Related variety, unrelated variety and technological breakthroughs : an analysis of US state-level patenting

Regional Studies
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/cres20

Related Variety, Unrelated Variety and Technological Breakthroughs: An analysis of US State-Level Patenting

Carolina Castaldi[a], Koen Frenken[b] & Bart Los[c]

[a] School of Innovation Sciences, Eindhoven University of Technology, PO Box 513, Eindhoven NL-5600MB, the Netherlands.

[b] Innovation Studies, Copernicus Institute of Sustainable Development, Utrecht University, PO Box 80115, Utrecht NL-3508TC, the Netherlands. Email:

[c] Faculty of Economics and Business, University of Groningen, PO Box 800, Groningen NL-9700 AV, the Netherlands. Email:

Published online: 20 Aug 2014.

To cite this article: Carolina Castaldi, Koen Frenken & Bart Los (2015) Related Variety, Unrelated Variety and Technological Breakthroughs: An analysis of US State-Level Patenting, Regional Studies, 49:5, 767-781, DOI: 10.1080/00343404.2014.940305

To link to this article: http://dx.doi.org/10.1080/00343404.2014.940305

Regional Studies, 2015
Vol. 49, No. 5, 767–781, http://dx.doi.org/10.1080/00343404.2014.940305

Related Variety, Unrelated Variety and Technological Breakthroughs: An analysis of US State-Level Patenting

CAROLINA CASTALDI*, KOEN FRENKEN† and BART LOS‡
*School of Innovation Sciences, Eindhoven University of Technology, PO Box 513, Eindhoven NL-5600MB, the Netherlands. Email: c.castaldi@tue.nl
†Innovation Studies, Copernicus Institute of Sustainable Development, Utrecht University, PO Box 80115, Utrecht NL-3508TC, the Netherlands. Email: k.frenken@uu.nl
‡Faculty of Economics and Business, University of Groningen, PO Box 800, Groningen NL-9700 AV, the Netherlands. Email: b.los@rug.nl

(Received January 2013; in revised form June 2014)

CASTALDI C., FRENKEN K. and LOS B. Related variety, unrelated variety and technological breakthroughs: an analysis of US state-level patenting, *Regional Studies*. This paper investigates how variety affects the innovation output of a region. Borrowing arguments from theories of recombinant innovation, it is expected that related variety will enhance innovation as related technologies are more easily recombined into a new technology. However, it is also expected 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 1977–99 and associated citation data, evidence is found for both hypotheses. This study thus sheds a new and critical light on the related variety hypothesis in economic geography.

Recombinant innovation Regional innovation Superstar patents Technological variety Evolutionary economic geography

CASTALDI C., FRENKEN K. and LOS B. 相关多样性、非相关多样性与技术突破：美国州层级的专利授予分析，*区域研究*。本文探讨多样性如何影响一个区域的创意产出。本文借用重组式创新理论的主张，预期相关多样性将会增进创新，因为相关技术更容易重组成为新的技术。但本文同时预期，非相关多样性能够增进技术突破，因为突破性的创新经常源自于连结过去不相关的技术，并开啓崭新的功能性与应用。本研究运用美国各州在1977年至1999年之间的专利数据和相关的引用数据，同时发现支持上述两项假说的证据。本研究因此对经济地理学中的相关多样性假说，提供了崭新且具批判性的洞见。

重组式创新 区域创新 明星专利 技术多样性 演化经济地理学

CASTALDI C., FRENKEN K. et LOS B. La variété reliée, la variété non reliée et les percées technologiques: une analyse de l'obtention de brevets au niveau des états aux É-U, *Regional Studies*. Cet article examine comment la variété influe sur l'innovation d'une région. S'appuyant sur les théories de l'innovation recombinante, on s'attend à ce que la variété reliée améliore l'innovation parce que l'on peut recombiner plus facilement les technologies reliées en nouvelle technologie. Cependant, on s'attend aussi à ce que la variété non reliée améliore les percées technologiques, étant donné que l'innovation radicale provient souvent du raccordement des technologies jusqu'alors sans rapport, ce qui offre des fonctionnalités et des applications tout nouvelles. À partir des données sur les brevets pour les états aux É-U pendant la période de 1977 à 1999 et des données de citation y associées, on a trouvé des preuves qui corroborent les deux hypothèses. Cette étude jette une lumière nouvelle et critique sur l'hypothèse de la variété reliée dans la géographie économique.

Innovation recombinante Innovation régionale Brevets vedettes Variété technologique Géographie économique évolutionniste

CASTALDI C., FRENKEN K. und LOS B. Verwandte Varietät, nichtverwandte Varietät und technologische Durchbrüche: eine Analyse der Patente auf US-Bundesstaatsebene, *Regional Studies*. In diesem Beitrag wird untersucht, wie sich Varietät auf die Innovationsleistung einer Region auswirkt. Unter Anlehnung an die Argumente der Theorien der rekombinanten Innovation gehen wir davon aus, dass verwandte Varietät die Innovation verbessert, da sich verwandte Technologien einfacher zu einer neuen Technologie kombinieren lassen. Allerdings gehen wir auch davon aus, dass nichtverwandte Varietät technologische

Durchbrüche verbessert, da radikale Innovation oft auf einer Kombination bisher nichtverwandter Technologien beruht, die völlig neue Funktionalitäten und Anwendungen ermöglicht. Anhand von Patentdaten für US-Bundesstaaten im Zeitraum von 1977 bis 1999 sowie mithilfe der zugehörigen Zitatdaten werden Belege für beide Hypothesen gefunden. Diese Studie lässt somit die Hypothese der verwandten Varietät in der Wirtschaftsgeografie in einem neuen und kritischen Licht erscheinen.

Rekombinante Innovation Regionale Innovation Superstar-Patente Technologische Varietät Evolutionäre Wirtschaftsgeografie

CASTALDI C., FRENKEN K. y LOS B. Variedad relacionada, variedad no relacionada y avances tecnológicos: un análisis de las patentes estatales en los Estados federales de EE.UU., *Regional Studies*. En este artículo investigamos qué efecto tiene la variedad en la capacidad innovadora de una región. Tomando prestados argumentos de teorías de la innovación recombinante, se prevé que la variedad relacionada aumente la innovación puesto que las tecnologías relacionadas se pueden volver a combinar más fácilmente en una nueva tecnología. Sin embargo, también se supone que con la variedad no relacionada aumenten los avances tecnológicos dado que la innovación radical muchas veces surge de combinar tecnologías no relacionadas previamente, descubriendo toda una serie de nuevas funcionalidades y aplicaciones. A partir de datos de patentes estatales de Estados Unidos durante el periodo de 1977 a 1999 y de datos de citación pertinentes, observamos evidencia de ambas hipótesis. Por consiguiente, este estudio aporta un enfoque nuevo y crítico a la hipótesis de la variedad relacionada en la geografía económica.

Innovación recombinante Innovación regional Patentes superestrella Variedad tecnológica Geografía económica evolutiva

JEL classifications: O31, R11

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. Both NIGHTINGALE (1998) and NOOTEBOOM (2000) stress that decision-makers in firms have limited cognitive capabilities, limiting their abilities to identify potentially fruitful combinations of pieces of knowledge that seem unrelated to their existing knowledge bases and/or to each other. 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, 2011; ANTONIETTI and CAINELLI, 2011; BRACHERT *et al.*, 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, particularly employment growth. In their analysis of the impact of related variety, 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 (TAVASSOLI and CARBONARA, 2014).[1] The present paper aims to develop further the notion of related variety and its effect on innovation. It does so within a theoretical framework that explicitly distinguishes between related and unrelated variety and predicts differential effects of the two types of variety on innovation processes. The authors take issue with the notion that related variety supports all kinds of innovation. Instead, it is argued that related variety is supportive of the bulk of innovations that incrementally build 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.

This paper's new framework is not incompatible with the original related variety framework by FRENKEN *et al.* (2007), since related variety is still expected to support innovation in general. Hence, in so far innovations lead to employment growth, the original related variety hypothesis still holds. Additionally, it is also expected that unrelated variety supports breakthrough innovations. Potentially, breakthrough innovations may have much more impact on employment growth than innovations more generally, since whole new industries can emerge out of breakthrough innovations in the long run (SAVIOTTI and FRENKEN 2008). Nevertheless, the present additional hypothesis does not necessarily contradict previous findings that unrelated variety does not support employment growth, since earlier studies only analysed the short-term effect of variety on employment. What is more, the employment effects of technological breakthroughs need not be found in the region of origin, since the successful commercialization of a breakthrough technology may well take place in regions other than the region from which it originated (BOSCHMA 1997; MURMANN 2003).

Within this new theoretical framework, two hypotheses are tested. The first 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 states that unrelated variety of the regional knowledge base supports the rare breakthrough innovations, while related variety does not have such an effect.

A criterion based on the numbers of citations to a patent is used 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 (USPTO) 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 US states. To construct variables regarding various types of variety of the regional knowledge base, technological classification schemes at different levels of aggregation were used, as designed by the USPTO. The actual construction of related- and unrelated variety variables is rooted in entropy statistics (FRENKEN *et al.*, 2007).

The 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 a spatially lagged research and development (R&D) variable, that is, the sum of R&D investments in neighbouring states.

The paper is structured as follows. The second section gives a brief overview of the theoretical concepts on the interplay of existing pieces of knowledge in recombinant innovation processes. The methods are introduced in the third section, which includes a discussion of the procedure adopted to distinguish between incremental innovations and breakthrough innovations. The fourth section shows how the numbers of produced breakthrough innovations vary across states and provides indications of differences in the variety of their knowledge bases, before testing the hypotheses using econometric estimation techniques. The fifth section concludes.

VARIETY, RECOMBINATION AND INNOVATION

Technological innovation is commonly understood to be a cumulative process in which most new artefacts are being invented by recombining 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, smart phones combine technologies related to batteries, chips, antennas, audio, video, display and the Internet. In this context, Schumpeter famously spoke of innovation as the bringing about of new combinations ('*Neue Kombinationen*'), an idea that continues to inspire evolutionary theorizing 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 localized 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 artefact 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).

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. The following two hypotheses will guide the remainder of this study:

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 inventions.

RESEARCH DESIGN

The hypotheses are tested 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). In a well-known early contribution to the literature, ACS *et al.* (1992) found evidence that patent counts are a noisy but useful indicator of innovative activity at the state level, by comparing patent counts to numbers of innovations identified in professional and trade journals.[2] Given that money was invested in advertising these innovations, it is likely that these corresponded to patents with a perceived high value. A more debated issue is how to quantify success in producing breakthrough innovations in a systematic way. Can this also be attained using patent statistics? 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 the 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 (HALL *et al.*, 2002). This study uses a refined methodology proposed by CASTALDI and LOS (2012) to identify what they term 'superstar patents'. The basic idea behind this methodology is to derive endogenously 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 the USPTO as either superstars or not.[3] This estimation relies on a modified version of the estimation routine in SILVERBERG and VERSPAGEN (2007), based upon the so-called Hill estimator (for more details, see Appendix A). Additionally, it was ensured that only patents with the same application year and belonging to an identical

technological subcategory were compared. USPTO patents have been classified by HALL *et al.* (2002) in six 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 National Bureau of Economic Research (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 present purposes, the aim is to count patents and superstar patents across regions. US patents included in the NBER database can be assigned to the US state of the first inventor. The state will be the definition of a region in this study.[4] For each state and each year from 1976 to 1999, there are the number of total granted patents applied for in that year at the USPTO by inventors in that state and also estimates of how many of the total patents are superstar patents.[5] As the hypotheses relate to explaining regional innovative output, this paper works with two dependent variables for each state i:

- The total number of granted patents with application year t, as a proxy for the general innovation performance of a state ($NUMPATENTS_{it}$).
- 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 ($SHARESUPER_{it}$).

It was chosen to consider shares of superstars rather than absolute numbers, since shares indicate 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 innovations in very specific places, while the production of regular innovations happens in many more places. Their descriptive results regarding this issue are in line with similar ones by EJERMO (2009) and indicate already that explaining regional performance in terms of breakthrough innovation requires different hypotheses than explaining regional innovative performance in more general terms.

The paper now turns to the explanatory variables. The key independent variables in the model will be measures of regional variety in innovative activity. Again, patent data are used, as patents indicate something about the technological fields in which states contribute innovations. In line with previous work, variety is measured 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 be the probability of event E_i occurring, with $i = 1, \ldots, n$. The entropy level H is given by:

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

with:

$$p_i \ln\left(\frac{1}{p_i}\right) = 0 \quad \text{if} \quad p_i = 0$$

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 were to have all its patents in a single patent class. If a patent were to 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 the 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 (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 one 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 ($i = 1, \ldots, n$) can be aggregated into a smaller number of sets of events $S_1, \ldots, S_G$ in such a way that each event exclusively falls in a single set S_g, where $g = 1, \ldots, G$. For the 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 six 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 present purposes, (4) implies that one 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.[6]

As mentioned above, the technological classification by HALL *et al.* (2002) is relied upon. Because CASTALDI and LOS (2012) focused on 31 subcategories (leaving out all patents in HALL *et al.*'s (2002) 'Miscellaneous' subcategories) in identifying superstar patents, one can only consider patents in six categories, 31 subcategories and 296 classes. Unrelated variety (UV) is measured as the entropy of the distribution of patents over one-digit categories, which states how diversified each state is across the six broad unrelated technological categories:

$$\mathrm{UV}_{it} = \sum_{k=1}^{6} s_{k,it} \ln\left(\frac{1}{s_{k,it}}\right) \tag{5}$$

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, *semi-related variety* (SRV) is defined as the weighted sum of two-digit entropies in each one-digit category. The decomposition theorem (4) implies that this is the difference between the entropy measure at the level of two-digit technological subcategories and UV itself:

$$\mathrm{SRV}_{it} = \sum_{l=1}^{31} s_{l,it} \ln\left(\frac{1}{s_{l,it}}\right) - \sum_{k=1}^{6} s_{k,it} \ln\left(\frac{1}{s_{k,it}}\right) \tag{6}$$

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. It is computed in a similar vein as SRV, but taking the difference between total entropy at the level of narrowly defined three-digit patent classes and two-digit technological subcategories:

$$\mathrm{RV}_{it} = \sum_{m=1}^{296} s_{m,it} \ln\left(\frac{1}{s_{m,it}}\right) - \sum_{l=1}^{31} s_{l,it} \ln\left(\frac{1}{s_{l,it}}\right) \tag{7}$$

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.

It should be stressed 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- 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, 2011; BOSCHMA *et al.*, 2012; HARTOG *et al.*, 2012).

Next to the entropy measures, one also takes into account each state's R&D expenditures (RD) as their key innovation input variable. R&D expenses give a measure of the scale of inventive efforts in each state. Historical R&D data are collected 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 1963–98. Until 1995, data are available only for odd years since the R&D survey was administered every other year. The values for even years are estimated using linear interpolation. Next, the figures are expressed in constant 2005 US dollars using gross domestic product (GDP) deflators.

The observations are pooled across states and years and each of the two dependent variables is modelled as a function of one-year lag independent variables, namely the 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 the two regression equations:

$$\begin{aligned}\mathrm{NUMPATENTS}_{it} = {} & \alpha^N + \beta_1^N \mathrm{UV}_{i,t-1} \\ & + \beta_2^N \mathrm{SRV}_{it-1} + \beta_3^N \mathrm{RV}_{it-1} \\ & + \gamma^N \mathrm{RD}_{it-1} + \delta^N \mathbf{d} + \varepsilon_{it}\end{aligned} \tag{8}$$

$$\begin{aligned}\mathrm{SHARESUPER}_{it} = {} & \alpha^S + \beta_1^S \mathrm{UV}_{i,t-1} \\ & + \beta_2^S \mathrm{SRV}_{it-1} + \beta_3^S \mathrm{RV}_{it-1} \\ & + \gamma^S \mathrm{RD}_{it-1} + \delta^S \mathbf{d} + \upsilon_{it}\end{aligned} \tag{9}$$

The vector $\mathbf{d}$ 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, the sample covers 51 US states for the years 1977–99. Missing values of the R&D variable (for a number of states these data are not available for periods of varying length) imply that there is a total of 877 observations.

The method relies on generalized linear model (GLM) regression methods to estimate (8) and (9). For

(8), a negative binomial model is estimated, given that NUMPATENTS is a count variable. For (9) a linear model can be estimated. Tests based on the model deviance (McCullagh and Nelder, 1989) are used to gauge the goodness of fit of the models and to compare the performance of nested models. Standardized coefficients are reported; for the case of the negative binomial model only the independent variables are standardized since the dependent one is a count.

RESULTS

Before turning to the tests of the hypotheses, it is important to give indications of the empirical importance of the differences being explained, 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-five 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 five was also 45%, but the composition of the top five changed slightly (California, Texas, New York, Michigan and New Hampshire).

A lot of variation is also found 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 1977–99. 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, however. Idaho and Minnesota averaged shares of 7.1% and 6.9%, while there are shares of 6.7%, 6.7% and 6.4% for California, New Mexico and Massachusetts, respectively.[7] At the bottom end are mainly found states that produced only a few patents in general, such as South Dakota (1.9%), Nevada (2.1%) and Arkansas (2.6%).

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 1977–99 assigned to Delaware amounted for 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 three 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 RV, a different top three is found: Indiana (1.83), Ohio (1.79) and

Table 1. *Variables and descriptive statistics* (N = *877*)

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

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 the data, the average amount of R&D expenditures over states grew from about US$1.75 billion in 1977 to about US$3.75 billion in 1999 (all amounts converted to constant prices in 2005). The top five states in terms of average R&D funds were California (US$28.9 billion), Michigan (US$11.0 billion), New York (US$10.0 billion), New Jersey (US$8.7 billion) and Massachusetts (US$6.7 billion). States like Wyoming (US$0.014 billion), South Dakota (US$0.015 billion) and North Dakota (US$0.032 billion) appear at the bottom.

The previous section argued that the entropy decomposition theorem allows one to quantify UV, SRV and RV in a way that allows for complete statistical independence of these variety measures. Empirically, however, the entropy measures may still be correlated. Fig. 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, 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 ($R^2 = 0.58$).

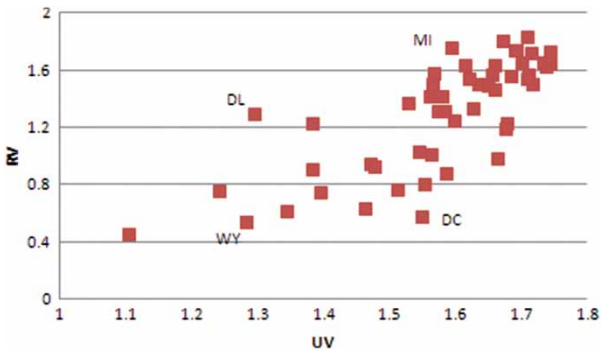

Fig. 1. *Related variety (RV) versus unrelated variety (UV)*
Note: Squares denote state averages for UV and RV over 1977–99

Fig. 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 is 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, it is found that Florida had 1,999 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 Fig. 1, also shows up in Table 2, which gives the pairwise (Pearson) correlations between the variables that enter the regression equations (8) and (9). Table 2 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 3 reports the results of maximum likelihood estimates of the regression models (8) and (9). For each equation, three nested models are actually estimated. 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 one controls 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 the entropy-based measures of variety are included. This last model allows one 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.

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,

Table 2. Correlation analysis (N = 877)

	NUMPATENTS	SHARESUPER	RD_{t-1}	UV_{t-1}	SRV_{t-1}
SHARESUPER	0.286**				
RD_{t-1}	0.847**	0.251**			
UV_{t-1}	0.258**	0.238**	0.238**		
SRV_{t-1}	0.205**	−0.015	0.271**	0.429**	
RV_{t-1}	0.461**	0.144**	0.378**	0.571**	0.599**

Note: **Significant at 5%.

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.[8] 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 first hypothesis that innovation in general benefits from diversification in related technologies.

If one looks at the estimates in the lower panel of Table 3, it can be seen that R&D is also strongly related to the shares of superstars in Model 1. The positive relation remains significant also in Models 2 and 3. Differences in the production of breakthroughs across states cannot be simply reduced to state-specific effects, such as 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. The second hypothesis that states with higher unrelated variety would outperform states with lower unrelated variety in terms of breakthrough innovation is thus confirmed. Semi-related variety is also found to be 'detrimental' for breakthroughs. If the recombination theory is applied, 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. On the other hand, the semi-related variety measure is a measure that was included because of the properties of the data classification. Notice that the key results about related and unrelated variety remain valid even when leaving aside the semi-related variety measure in the model estimations (Model 4).

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, it was tested whether not only R&D efforts of the state itself but also of neighbouring states have played a role. An adjacency matrix was constructed where two states are defined as neighbours if they share a border. The

Table 3. Generalized linear model (GLM) regression results for the models explaining the total number of patents and the share of breakthrough innovations per state (standardized estimates)

	Model 1		Model 2		Model 3		Model 4	
	b	*p*-value	*b*	*p*-value	*b*	*p*-value	*b*	*p*-value
Dependent variable: NUMPATENTS								
RD_{t-1}	0.910	0.000	0.068	0.540	0.087	0.457	0.093	0.425
State dummies			Yes		Yes		Yes	
Time trend			0.301	0.000	0.298	0.000	0.303	0.000
UV_{t-1}					−0.084	0.330	−0.086	0.324
SRV_{t-1}					−0.046	0.529		
RV_{t-1}					0.325	0.022	0.322	0.023
Deviance	791		44		37		37	
d.f.	875		824		821		822	
Dependent variable: SHARESUPER								
RD_{t-1}	0.216	0.000	0.167	0.004	0.197	0.001	0.210	0.000
State dummies			Yes		Yes			
Time trend			0.378	0.000	0.334	0.000	0.345	0.000
UV_{t-1}					0.118	0.006	0.117	0.007
SRV_{t-1}					−0.103	0.005		
RV_{t-1}					0.085	0.233	0.078	0.275
Deviance	611		230		226		228	
d.f.	875		824		821		822	

Table 4. Generalized linear model (GLM) regression results for the models including a spatial variable (R&D of neighbouring states). Coefficient estimates are standardized

	Model 1		Model 2		Model 3		Model 4	
	b	*p*-value	*b*	*p*-value	*b*	*p*-value	*b*	*p*-value
Dependent variable: NUMPATENTS								
RD_{t-1}	0.820	0.000	0.084	0.511	0.101	0.455	0.109	0.421
$RDneighbours_{t-1}$			−0.014	0.904	0.005	0.964	0.013	0.914
State dummies			Yes		Yes		Yes	
Trend			0.281	0.000	0.272	0.000	0.273	0.000
UV_{t-1}					−0.061	0.522	−0.059	0.537
SRV_{t-1}					−0.046	0.576		
RV_{t-1}					0.309	0.065	0.311	0.063
Deviance	682		44		25		25	
d.f.	692		640		637		638	
Dependent variable: SHARESUPER								
RD_{t-1}	0.211	0.000	0.180	0.005	0.223	0.001	0.238	0.001
$RDneighbours_{t-1}$			0.061	0.263	0.048	0.379	0.064	0.379
State dummies			Yes		Yes		Yes	
Trend			0.342	0.000	0.290	0.000	0.293	0.000
UV_{t-1}					0.169	0.000	0.173	0.000
SRV_{t-1}					−0.095	0.014		
RV_{t-1}					0.023	0.774	0.027	0.736
Deviance	482		164		159		161	
d.f.	692		640		637		638	

variable *RDneighbors*, which equals the R&D efforts of all neighbouring states taken together, was then constructed. 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 are reassuring that spatial dependence effects are not relevant at the state level.

DISCUSSION

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. It was argued that breakthrough innovations will most likely depend on technological variety in a way that is different from innovation in general. To produce 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.

This paper used patent data from the USPTO 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, entropy statistics were used to construct variables reflecting unrelated variety, semi-related variety and related variety. By including these as independent variables in a regression framework, the hypotheses could be tested. It was 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. The 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.

A key conclusion from this study holds that the alleged opposition between related and unrelated variety can be misleading, since both types of variety can lead to innovation. Related variety would raise the likelihood of innovations in general, while 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 that the relatedness structure among technologies is evolving, albeit slowly, in a way that is driven by radical innovation that renders previously unrelated technologies to become related (Fig. 2).

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.

A dynamic view on related and unrelated variety would suggest a further research agenda on the topic. In particular, one would be interested to understand at what pace technological relatedness is indeed changing. Furthermore, one can investigate if recombinant innovations across unrelated technologies are indeed driving the fundamental changes in relatedness, and whether firms and regions pioneering such recombinant innovations also thrive economically in the long run. Indeed, further investigations in the mechanisms underlying the evolving nature of technological relatedness are considered to be among the most interesting and challenging research avenues for the future.

The new framework also has potential implications for regional policy initiatives. In particular, the role of variety in regional development is linked to the smart specialization strategy framework pursued by regions supported by the European Commission (Foray *et al.*, 2009; McCann and Ortega-Argiles, 2013). One important part of the smart specialization concept holds that regions should build on related variety to support regional development in the long run (Foray, 2014). By combining knowledge and competences from related sectors or technologies, new activities can emerge in a continuous process of related diversification. Clearly, this is in line with past empirical research on the role of related sectoral variety on employment growth as well as the present study showing the role of related technological variety on patenting. However, this study also suggests that regions should also aim to exploit possible connections between sectors and technologies that are (currently) unrelated in attempt to find innovations that would make them more related. A full discussion of policy instruments that can be helpful to exploit unrelated variety is, however, beyond the scope of the current paper.

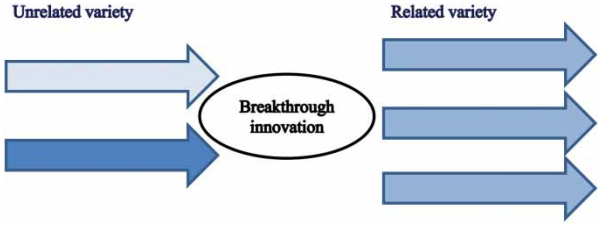

Fig. 2. Breakthrough innovation turning unrelated variety (UV) into related variety (RV)

It goes without saying that further studies are required to probe the validity of the 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 four ways. First, future studies could replicate this 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 links between variety and different types of innovation could be analysed at a lower level of geographical aggregation. As long as R&D data at the level of metropolitan statistical areas (MSAs) are not available, one would have to resort to alternative approaches that do not use knowledge production functions. Fourth, this type of research could be done for companies rather than for regions. If innovation is mainly seen as a firm-specific process in which externalities among regional clusters play a smaller role, the distinction between unrelated variety and related variety could be linked to Teece's (1996) notion that different archetypes of companies are better at specific types of innovation than others. His 'multi-product integrated hierarchies', for example, will generally have firm-specific knowledge bases with a higher degree of unrelated variety than his 'high-flex Silicon Valley types'.

Acknowledgements – The authors thank the editors and anonymous referees for help in improving this paper. They also acknowledge the comments received at the Geography of Innovation Conference in Utrecht (January 2014), at the Science and Policy Research Unit (SPRU) research seminar series (March 2014), and at the European Regional Science Association (ERSA) conference in Palermo (August 2013).

APPENDIX A: USE OF HILL ESTIMATORS TO IDENTIFY SUPERSTAR PATENTS (BASED UPON CASTALDI AND LOS, 2012)

Empirical examination of patent data shows that Pareto distributions are superior at matching the observed frequency distributions in the right tail of the distribution of the 'value' of innovations (Scherer *et al.*, 2000), also when the value is measured with the numbers of citations received by patents (Silverberg and Verspagen, 2007).

To illustrate this with two technologies from the sample, Fig. A1 was generated by ordering all patents with application year 1976 assigned to subcategory 12 ('biotech'), according to the numbers of citations they received in the period 1976–2006 and similarly for subcategory 22 ('optics'). The numbers of citations are depicted along the horizontal axis. The numbers of

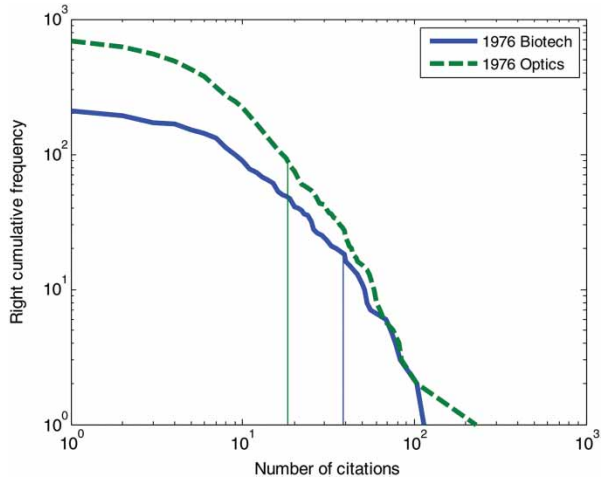

Fig. A1. *Fat tails in numbers of forward citations*
Source: Authors' computations on NBER Patent-Citations Datafile, citations received in 1976-2006. Estimated cutoff points between lognormal distributions and Pareto distributions (vertical lines) obtained by Drees-Kaufman-Lux procedure.

patents with an equal or higher number of citations than the value depicted on the horizontal axis are indicated along the vertical axis. Since both axes have a logarithmic scale, a Pareto distribution appears as a straight, downward sloping line. Exponential distributions (such as the lognormal) show curvature. For both technologies, a mixed distribution depicts indeed the observed frequencies in Fig. A1 more accurately than one type of distribution over the whole range. For less-cited patents, lognormal distributions fit the evermore steeply declining curves better. The rightmost parts of the curves are approximately linear, reflecting Pareto distributions.

Results from extreme value statistics (COLES, 2001) allow the numbers of citations that correspond to the cut-off point to be estimated. Given a comparable set of patents, i.e. applied in the same year and in the same technological field, CASTALDI and LOS (2012) call a patent a superstar patent if it received the cut-off point number of citations or more.

Following SILVERBERG and VERSPAGEN (2007), they estimate the cut-off point by using an estimator for the essential parameter of the Pareto distribution. If the tail follows this distribution $F(x) = 1 - x - \alpha$, a maximum likelihood estimator of the parameter α can be obtained using the Hill estimator (HILL, 1975). Given the rank-order statistics of the sample $X(1) \geq X(2) \geq \ldots \geq X(n)$, the Hill estimator of the inverse of α is obtained as:

$$\bar{\gamma} = \hat{\alpha}^{-1} = \left(\frac{1}{k}\right) \sum_{i=1}^{k} (\ln X_{(i)} - \ln X_{(k+1)})$$

Note that the parameter α reflects the magnitude of the negative slope of the straight line characterizing the Pareto distributions in Pareto-plots like Fig. A1.

The value of the Hill estimator is a function of k, the number of observations included in the tail. The slope parameter of the Pareto distribution is initially estimated for an extremely small subsample, which contains the most highly cited patents only. Next, the subsample is extended with the most cited patent that did not belong to the initial subsample and the Hill estimator is again computed. This procedure is repeated for a successively growing subsample of well-cited patents. As long as these growing subsamples remain drawn from a Pareto distribution indeed, the estimated slopes will remain relatively stable. This changes, however, as soon as patents are added that are well-cited, but belong to the lognormally distributed part of the set of patents. This can be easily visualized with the aid of a so-called Hill plot: the sequence of estimated slopes starts to show a saw-toothed pattern, and each added patent causes a swing in the estimated slopes. The Hill plot can be used to get an idea of the value at which the Hill estimates stabilizes. If the underlying distribution is Paretian, the Hill estimates will stabilize at a certain value. But if the distribution is not overall Paretian, including observations from the central part of the distribution will decrease the validity of the estimator. A method is then needed to estimate the 'optimal' value of the parameter k.

In the computationally convenient procedure adopted by DREES and KAUFMANN (1998), the length of the right tail is first set to one observation. Next, the most likely length is found by examining the fluctuations in the value of the Hill estimator when adding more observations to the tail. Such fluctuations emerge if Hill estimators are applied to distributions that are not Pareto. If a predetermined threshold value is exceeded by the fluctuation, an estimate for k is found. CASTALDI and LOS (2012) use a slightly modified version of this Drees–Kaufmann estimator, proposed by LUX (2001): in this version the stopping rule is modified with a higher threshold so that the tail includes fewer observations from the central part of the distribution.

NOTES

1. Actually, invention is the focus here since issues of successful commercialization are not addressed, but technological attainments are the sole focus. Throughout, the paper uses the terms 'innovation' and 'invention' interchangeably since the theory of recombinant innovation has been framed in terms of innovation rather than invention.
2. Subsequently, patent statistics have often served as a source of indicators for regional inventive activity (e.g. BOTTAZZI and PERI, 2003; FISCHER and VARGA, 2003; EJERMO, 2009). There are good arguments to study smaller geographical units than the state level. California, for example, contains a number of metropolitan areas (the Bay Area with San Francisco and Silicon Valley; the state capital Sacramento; and the Los Angeles agglomeration, among others). Most probably, these agglomerations are

geographically too distant from each other to allow for frequent knowledge spillovers (e.g., THOMPSON, 2006). Many other states, though, like Oregon, Illinois and Massachusetts, are dominated by a single large agglomeration. In such cases, the variety characteristics of the regional knowledge bases at the state level will be very similar to those of the dominant consolidated metropolitan statistical area (CMSA). In view of the fact that this analysis entails the estimation of augmented regional knowledge production functions, analyses for smaller geographical units cannot be done. Data on R&D expenditures, the most important inputs into knowledge production processes, are only available at the state and national levels.

3. An interesting alternative approach was chosen by DAHLIN and BEHRENS (2005). In identifying radical inventions in tennis racket technology, they focused not only on the numbers of citations the associated patents received, but also to what extent citations in these patents to prior art were dissimilar from existing patents. The identified patents were largely successfully confronted with expert opinions afterwards.
4. 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 with smaller spatial units of analysis, state-level analysis also has the advantage of having a substantial number of breakthrough innovations per state.
5. The original NBER Patent Citation database covers all patents granted at the USPTO in 1975–99. 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, the 2002 database is used.
6. As for the Herfindahl index, entropy values are biased for small numbers of patent counts (HALL, 2005).
7. 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; CASTALDI and LOS, 2012) had an even higher SHARESUPER (16.4%) for 1992, but this observation could not be included in the sample since R&D data for this state were lacking for 1991–93.
8. Additional tests (available from the authors upon request) were performed by excluding the state dummies from the regressions. The two key hypotheses remain confirmed and the overall results do not change dramatically.

REFERENCES

ACS Z. J., AUDRETSCH D. B. and FELDMAN M. P. (1992) Real effects of academic research: comment, *American Economic Review* **82**, 363–367.

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* **22**, 521–543. doi:10.1002/smj.176

ALMEIDA P. and KOGUT B. (1999) Localisation of knowledge and the mobility of engineers in regional networks, *Management Science* **45**, 905–917. doi:10.1287/mnsc.45.7.905

ANTONIETTI R. and CAINELLI G. (2011) The role of spatial agglomeration in a structural model of innovation, productivity and export: a firm-level analysis, *Annals of Regional Science* **46**, 577–600. doi:10.1007/s00168-009-0359-7

ARTHUR W. B. (2007) The structure of invention, *Research Policy* **36**, 274–287. doi:10.1016/j.respol.2006.11.005

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

BECKER M. C., KNUDSEN T. and SWEDBERG R. (2012) Schumpeter's theory of economic development: 100 years of development, *Journal of Evolutionary Economics* **22**, 917–933. doi:10.1007/s00191-012-0297-x

BISHOP P. and GRIPAIOS P. (2010) Spatial externalities, relatedness and sector employment growth in Great Britain, *Regional Studies* **44**, 443–454. doi:10.1080/00343400802508810

BOSCHMA R. A. (1997) New industries and windows of locational opportunity. A long-term analysis of Belgium, *Erdkunde* **51**, 12–22. doi:10.3112/erdkunde.1997.01.02

BOSCHMA R. A. and IAMMARINO S. (2009) Related variety, trade linkages and regional growth in Italy, *Economic Geography* **85**, 289–311. doi:10.1111/j.1944-8287.2009.01034.x

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

BOTTAZZI L. and PERI G. (2003) Innovation and spillovers in regions: evidence from European patent data, *European Economic Review* **47**, 687–710. doi:10.1016/S0014-2921(02)00307-0

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. Halle Institute for Economic Research (IWH), Halle.

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**, 439–468. doi:10.1093/jeg/lbp008

CASTALDI C. and LOS B. (2012) Are new 'Silicon Valleys' emerging? The changing distribution of superstar patents across US states. Paper presented at the Danish Research Unit for Industrial Dynamics (DRUID) Summer Conference 2012.

COLES, S. (2001) *An Introduction to Statistical Modeling of Extreme Values*. Springer, London.

DAHLIN K. B. and BEHRENS D. M. (2005) When is an invention really radical?, *Research Policy* **34**, 717–737. doi:10.1016/j.respol.2005.03.009

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**, 843–863. doi:10.1093/jeg/lbq028

DOSI G. (1982) Technological paradigms and technological trajectories, *Research Policy* **11**, 147–162. doi:10.1016/0048-7333(82)90016-6

Drees H. and Kaufmann E. (1998) Selecting the optimal sample fraction in univariate extreme value estimation, *Stochastic Processes and their Applications* **75**, 149–172. doi:10.1016/S0304-4149(98)00017-9

Ejermo O. (2005) Technological diversity and Jacobs' externality hypothesis revisited, *Growth and Change* **36**, 167–195. doi:10.1111/j.1468-2257.2005.00273.x

Ejermo O. (2009) Regional innovation measured by patent data: does quality matter?, *Industry and Innovation* **16**, 141–165. doi:10.1080/13662710902764246

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

Fischer M. and Varga A. (2003) Spatial knowledge spillovers and university research: evidence from Austria, *Annals of Regional Science* **37**, 303–322. doi:10.1007/s001680200115

Fleming L. (2001) Recombinant uncertainty in technological space, *Management Science* **47**, 117–132. doi:10.1287/mnsc.47.1.117.10671

Foray D. (2014) *Smart Specialisation*. Innovation for Growth (i4 g) Policy Brief No. 8. European Commission.

Foray D., David P. A. and Hall B. (2009) *Smart Specialisation – The Concept*. Knowledge Economists Policy Brief No. 9. European Commission.

Frenken K. (2007). Entropy statistics and information theory, in Hanusch H. and Pyka A. (Eds) *The Elgar Companion to Neo-Schumpeterian Economics*, pp. 544–555. Edward Elgar, Cheltenham.

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**, 635–649. doi:10.1093/jeg/lbm018

Frenken K., Izquierdo L. and Zeppini P. (2012) Branching innovation, recombinant innovation and endogenous technological transitions, *Environmental Innovation and Societal Transitions* **4**, 25–35. doi:10.1016/j.eist.2012.06.001

Frenken K., van Oort F. G. and Verburg T. (2007) Related variety, unrelated variety and regional economic growth, *Regional Studies* **41**, 685–697. doi:10.1080/00343400601120296

Glaeser E., Kallal H. D., Scheinkman J. A. and Shleifer A. (1992) Growth in cities, *Journal of Political Economy* **100**, 1126–1152. doi:10.1086/261856

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. doi:10.1007/BF02017763

Hall B. H. (2005) A note on the bias in Herfindahl-type measures based on count data, *Revue d'Economie Industrielle* **110**, 149–156. doi:10.3406/rei.2005.3076

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 Trajtenberg M. (Eds) *Patents, Citations & Innovations*, pp. 403–459. MIT Press, Cambridge, MA.

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**, 459–476. doi:10.1080/13662716.2012.718874

Hill B. M. (1975) A simple general approach to inference about the tail of a distribution, *Annals of Statistics* **3**, 1163–1173. doi:10.1214/aos/1176343247

Iammarino S. (2005) An evolutionary integrated view of regional systems of innovation. Concepts, measures and historical perspectives, *European Planning Studies* **13**, 497–519. doi:10.1080/09654310500107084

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

Lux T. (2001) The limiting extremal behaviour of speculative returns: an analysis of intra-daily data from the Frankfurt Stock Exchange, *Applied Financial Economics* **11**, 299–315. doi:10.1080/096031001300138708

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 No. 12.03. Utrecht University, Utrecht.

McCann P. and Ortega-Argiles R. (2013) Transforming European regional policy: a results-driven agenda and smart specialization, *Oxford Review of Economic Policy* **29**, 405–431. doi:10.1093/oxrep/grt021

McCullagh P. and Nelder J. A. (1989) *Generalized Linear Models*. Chapman & Hall, London.

Murmann J. P. (2003) *Knowledge and Competitive Advantage. The Co-Evolution of Firms, Technology, and National Institutions*. Cambridge University Press, Cambridge.

Neffke F., Henning M. and Boschma R. (2011) How do regions diversify over time? Industry relatedness and the development of new growth paths in regions, *Economic Geography* **87**, 237–265. doi:10.1111/j.1944-8287.2011.01121.x

Nightingale P. (1998) A cognitive theory of innovation, *Research Policy* **27**, 689–709. doi:10.1016/S0048-7333(98)00078-X

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

NSF (2012) *Industrial Research and Development Information System, Historical Data* (available at: 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 Libecap G. D. and Hoskinson S. (Eds) *Entrepreneurship and Global Competitiveness in Regional Economies: Determinants and Policy Implications*, pp. 185–217. Emerald Group, Bingley.

Rigby D. L. and Essletzbichler J. (2006) Technological variety, technological change and a geography of production techniques, *Journal of Economic Geography* **6**, 45–70. doi:10.1093/jeg/lbi015

Saviotti P. P. and Frenken K. (2008) Trade variety and economic development of countries, *Journal of Evolutionary Economics* **18**, 201–218. doi:10.1007/s00191-007-0081-5

Scherer F. M., Harhoff D. and Kukies J. (2000) Uncertainty and the size distribution of rewards from innovation, *Journal of Evolutionary Economics* **10**, 175–200. doi:10.1007/s001910050011

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. doi:10.1016/j.jeconom.2006.10.017

Singh J. and Fleming L. (2010) Lone inventors as sources of technological breakthroughs: myth or reality?, *Management Science* **56**, 41–56. doi:10.1287/mnsc.1090.1072

Smith K. (2005) Measuring innovation, in Fagerberg J., Mowery D. C. and Nelson R. R. (Eds) *The Oxford Handbook of Innovation*, pp. 148–177. Oxford University Press, New York, NY.

Stuart T. and Sorenson O. (2003) The geography of opportunity: spatial heterogeneity in founding rates and the performance of biotechnology firms, *Research Policy* **32**, 229–253. doi:10.1016/S0048-7333(02)00098-7

Tavassoli M. H. and Carbonara N. (2014) The role of knowledge variety and intensity for regional innovation, *Small Business Economics*, 1–17.

Teece J. D. (1996) Firm organization, industrial structure, and technological innovation, *Journal of Economic Behavior and Organization* **31**, 193–224. doi:10.1016/S0167-2681(96)00895-5

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

Thompson P. (2006) Patent citations and the geography of knowledge spillovers: evidence from inventor- and examiner-added citations, *Review of Economics and Statistics* **88**, 383–388. doi:10.1162/rest.88.2.383

Trajtenberg M. (1990) A penny for your quotes: patent citations and the value of innovations, *RAND Journal of Economics* **21**, 172–187. doi:10.2307/2555502

Weitzman M. L. (1998) Recombinant growth, *Quarterly Journal of Economics* **113**, 331–360. doi:10.1162/003355398555595

Van den Bergh J. (2008) Optimal diversity: increasing returns versus recombinant innovation, *Journal of Economic Behavior and Organization* **68**, 565–580. doi:10.1016/j.jebo.2008.09.003

Van Zeebroeck N. (2011) The puzzle of patent value indicators, *Economics of Innovation and New Technology* **20**, 33–62. doi:10.1080/10438590903038256