (8)$$

$$SHARESUPER_{it} = \alpha^S + \beta_1^S UV_{i,t-1} + \beta_2^S SRV_{it-1} + \beta_3^S RV_{it-1} + \gamma^S RD_{it-1} + \delta^S \mathbf{d} + \upsilon_{it} \quad (9)$$

The vector **d** contains dummies to capture time-invariant state-specific effects and a variable to capture trends over time. Given that R&D data are available until 1998, our sample covers 51 US states for the years 1977-1999. Missing values of the R&D variable (for a number of states these data are not available for periods of varying length) imply that we have a total of 877 observations.

We rely on generalized linear model regression methods to estimate (8) and (9). For (8), we estimate a Negative Binomial model, given that *NUMPATENTS* is a count variable. For (9) we can estimate a linear model. We use tests based on the model deviance (McCullagh and Nelder 1989) to gauge the goodness of fit of the models and to compare the performance of nested models.

4. Results

Before turning to the tests of our hypotheses, it is important to give indications of the empirical importance of the differences we attempt to explain, and to give some ideas about statistical properties of the explanatory variables. Table 1 gives some descriptive statistics, computed over all 877 observations.

The output of patents (*NUMPATENTS*) varies strongly across states and years. In 1990, South Dakota only produced 12 patents, whereas California churned out as many as 15,404 in 1997. The average number of patents by state grew rather steadily from 567 in 1977 to 1169 in 1999. This modest growth in combination with the absence of wild swings implies that most of the variation in *NUMPATENTS* is in the "across states" dimension. In 1977, the top-5 patent producers in that year (California, New York, New Hampshire, Indiana and Pennsylvania) produced as much as 45% of all patents considered. In 1999, the share of the top-5 was also 45%, but the composition of the top-5 changed slightly (California, Texas, New York, Michigan and New Hampshire).

Table 1: List of the variables and descriptive statistics (N=877).

Variable	Description	Min	Max	Mean	Std. Dev.
NUMPATENTS	**Total number of USPTO patents applied in year *t* assigned to inventors located in the state**	**12**	**15,404**	**887.66**	**1402.37**
SHARESUPER	**Share (in %) of superstar patents in total patents for year *t* and state *i*.**	**0.00**	**12.21**	**4.34**	**1.95**
UV	**Entropy at 1-digit level technological categories**	**0.79**	**1.78**	**1.61**	**0.13**
SRV	**Entropy at 2-digit-level subcategories minus entropy at 1 digit level categories**	**0.61**	**1.64**	**1.38**	**0.14**
RV	**Entropy at 3-digit-level classes minus entropy at 2 digit level subcategories**	**0.09**	**1.93**	**1.37**	**0.35**
RD	**Total R&D expenditures (in thousands of 2005 US$)**	**2000**	**41,561,000**	**2,886,000**	**4,821,000**

We also find a lot of variation with respect to the second dependent variable, the share of superstar patents in all patents (*SHARESUPER*). A substantial number of states almost never produce a superstar patent. Alaska, South Dakota, Wyoming and Nevada generated less than one superstar patent per year over the period 1977-1999. At the other end of the spectrum, California managed to generate more than 11,500 superstar patents over this period. On average, California was not the state with the strongest specialization in the production of superstar patents, though. Idaho and Minnesota averaged shares of 7.1% and 6.9%, while we find shares of 6.7%, 6.7% and 6.4% for California, New Mexico and Massachusetts, respectively.[5] At the bottom end, we mainly find states that produced only few patents in general, such as South Dakota (1.9%), Nevada (2.1%) and Arkansas (2.6%).

[5] The maximum *SHARESUPER* of 12.1% in the sample was recorded for New Mexico in 1992. Idaho (which produced a high number of superstar patents in semiconductor technology (see CASTALDI and LOS, 2012) had an even higher *SHARESUPER* (16.4%) for 1992, but this observation could not be included in our sample since R&D data for this state were lacking for 1991-1993.

Unrelated variety (*UV*) remained relatively constant over time, at around 1.60. The maximum entropy for a situation with six technological categories is ln(6) = 1.79, so 1.60 implies that most states had a very diversified patent production at this level of aggregation. In a few states, though, much less variety could be found. Alaska, Nevada and Wyoming are examples of states that did not generate many patents, and it could be expected that their patents could not cover the entire technological range to a substantial extent. The situation is different for Delaware and Idaho, however. These states produced as many as about 300 patents per year on average, but have average *UV* values of 1.30 and 1.39, respectively. Patents in Chemicals as a fraction of all patents over the period 1977-1999 assigned to Delaware amounted to as much as 57% (mainly due to DuPont's activities), while patents in Electrical and Electronic accounted for almost 49% of all patents in Idaho (as a consequence of Micron's inventive capabilities). New York, Connecticut and Minnesota are the states with the highest average over years for *UV*, in the 1.74-1.75 range.

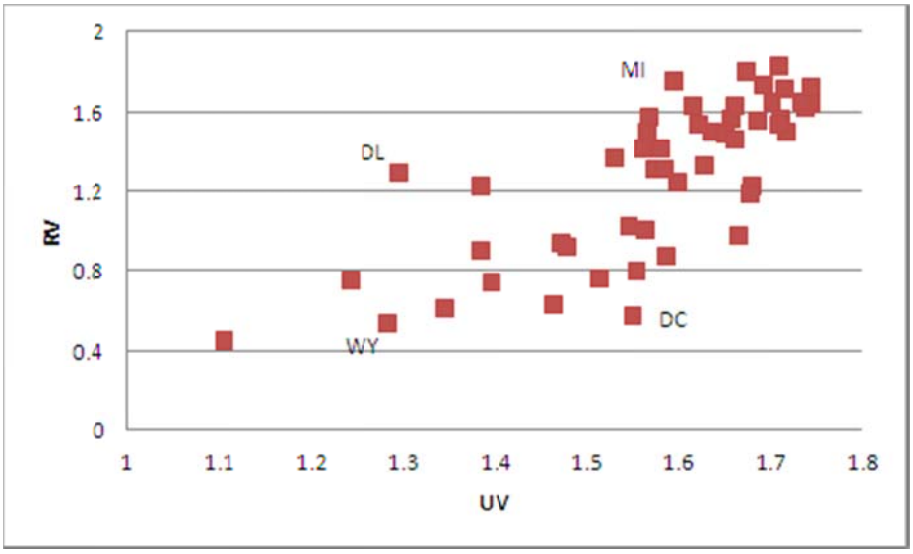
For *SRV* and *RV*, the maximum attainable values (given the numbers of technological subcategories and classes) are ln(31) - ln(6) = 1.64, and ln(296) - ln(31) = 2.16, respectively. As Table 1 reveals, the actual averages over states and years for these variables are 1.38 and 1.37. These averages were again relatively stable, with a slight decline in *SRV* over the last six to seven years of the period under investigation. The top-3 states in terms of average *SRV* were California (1.53), Colorado (1.50) and New York (1.49). New Hampshire is the prime example of a heavy producer of patents with little semi-related variety. With an average *SRV* of 1.29 it belongs to the bottom-15 of states, besides states that do not produce many patents, Delaware and Idaho. Turning to related variety (*RV*), we find a different top-3: Indiana (1.83), Ohio (1.79) and Michigan (1.75). Idaho (0.90), Rhode Island (0.98) and New Jersey (1.00) are examples of states that produce sizable numbers of patents, but with little related variety. These examples strengthen the impression conveyed by the last two columns of Table 1, which show that the coefficient of variation (standard deviation divided by mean) increases with the level of technological detail at which variety is measured.

R&D budgets went up over time. In our data, the average amount of R&D expenditures over states grew from about US$1,750 million in 1977 to about US$3,750 million in 1999 (all amounts converted to constant prices in 2005). The top-5 states in terms of average R&D funds were California (28.9 billion), Michigan (11.0), New York (10.0), New Jersey (8.7) and Massachusetts (6.7). States like Wyoming (0.014 billion), South Dakota (0.015) and North Dakota (0.032) appear at the bottom.

In the previous section, we argued that the entropy decomposition theorem allows us to quantify *UV*, *SRV* and *RV* in a way that allows for complete statistical independence of these variety measures. The framework does not impose such independence, so it might be insightful to see how these variables correlate in the sample. Figure 1 contains observations

for all 51 states. The horizontal axis indicates the average value of *UV* over the entire period (including observations that had to be removed from the regression analysis as a consequence of missing data for *RD*), while the average values for states for *RV* are reflected by the vertical axis. The scatterplot shows that there is a clear positive relation between the two variables in line with previous findings (FRENKEN et al., 2007; QUATRARO, 2010; QUATRARO, 2011; BOSCHMA et al., 2012; HARTOG et al., 2012). An increase of 0.1 in *UV* implies (on average) an increase of 0.22 in *RV*. This hardly changes if only the 30 states with the highest values of *UV* are taken into account (0.21). The explanatory power of a simple model of *RV* with *UV* and a constant intercept as independent variables is not extremely high, though (R^2=0.58).

Figure 1: Related variety vs. unrelated variety



Note: Squares denote state averages for *UV* and *RV* over 1977-1999.

Figure 1 reveals some examples of states with similar average unrelated variety levels, but which had very different levels of related variety. Wyoming and Delaware are examples of such states with very low levels of *UV*, while Washington DC and Michigan show such differences in *RV* at higher levels of *UV*. An example from 1999 can be illustrative. In that year, Iowa had an *UV* of 1.70 and Florida's *UV* amounted to 1.71, which indicates that these states were diversified to the same extent if the six technological categories are considered. Since the maximum attainable *UV* is 1.79, both states can be considered as having a fairly high degree of unrelated variety. Examining the 296 patent classes on which the RV variable is based, we find that Florida had 1999-patents in as many as 217 classes, whereas Iowa's patents were present in only 138 classes. Apparently, Iowa's patents were much more clustered in relatively few classes within the categories than Florida's, which is clearly reflected in the RVs for both states (Florida: 1.72; Iowa: 1.26).

The positive, but far from perfect linear relationship between *UV* and *RV* as depicted in Figure 1 also shows up in Table 2, which gives the pairwise (Pearson) correlations between the variables that enter our regression equations (8) and (9). The table indicates that positive relationships of about equal strength are also found for pairwise comparisons of *UV* and *RV* with *SRV*. Overall, the results indicate that almost all variables are weakly correlated with each other. The correlations for R&D clearly show that R&D efforts explain a large part of variation in total innovative output (*NUMPATENTS*), but have much less of an impact on the share of breakthrough innovations (*SHARESUPER*).

Table 2: Correlation analysis (*N*=877)

	NUMPATENTS	*SHARESUPER*	RD_{t-1}	UV_{t-1}	SRV_{t-1}
SHARESUPER	.286**				
RD_{t-1}	.847**	.251**			
UV_{t-1}	.258**	.238**	.238**		
SRV_{t-1}	.205**	-.015	.271**	.429**	
RV_{t-1}	.461**	.144**	.378**	.571**	.599**

**: Significant at 5%

Table 3 reports the results of maximum likelihood estimates of the regression models (8) and (9). For each equation, we actually estimated three nested models. Model 1 is a baseline model including only the R&D variable and basically capturing the relation between R&D efforts as innovation inputs and patent counts as proxies for innovation outputs. Model 2 refines Model 1 by inserting state dummies and a time trend. Thereby we control for state-specific fixed effects and a possible positive trend in the intensity of innovative activity. Finally, Model 3 is a complete model, in which our entropy-based measures of variety are included. This last model allows us to test the two main hypotheses of this study.

For both equations, the Chi-square tests based on the difference of the models' Deviance indicate that Model 2 significantly improves upon the goodness of fit of Model 1 and Model 3 significantly improves upon Model 2.

Table 3: GLM regression results for the models explaining the total number of patents and the share of breakthrough innovations per state

DV: *NUMPATENTS*	Model 1		Model 2		Model 3	
	b	p-value	b	p-value	b	p-value
RD_{t-1}	0.189	0.000	0.014	0.540	0.018	0.457
state dummies			yes		yes	
time trend			0.045	0.000	0.045	0.000
UV_{t-1}					-0.493	0.330
SRV_{t-1}					-0.281	0.529
RV_{t-1}					0.805	0.022
Deviance	791		44		37	
df	875		824		821	

DV: *SHARESUPER*	Model 1		Model 2		Model 3	
	b	p-value	b	p-value	b	p-value
RD_{t-1}	0.102	0.000	0.078	0.004	0.092	0.001
state dummies			yes		yes	
time trend			0.129	0.000	0.114	0.000
UV_{t-1}					1.563	0.006
SRV_{t-1}					-1.414	0.005
RV_{t-1}					0.475	0.233
Deviance	3125		1178		1158	
df	875		824		821	

State-level inventive output measured by the total number of patents is positively related to R&D efforts in Model 1, as expected. When state dummies and a time trend are included, the significance of R&D vanishes. This is most probably due to the fact that R&D expenditures vary strongly in terms of levels across states and have grown rather steadily over time, for virtually all states. As a result, the state dummies and the time trend already explain the major differences in R&D efforts and since state dummies and time trend are also strongly significantly related to patent performance, the residual effect of R&D is not significant. Model 3 reveals a significant relation between total patents production *NUMPATENTS* and related variety *RV*, while the unrelated and semi-related variety variables *UV* and *SRV* are not significant. This evidence supports the hypothesis that innovation in general benefits from diversification in related technologies.

If we look at the estimates in the lower panel of Table 3, we see that R&D is also strongly related to the shares of superstars in Model 1. The positive relation remains significant also in Model 2 and Model 3. Differences in the production of breakthroughs across states cannot be simply reduced to state-specific effects, like size. The estimates for Model 3 indicate that both *RD* and *UV* help in explaining those differences. On average, states that are more specialized

in breakthroughs are more diversified across unrelated technologies. Our hypothesis that states with higher unrelated variety would outperform states with lower unrelated variety in terms of breakthrough innovation is thus confirmed. We also find semi-related variety to be 'detrimental' for breakthroughs. If we apply our recombination theory this would suggest that, conditional on a given level of unrelated variety, the more specialized the knowledge in selected subcategories within large technological categories, the more likely is recombination across categories. A lot of focused technological knowledge in diverse technology appears to enhance the specialization of states in producing relatively many breakthrough innovations

Table 4: GLM regression results for the models including a spatial variable (R&D of neighboring states).

DV: *NUMPATENTS*		Model 1		Model 2		Model 3
	b	p-value	b	p-value	b	p-value
RD_{t-1}	0.170	0.000	0.017	0.511	0.021	0.455
$RDneighbours_{t-1}$			-0.002	0.904	0.001	0.964
state dummies			yes		yes	
trend			0.042	0.000	0.041	0.000
UV_{t-1}					-0.358	0.522
SRV_{t-1}					-0.280	0.576
RV_{t-1}					0.764	0.065
Deviance	682		44		25	
df	692		640		637	

DV: *SHARESUPER*		Model 1		Model 2		Model 3
	b	p-value	b	p-value	b	p-value
RD_{t-1}	0.098	0.000	0.084	0.005	0.105	0.001
$RDneighbours_{t-1}$			0.015	0.263	0.012	0.379
state dummies			yes		yes	
trend			0.117	0.000	0.099	0.000
UV_{t-1}					2.240	0.000
SRV_{t-1}					-1.292	0.014
RV_{t-1}					0.127	0.774
Deviance	2469		839		814	
df	692		640		637	

Regressions on spatial units of analysis can be subject to spatial dependence effects. To get an idea of the robustness of the results reported in Table 3, we tested whether not only R&D efforts of the state itself but also of neighboring states have played a role. We constructed an adjacency matrix where two states are defined as neighbors if they share a border. We then constructed the variable *RDneighbors*, which equals the R&D efforts of all neighboring states

taken together. The results of the new estimates are reported in Table 4. The number of observations gets reduced to 693, since the missing values in the R&D variables translate into even more missing values for *RDneighbors*. The additional variable turns out to be not significant, while the other estimates do not change qualitatively, except for *RV* becoming marginally insignificant at 5% in the modified version of (8). All in all, the additional estimations reassure us that spatial dependence effects are not relevant at the state level.

5. Concluding remarks

In many recent studies, empirical support has been established for positive relationships between the related variety present in a region and its economic performance. Implicitly, these studies assume that the two variables considered are linked to each other via innovation. Not much work has been done, however, on directly investigating the impact of technological variety on innovation performance. The theory of recombinant innovation provides a framework from which testable hypotheses in this respect can be derived. We argued that breakthrough innovations will most likely depend on technological variety in a way that is different from innovation in general. For producing a breakthrough innovation, recombination of very different types of technological knowledge is needed, while more incremental innovation (along well-defined technological trajectories) would benefit mainly from recombining knowledge about closely related topics.

In this paper, we used patent data from the US Patent and Trademark Office regarding inventions in US states, and used statistical regularities in the numbers of citations that patents receive to distinguish between breakthrough innovations and more regular innovations. Having complete information on the classifications of these patents at three levels of technological aggregation, we used entropy statistics to construct variables reflecting unrelated variety, semi-related variety and related variety. Including these as independent variables in a regression framework, we could test our hypotheses. We found that a high degree of unrelated variety affects the share of breakthrough innovation in a state's total innovation output positively, while semi-related variety has a negative effect. As hypothesized, related variety does not influence breakthrough innovation, but has a clear positive effect on innovation output in general. Our models include control variables, time trends and dummies to capture time-invariant state-specific effects. The results also appeared robust against inclusion of spatial effects.

It goes without saying that further studies are required to probe the validity of our findings regarding the differential effects of related variety and unrelated variety on the types of innovation processes they support. This can be done in at least two ways. First, future studies could replicate our study for regions in different countries. Second, given the limitations of

patent data, one could attempt to test the theoretical framework by using other proxies for innovation, breakthrough innovation and related and unrelated variety. Third, the technological relatedness of regions is a dynamic concept, which changes according to the specific point in time chosen by retrospective research. We consider further investigations in the mechanisms underlying the evolving nature of technological relatedness as among the most interesting and challenging research avenues.

References

AHUJA, G. and LAMPERT, C.M. (2001) Entrepreneurship in the large corporation: a longitudinal study of how established firms create breakthrough inventions, *Strategic Management Journal* 2(6-7), 521–543.

ALMEIDA, P. and KOGUT, B. (1999) Localisation of knowledge and the mobility of engineers in regional networks, *Management Science* 45(7), 905–917.

ANTONELLI, C., KRAFFT, J. and QUATRARO, F. (2010) Recombinant knowledge and growth: The case of ICTs, *Structural Change and Economic Dynamics* 21, 50–69.

ANTONIETTI, R. and CAINELLI, G. (2011) The role of spatial agglomeration in a structural model of innovation, productivity and export, *Annals of Regional Science* 46, 577–600.

ARTHUR, W.B. (2007) The structure of invention, *Research Policy* 36(2), 274-287.

BASALLA, G. (1988) *The Evolution of Technology*, Cambridge University Press.

BECKER, M.C., KNUDSEN, T. and SWEDBERG, R. (2012) Schumpeter's *Theory of Economic Development*: 100 years of development, *Journal of Evolutionary Economics* 22(5), 917-933.

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

BOSCHMA, R.A. (2005) Proximity and innovation. A critical assessment, *Regional Studies* 39(1), 61–74.

BOSCHMA, R.A. and FRENKEN, K. (2006) Why is economic geography not an evolutionary science? Towards an evolutionary economic geography, *Journal of Economic Geography* 6(3), 273–302.

BOSCHMA, R.A. and IAMMARINO, S. (2009) Related variety, trade linkages and regional growth, *Economic Geography* 85(3), 289–311.

BOSCHMA, R.A., MINONDO, A. and NAVARRO, M. (2012) Related variety and regional growth in Spain, *Papers in Regional Science* 91(2), 241–256.

BRACHERT, M., KUBIS, A. and TITZE, M. (2011) Related variety, unrelated variety and regional functions: Identifying sources of regional employment growth in Germany from 2003 to 2008. IWH-Diskussionspapiere, No. 2011,15

BRESCHI, S. and LISSONI, F. (2009) Mobility of skilled workers and co-invention networks: an anatomy of localized knowledge flows. *Journal of Economic Geography*, 9(4), 439–468.

CASTALDI, C. and LOS, B. (2012) *Are New 'Silicon Valleys' Emerging? The Changing Distribution of Superstar Patents across US States*, DRUID Summer Conference 2012.

DESROCHERS, P. and LEPPÄLÄ, S. (2011) Opening up the 'Jacobs Spillovers' black box: local diversity, creativity and the processes underlying new combinations, *Journal of Economic Geography* 11(5), 843–863.

DOSI, G. (1982) Technological paradigms and technological trajectories, *Research Policy* 11, 147-162.

EJERMO, O. (2005) Technological diversity and Jacobs' externality hypothesis revisited, *Growth and Change* 36(2), 167–195.

ESSLETZBICHLER, J. (2007) Diversity, stability and regional growth in the United States 1975–2002. In: Frenken K (ed.), *Applied Evolutionary Economics and Economic Geography*. Cheltenham: Edward Edgar, pp. 203-229.

FLEMING, L. (2001) Recombinant uncertainty in technological space, *Management Science* 47(1), 117-132.

FRENKEN, K. (2007). Entropy statistics and information theory, in H. Hanusch and A. Pyka (eds.), *The Elgar Companion to Neo-Schumpeterian Economics*, Cheltenham, UK and Northampton MA: Edward Elgar, pp. 544-555.

FRENKEN, K. and BOSCHMA, R.A. (2007) A theoretical framework for evolutionary economic geography: industrial dynamics and urban growth as a branching process, *Journal of Economic Geography* 7(5), 635–649.

FRENKEN, K., IZQUIERDO, L. and ZEPPINI, P. (2012) Branching innovation, recombinant innovation and endogenous technological transitions, *Environmental Innovation and Societal Transitions* 4, 25-35.

FRENKEN, K., VAN OORT, F.G. and VERBURG, T. (2007) Related variety, unrelated variety and regional economic growth. *Regional Studies*, 41(5), 685–697.

GLAESER, E., KALLAL, H.D., SCHEINKMAN, J.A. and SHLEIFER, A. (1992) Growth in cities, *Journal of Political Economy* 100(6), 1126-52.

GRUPP, H. (1990) The concept of entropy in scientometrics and innovation research. An indicator for institutional involvement in scientific and technological developments, *Scientometrics* 18, 219-239

HALL, B.H., JAFFE, A.B. and TRAJTENBERG, M. (2002) The NBER patent-citations data file: lessons, insights, and methodological tools, in: Jaffe, A.B. and M. Trajtenberg, *Patents, Citations & Innovations* (Cambridge MA: MIT Press), 403-459.

HARTOG, M., BOSCHMA, R. and 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.

HIDALGO, C.A., KLINGER, B., BARABASI, A.-L. and HAUSMANN, R. (2007) The product space and its consequences for economic growth, *Science* 317, 482–487.

IAMMARINO, S. (2005) An evolutionary integrated view of regional systems of innovation. Concepts, measures and historical perspectives, *European Planning Studies* 13(4), 497–519.

JACOBS, J. (1969) *The Economy of Cities*. New York: Vintage Books.

MAMELI, F., IAMMARINO, S. and BOSCHMA, R. (2012) Regional variety and employment growth in Italian labour market areas: services versus manufacturing industries, *Papers in Evolutionary Economic Geography* 12.03, Utrecht University

MCCULLAGH, P. and NELDER, J.A. (1989) *Generalized Linear Models*, London: Chapman and Hall.

NEFFKE, F., HENNING, M., BOSCHMA, R. (2011), How do regions diversify over time? Industry relatedness and the development of new growth paths in regions, *Economic Geography* 87 (3), 237–265.

NOOTEBOOM, B. (2000) *Learning and Innovation in Organizations and Economies*. Oxford: Oxford University Press.

NSF (2012), Industrial Research and Development Information System, Historical data, http://www.nsf.gov/statistics/iris/excel-files/historical_tables/h-21.xls

QUATRARO, F. (2010) Knowledge coherence, variety and productivity growth: Manufacturing evidence from Italian regions, *Research Policy* 39, 1289–1302.

QUATRARO, F. (2011) Knowledge structure and regional economic growth: The French Case, in: G.D. Libecap and S. Hoskinson (eds.), *Entrepreneurship and Global Competitiveness in Regional Economies: Determinants and Policy Implications*, Emerald Group Publishing Limited, pp. 185–217.

RIGBY, D.L. and ESSLETZBICHLER, J. (2006) Technological variety, technological change and a geography of production techniques, *Journal of Economic Geography* 6, 45–70.

SAVIOTTI, P.P. and FRENKEN, K. (2008) Trade variety and economic development of countries, *Journal of Evolutionary Economics* 18(2), 201–218.

SILVERBERG, G. and VERSPAGEN, B. (2007) The size distribution of innovations revisited: An application of extreme value statistics to citation and value measures of patent significance, *Journal of Econometrics* 139, 318–339.

SINGH, J. and FLEMING, L. (2010) Lone inventors as sources of technological breakthroughs: myth or reality?, *Management Science* 56 , 41-56.

SMITH, K. (2005) Measuring innovation, in Fagerberg, J., Mowery, D.C. and Nelson, R.R. (eds.), *The Oxford Handbook of Innovation*, Oxford University Press.

STUART, T. and SORENSON, O. (2003) The geography of opportunity: spatial heterogeneity in founding rates and the performance of biotechnology firms, *Research Policy* 32(2), 229–253.

TAVASSOLI, M.H. and CARBONARA, N. (2012) The role of knowledge on the innovative capability of Swedish regions. *Paper presented at the ERSA conference*, Bratislava, August, 2012.

THEIL, H. (1972) *Statistical Decomposition Analysis*, Amsterdam: North-Holland.

TRAJTENBERG, M. (1990) A penny for your quotes: patent citations and the value of innovations, *RAND Journal of Economics* 20, 172-187.

VAN DEN BERGH, J. (2008) Optimal diversity: Increasing returns versus recombinant innovation, *Journal of Economic Behavior and Organization* 68 (3-4), 565-580.

WEITZMAN, M.L. (1998) Recombinant growth, *The Quarterly Journal of Economics* 113(2), 331-360.

VAN ZEEBROECK, N. (2011) The puzzle of patent value indicators, *Economics of Innovation and New Technology* 20(1), 33-62.