DISSERTATION

PATENTS, KNOWLEDGE CREATION, AND SPILLOVERS IN GENETICS FOR AGRICULTURE AND NATURAL RESOURCES

Submitted by

Ghulam Samad

Graduate Degree Program in Ecology

In partial fulfillment of the requirements

For the Degree of Doctor of Philosophy

Colorado State University

Fort Collins, Colorado

Summer 2020

Doctoral Committee:

Advisor: Gregory D. Graff

Keith E Maskus
Stephan Weiler
Mevin Hooten

ABSTRACT

PATENTS, KNOWLEDGE CREATION, AND SPILLOVERS IN GENETICS FOR AGRICULTURE AND NATURAL RESOURCES

Increasing food, energy, and resource demand by growing global population is putting unprecedented pressure on agriculture and natural resource systems. Innovation in agriculture, energy, and other resource intensive industries contributes enormously to productivity and sustainability gains. Innovation in genetic resources and biological systems is a particularly promising yet controversial area of such innovation. Generally, it has been observed that regional clustering (economies of agglomeration) plays an important role in driving innovation. To what extent do we observe regional clustering to play a role in innovation in these industries? Especially given that production is highly diffused geographically, and research and technology are seen as highly globalized (global public goods vs. global monopolies by MNCs). The overarching questions address by this study are the following: (1) What do patents reveal about geographic patterns of knowledge creation and spillovers? (2) What economic and policy factors drive invention activity at the regional scale? And indirectly, (3) What is the role of regional clustering in driving innovations for food security and sustainability? To address these overarching objectives this study is mainly separated into three parts.

The first part delves into three related questions: (1) How have biological inventions for use in primary resource-intensive industries been spatially distributed across the United States? And, in particular, to what degree have they been geographically concentrated? (2) What are the time-space dynamics of biological inventions for these industries? To what extent does the

concentration of previous inventions effect where new inventions arise? And, (3) based on these insights, can we identify primary innovation clusters in the U.S. for these industries? This study draws on detailed information on inventor address from about 34,000 patented inventions as indicators of innovation and entrepreneurship in three closely related industries: (1) agriculture, (2) bioenergy, and (3) environmental management. To address these questions three approaches are used mapping, Moran I and regression analysis. Results indicate these biological inventions are distributed across the U.S, but highly concentrated clusters are formed in urban regions. Moreover, a spatial clustering pattern clearly exists. In term of concentration of biological inventions for these industries, a rural-urban division exists. Inventions do not tend to concentrate near production activities but tend to concentrate in urban area. The number of inventions in an area in prior years has a significant impact on the number of current year inventions. This relationship represents the localized spillover phenomenon. While we do see inventions in rural areas, rural areas do not appear to be the hotspots of innovation in agricultural, energy, or environmental biotechnologies.

The second part of this dissertation explores the covariates of regional concentration of these biological inventions for agriculture, energy, and environment in the United States. First, the geographic patterns of these inventions are analyzed using negative binomial panel regression of patented inventions by region, to identify the density of inventions overall as well as the space-time dynamics of invention cumulativeness. We find that inventions have been spatially concentrated in about 30 major metropolitan clusters, and that spatial distribution has remained remarkably stable over time. Factors of population, earnings, and farm income are correlated with their invention counts. As a first rule, these inventions are created in higher population urban regions. Although, among regions of similar population inventions are more likely closer to

agricultural production. Results clearly show the emergence of largely urban innovation clusters in agriculture and resource industries.

The third part of this dissertation broadens the scope to explore the spatial distribution and covariates of regional invention activity across Organization for Economic Cooperation and Development (OECD) countries. Three approaches are used mapping, Moran I and regression analysis to analyse the spatial distribution and covariates across OECD. The results showed that while inventions are distributed across the OECD, there again appear to be concentrated clusters in larger urban regions (another broader set of top 30 clusters). Moreover, the number of inventions made in prior years has significant explanatory power on the number of current year inventions, by region. This represents the localized spillover phenomenon. In addition, region size (as measured by population) and level of economic activity (as measured by regional income) do not appear to be related to the count of inventions for these industries. R&D expenditures (regional) and an IP index (which is national in nature but is applied to regions for this study) are strongly related to biotech invention activity for these industries. A rural-urban division does appear to exist. Finally, these invention counts appear to be negatively correlated with gross value added of agriculture by region across OECD countries.

ACKNOWLEDGEMENTS

I have no words to thank Prof. Gregory D. Graff for everything in this hard journey. It was impossible to have a Ph.D. degree without Greg's continuous support and guidance. A couple of times, I thought to quit and return to home without a Ph.D., but he always provided encouragement and showed ways to resolve the problems. He is a great mentor, academician, and a researcher who knows how to teach and how to train his students. He knows invention dynamics. Lord bless him and his lovely family.

I am also thankful to Prof. Keith E Maskus, who inspired me to advance my career in this field. I am again thankful that he accepted to be in my committee. His International Trade and Research Seminar courses, which I was fortunate to be able to attend at the University of Colorado, Boulder, were an opportunity to learn about the theoretical underpinnings of international trade, and start developing my dissertation concept note.

I am also grateful to my other committee members Prof. Stephan Weiler, for his valuable comments on regional dynamics of innovations, and Prof. Mevin Hooten, for his valuable feedback on statistical and ecological knowledge.

I would like to thank the faculty of Department of Agricultural and Resource Economics, especially Dawn Thilmany as an academic advisor for the initial two years. I also thank the administrative staff and my lovely friends and class fellows especially Aaron Hrozencik, Jada Thompson, Jason Holderieath, Jenny Apriesning, Anne Byrne, Andy Stewart, and Yoo Hwan Lee for their love and support.

Finally, I am thankful to my parents, sisters, brothers in law, nephew, and nieces for their endless support and love.

DEDICATION

I dedicate this dissertation to the Fulbright Program which provided financial opportunity, and to the lovely Fort Collins community for their love.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1. OVERALL INTRODUCTION

Governments around the world have long invested directly in research and have created state and national policies to foster greater commercial R&D to improve productivity and sustainability of essential resource intensive industries such as agriculture, natural resources, energy, and environment. More recently, the rapid growth in genetics and molecular biology has led to a boom in biotechnologies with wide scope for application in these industries. These have been viewed by many as part of what has been collectively identified "agbiotech" from the 1980s and 1990s followed by "biofuels" and "clean tech" in the late 1990s and early 2000s. Inventions in genetics and molecular biology applied in resource intensive industries are considered an important factor for food security, economic development, and increasingly for environmental sustainability, including climate change adaptation and mitigation. Inventions in these fields help meet growing demand for food given increasing populations, demand that is putting unprecedented pressure on agriculture and natural resource systems. Physiological stressors—such as drought, degraded soils, and extreme temperatures—limit productivity, profitability, and sustainability. Increasing productivity and reducing waste are two of the core strategies recommended by Foley et al. (2011) that can be achieved directly by the application of genetics and molecular biology to improve farming practices and natural resource systems.

Innovation in agriculture, energy, and other resource intensive industries contributes enormously to productivity and sustainability gains. Generally, it has been observed that regional clustering (economies of agglomeration) plays an important role in driving innovation. Then, to what extent do we observe regional clustering to play a role in innovation in these industries given

that the economic activity of production is highly diffused geographically, and research & technology are seen as highly globalized.

Much of the economic analysis of research spending and technology policy in agriculture and natural resources has taken a decidedly neoclasical perspective. The presumption appears to be that, given the right mix of spending and policy incentives, new knowledge and technologies arise stochastically from R&D activities across national innovation systems and then disseminate quickly and broadly, often as global public goods, unless some form of intellectual property protection hinders their otherwise free path to widespread utilization. Where there is a regional aspect to innovation, it is assumed to play a role in capturing and adapting these globally available R&D outputs to local agroecological and market conditions. Agricultural and resource economists have given less regard to the internal, regional dynamics of the creation of innovations, having paid less attention to the burgeoning literature on the economies of agglomeration or "clustering" in driving commercial innovation.

This study addresses three overarching questions. (1) What do patents reveal about geographic patterns of knowledge creation and spillovers in the biosciences for these industries (agriculture, energy, and environment) in which the human capital and production processes are necessarily geographically dispersed? (2) What economic and policy factors drive invention activity at the regional scale? And (3) What is the role of regional clustering in driving innovations for food security and sustainability?

To the extent that co-location creates advantage and drives innovation, this has important policy implications. Policies need to take into account the structure and dynamic nature of clustering in order to support and encourage innovation. For agriculture, these policy implications have an additional twist. To the extent that the natural constituencies and political base for

agricultural industry tends to be rural, they may be less attuned or sympathetic to funding and supporting innovation activities that will tend to agglomerate, which generally means they will locate in urban rather than rural areas.

This dissertation utilizes detailed information from inventor address data from patent publications that make up patent families, as an indicator of the location of invention. We draw upon the International Science and Technology Policy and Practice (InSTePP) Global Genetics Database, developed at the University of Minnesota from Thomson Innovation (TI) patent data, covering years 1970-2010. While the InSTePP database covers all fields of biotechnology, this present analysis is focused on biological inventions applied in the three closely related industries of (1) agriculture, (2) bioenergy and bioresources, and (3) environmental technologies, as based on Derwent World Patent Index (DWPI) Manual Code Classifications. Lead inventor zip code, city, state, and country are extracted from the InSTePP data and then analyzed at the level of Metropolitan Statistical Areas and the more rural Micropolitan Statistical Areas, to explore the degree and dynamics of spatial concentration, as well as the extent to which innovation is associated with the basic factors of population, level of economic activity, and level of agricultural production.

This dissertation is organized as follows. The next section briefly reviews the background and literature on clusters and the utilization of inventor address data from patents. Section 3 describes the global data set, presents the data cleaning process, and general descriptive statistics globally. Section 4 provides an overview the spatial distribution of these inventions in the United States and their evolution over time. Factors associated with cluster growth in the U.S. are analyzed in section 5. Section 6 expands the view and explores the spatial distribution and covariates of

clustering of invention across 17 countries of the Organization of Economic Cooperation and Development (OECD). Section 7 provides overall discussion and conclusions.

2. BACKGROUND AND LITERATURE REVIEW

2.1 History of innovation for agriculture and natural resource industries

Given what we know from the literature, what should we expect about the urban versus rural distribution of research and development (R&D) and innovation for agriculture? Agriculture was one of the earliest industries for which an innovation system was established in the United States. Amidst the chaos of the Civil War, the Morrill Act of 1862 funded the establishment of Land Grant colleges by each state, with a charter "to teach such branches of learning as are related to agriculture and the mechanic arts." While the Land Grant colleges were initially focused on education, they were intentionally dispersed across what were largely rural and agricultural regions throughout the United States. The research component of the system was only added 25 years later, by the Hatch Act of 1887, which established funding to support an agricultural experiment station in each state. Most states chose to integrate their state agricultural experiment station with their Land Grant college, thereby creating a broad network of agricultural research institutions across largely rural regions of the country (Huffman and Evenson, 2006). These Land Grant colleges and other related technical institutions (the polytechnics and the Schools of Mines) expanded their mission from agriculture and "the mechanical arts" to other natural resource industries, including energy, and, only much later, to resource management and conservation.

Thus, at about the same time Alfred Marshall began his career as an economist (1865), but still several decades before he began to articulate his theories of industrial agglomeration, policymakers in the United States were already grappling with some of the basic principles that he would later bring to light: the need for local pools of skilled labor and mechanisms for knowledge spillovers (Marshall, 1890). Moreover, in the context of agriculture, they were grappling with how

to manage policies to encourage innovation and economic development in the face of what still appears to be a paradox: How to realize the advantages that arise from economies of agglomeration in an industry in which the human capital and production processes are necessarily geographically dispersed? The form of the Land Grant system appears to have adapted to the realities of the industry, providing a diffused research and development network for a diffused industry.

Agricultural production has always been geographically diffuse, largely due to heavy dependence upon natural capital, including land and water resources. Even with agricultural production today in the United States at $396 billion, it is still very dispersed. It contributes to the economy of all 50 states, in 40 states it accounts for more than $1 billion, and in 13 states it accounts for more than $10 billion. The largest concentration by value is in California, which, at $48 billion, still only accounts for 12 percent of U.S. gross receipts (USDA Economic Research Service, 2018).

Natural resource production is similarly diffused across largely rural areas. Rich diversity of natural resources are contributing into the U.S. economy. Similarly, abundance of natural resources also provides comparative advantage to the U.S. A unique economic value for all the natural resources in the U.S. is hard to figure out. However, their distribution and extraction are available across the U.S. Nonrenewable energy (coal, crude oil, and natural gas) is produced across the U.S. predominantly in rural areas of Texas, North Dakota, California, Wyoming, Nevada, Alaska, and Oklahoma. Similarly, renewable energy (geo-thermal, wind energy, and hydropower) are mostly produced in California, Washington, and Oregon. Forestry is one of the major economic activities in the U.S. Its cultivation and harvesting is spanned out across the U.S. Under-harvesting of the forest causing significant loses to the biodiversity and threaten the timber industry.

Thus, it may seem natural that, in Paul Krugman's stylized two-sector model in his influential 1991 treatise on economic geography, one economic sector is characterized with constant returns to scale and workers evenly distributed in space, and he calls that sector "agriculture." The other sector, called "manufacturing," is characterized by increasing returns and what may be considered Marshallian externalities, with incentives thus leading to agglomeration. In the agricultural sector, there are no such incentives and thus no agglomeration (see Acs, Anselin, and Varga, 2002).

An empirical line of work tracking patterns of innovation in patent data (Malerba and Orsenigo, 1990, 1996; Breschi, Malerba, and Orsenigo, 2000; Breschi, 2010) seeks to distinguish the spatial patterns of growth of innovative activities, building upon the classic distinction between Schumpeterian Mark 1 (widening and diffused growth in innovation) versus Schumpeterian Mark II (deepening and concentrated growth in innovation) initially dubbed by Nelson and Winter (1982). In these results, agriculture is consistently identified with the Schumpeterian Mark 1 camp, with low concentration of innovative activities, relatively small size of innovating firms, low stability in the hierarchy of innovating firms, and high rates of new innovators in the patent data (Malerba and Orsenigo, 1996). Agriculture is one of the sectors that does not show signs of spatial agglomeration and is assumed that spatial proximity does not play a role in innovation (Breschi, 2010).

Whether it is because the publicly-funded Land Grant university system has already effectively addressed the spatial aspects of innovation for the industry in the United States, or because the intrinsic capital structure of agricultural production resists agglomeration, agricultural economists that study innovation in the industry have typically looked at it from the perspective of neoclassical theory (Sunding and Zilberman, 2001; Alston et al, 2010). Fundamental to that

perspective is the assumption that the existing stock of knowledge is intrinsically a public good, that is (or at least should be) freely accessible for any other innovator or producer to apply to their particular problem. Related to this, it is generally assumed that transmission of knowledge is free and costless, virtually globally. Given the experiences of the Green Revolution in the 1960s and 1970s, as genetically improved varieties of corn, wheat, and rice spread rapidly around the world, such assumptions appeared perfectly reasonable. That is not to say that geography is assumed not to matter. Rather, the primary economic question regarding regional agricultural innovation systems focused on their capacity to capture and adapt ideas to local conditions, mirroring the literature on the firm's ability to absorb, internalize, and utilize existing stocks of knowledge (Cohen and Levinthal, 1989). In short, questions of economies of agglomeration, whether in production, or in innovation, have not taken serious hold in agricultural and resource economics.

Yet, the burgeoning contemporary agricultural value chain—what may be thought of as the land-water-agricultural-food-beverage-bioenergy complex—consists of much more than simply the on-farm step of agricultural production. Innovation in inputs for on-farm production have, over the last century, gone through a series of technological revolutions, including mechanization, the advent of agricultural chemicals, scientific breeding, and more recently the concurrent revolutions in genomics and biotechnology and in data, information, and "precision" automation of on-farm production. At one time, virtually all of the inputs required for agricultural production could be sourced on the farm—including the land, the labor, the soil nutrients, the genetic inputs, the modes of traction and transport, the energy, and the mechanical implements. Over the last century, provision of agricultural inputs have successively been outsourced to industries specializing in their production, provision, and, in that vein, in the innovation necessary for their technological

improvement and productivity enhancement. The resulting complex vertical industry organization that has evolved thus defies simple categorization by industry sector codes.

The empirical work on patterns of innovation provides a glimpse into this complexity. In addition to "agriculture" these studies also included, separately, some of the agricultural input sectors. Not surprisingly, they tend to exhibit more of the characteristics of manufacturing industries. Malerba and Orsenigo (1996) find mixed evidence for "agricultural chemicals," but they squarely place "organic chemicals" and "bio- and genetic engineering" in the Schumpter Mark II camp exhibiting more concentrated innovation activity. In Breschi, Malerba, and Orsenigo's (2000) update of the analysis "agricultural chemicals" had earned its placement in the Schumpter Mark II camp as well. In fact, an entire literature has recognized and analyzed the dynamics of clustering in biotechnology (Audretsch and Stephan, 1996; Zucker and Darby, 1996; and many others citing these). While such clustering has been explored for the subset of the biotechnology industry that applies to agriculture and natural resources (Ryan and Phillips, 2004); however, these trends have not been thoroughly documented and empirically analyzed.

2.2 Marshallian clustering

The burgeoning literature in economic geography, regional science, and urban economics mainly builds on Marshall (1890, 1920) and Krugman (1991). Marshal's (1920) theories have provided the foundation to study geographic and spatial concentration of innovative activities.

Ellison et al. (2010) explain Marshall's theories of industrial agglomeration, identifying the following Marshallian forces: (a) proximity to customers and suppliers; (b) labor market pooling, and (c) intellectual or technology spillovers. Proximity to suppliers and customers helps firms have easy access to inputs or shipping good to downstream customers. Ellison et al argue that firms trade off the distance between customer and suppliers based on the cost of inputs and

finished goods. The risk sharing properties of the large skilled labor pool is emphasized. If, for some reason, some firms are more productive than others are, then workers can switch among them, reducing variance in wages. The point Ellison et al. emphasize is that clustering facilitates worker-firm matching. The agglomeration may also be formed due to natural advantages, and nature of industries. For example, agriculture and natural resources industries might be co-agglomerated in rural or urban peripheries. Similarly, industries like oil refining, and ship building might opt for coastal locations. Highly dispersed and rural industries challenge the question of whether the Silicon Valley innovation model applies to these types of industries.

The final reason firms cluster is to speed the flow of ideas. As workers learn new skills, and researchers make new discoveries, by co-locating researchers and managers of a firm are able to gain access to less formal or localized information exchanges gaining at least a temporal advantage. Ellison et al cite one additional reason that does not follow from the original Marshallian theories. Some industries see natural cost advantages in some regions, perhaps due to quality of capital endowments or strategic location near a port, and as they flourish, other firms then follow.

2.3 Knowledge as a factor in clustering

Audretsch and Feldman (1996) examine the geography of innovation in its own right. They start by arguing that knowledge externalities are more prevalent in industries where new economic knowledge (industry R&D, university R&D, and skilled labor) play a greater role. In their analysis, it is the spatial limitation of spillovers of new economic knowledge that drives the concentration of innovation activities.

Glaeser and Resseger (2010) examine the relationship between cities and skills. They consider agglomeration a result of the spread of knowledge within cities. They do not view

agglomeration to be due to good governance, easy access to ports or harbors, or easy access to capital. They find a strong connection between worker productivity and metropolitan area population. The result is that economies of agglomeration are strong for cities with higher skill levels and almost non-existent in less skilled areas. Urban population density is important because proximity spreads knowledge and skills, making workers more productive and entrepreneurs more successful. Their results suggest a strong complementarity between city size, learning, and skills, with agglomeration effects stronger for cities with more skills.

Scholars have shown that the geographic and spatial concentration of knowledge production processes are driven by several key factors, including level of opportunity[1], appropriability[2], cumulativeness[3], and knowledge base[4], all of which vary across sectors (Breschi, 2000). Sun (2000) shows that in China the relatively high levels of economic development of coastal provinces and large centers of population in inland provinces are the most significant factors explaining the spatial clustering of innovations within China where state-owned and collectivized industries still dominate in measures of innovation. Usai (2011) finds that innovative activities in OECD countries are also clustered, with spatial concentration increasing with the passage of time.

Boschma and Fornahl (2011) recognize the dynamic nature of innovation clusters, and to understand clusters' emergence consider them as a product of path dependence process. An appropriate analytical framework is still not available to evaluate cluster evolution. Considering

[1] Ease of innovating for any given amount of money invested in search

[2] Effectiveness of various means of protecting innovations from imitations

[3] Degree of persistence or serial correlation among subsequent innovations

[4] Tacit vs. codified, complex vs. simple

and exploring co-evolutionary process and network dynamics in clusters are important future research challenges described by Boschma and Fornahl (2011). Ter Wal (2013) posits that due to the inclusive nature of clusters, networks of local collective learning (co-inventorship) is not exhibited. Geographic orientation, connectivity, average path length, and clustering coefficients are the important network properties to detect the emergence of the local collective learning.

2.4 Inventor address data from patents to study clustering and knowledge spillovers

Scholars have specifically used inventor address information in patent data to study the geography of knowledge clustering and spillovers (Jaffe, Trajtenberg, and Henderson, 1993; Acs, Anselin, and Varga, 2002; Thompson and Fox-Kean, 2005; Known, Lee, Lee, and Oh, 2017). Patent data may contain some or all of the following inventor address information: country, city/town, state, postal or zip code, and street address. The usage of this information is complicated by the fact that patents often have multiple inventors and these inventors lives in different areas. Jaffe, Trajtenberg, and Henderson used two procedures to resolve multiple inventors' addresses in patent data. First, U.S. cities were assigned to counties based on an available city directory, and inventors were thereby allocated to a Metropolitan Statistical Area (MSA) based on state and county. Second, those patents with more dispersed inventor locations were assigned to a state and MSA based on a plurality of inventors. For example, for a patent with one inventor in Bethesda Maryland, one in Alexandria, Virginia, and one in rural Virginia the patent will be assigned to Virginia for its state and to Washington DC for its MSA.

Thompson and Fox-Kean (2005) converted the towns or cities and states to counties and then to 17 Consolidated Metropolitan Statistical Areas (CMSA) based on Office of Social and Economic Data Analysis (OSEDA) of the University of Missouri in order to assign geographic locations to patents. To assign a unique geographic location to each patent, a single inventor was

randomly selected from the list of domestic inventors. Known, Lee, Lee, and Oh, (2017), mapped each patent to one of 17 CMSAs following the same methodology.

Researchers have considered different units of analyses within the national (e.g. U.S, China), and regional levels (e.g. Europe, OECD) to study the space-time dynamics of biological inventions. For the U.S. spatial units of analyses that have been explored include state, county, and metropolitan areas. These geographic demarcations of state and county creates biases in the analyses and cannot measure accurately the space-time nature of knowledge creation and spillovers. Lim (2003) considers limiting the number of metropolitan areas not helping to improve the significance of the results of the study.

Similarly, scholars have considered questions of the aggregate level of technological categories to measure the geography of innovative activities. Aggregating patents across all technological categories cannot measure the spillovers phenomenon meaningfully (Known, Lee, Lee, and Oh, 2017).

2.5 Lessons for spatial analysis from the ecological literature on species distribution

An important field in the ecology literature provides a potential interface to the economics of innovation. Insights on patent distribution can be drawn from species distribution theories. For example, Ideal Free Distribution (IFD) theory (animals distribute themselves among several patches of resources) is not different from inventor cluster emergence and evolution. Ecological fallacy (ecological inference fallacy) is a logical fallacy in the interpretation of statistical data where inferences about the nature of individuals are deduced from inference for the group to which those individuals belong.

Hefley et al. (2016) posit that location error (when true location is different than reported location) causes unreliable inferences concerning species habitat relationships. For the whooping

crane resource selection, they find that location error can cause up to five-fold change in coefficient estimates. If the related internal and external forces or processes are not determined accurately, then the ecological models will give bias or statistically insignificant results. It is very challenging to detect general forces or processes in ecology. However, deep understanding of these ecological forces or process requires knowing "what laws of nature are and what roles they are supposed to play in scientific theory."

This overlap presents a fascinating area to be explored for spatial analysis of patent data. We are not aware of studies that have discussed the importance of patent location error and how it might influence our understanding of knowledge creation and spillovers. We believe that location error in patent data should affect results about knowledge spillovers. For example, if a patent is registered in the periphery instead of a hub/cluster then this patent has a minimal level of knowledge creation and spillover instead of a patent close to the cluster. If we consider such patents in our model with location errors, then we obtain biased coefficients. To minimize location error, we need to consider patent data at zip code and city levels.

3. GLOBAL PATENT DATA DESCRIPTION

Patents have long been considered as useful indicators of innovation activity (Schmookler, 1954; Griliches, 1990; Hall, Jaffe, and Trajtenberg, 2001; Acs, Anselin, and Varga, 2002). Yet, there are well established limitations to the use of patent data as well. Patents typically represent early stage technologies, before they are translated into commercial applications, and thus they can be difficult to attribute to specific industries. Moreover, patents are highly variable in value or importance. Still, among the available alternatives, patents provide a uniquely comprehensive view across multiple parts of the innovation system—including academic, entrepreneurial, and corporate R&D—across industries, and across geographies.

3.1 The InSTePP Global Genetics Patent Database

This study utilizes detailed information about patent families as an indicator of invention. We use the International Science and Technology Policy and Practice (InSTePP) Global Genetics Patent Database developed at the University of Minnesota from Thomson Innovation (TI) patent data (covering years 1970-2010), which is broadly classified into industries and technologies. The detailed classification, rationale of the data, and its structure have been defined by Graff et al. (2014). This analysis is focused on biological inventions applied in three closely related industries (1) agriculture, (2) bioenergy and bioresources, and (3) environmental remediation are extracted based on Derwent World Patent Index (DWPI) Manual Code Classifications. To the extent that inventor address data are available in the patent records, lead inventor zip code, city, state, and country have been extracted from the InSTePP data base. To study the dynamic nature of knowledge creation and its agglomeration, we explore analysis at the level of zip codes (4,979) and cities (29,217) to improve the significance of the findings.

The InSTePP Global Genetics Patent Database is a comprehensive compilation of all patent documents that contain biological sequence information—including nucleotide sequences and protein or amino acid sequences—and related filings in biological subject matters, thus targeting with a high degree of accuracy inventions across the full range of biotechnology and genetics. The collection identifies 1,093,038 inventions, from 1970-2010, represented by patent families with filings in 94 countries (Graff, Philips, and Pardey, 2015). For this study, we select those biotechnology and genetics inventions identified by Derwent World Patent Index (DWPI) Manual Code designations to be associated with industrial applications in agriculture, energy, and environmental management.[5]

Five steps were performed to develop and clean the InSTePP Global Genetics database (Graff, Philips, and Pardey, 2015). A number of core queries were run over the full text collections of Thomson Innovation (TI)—including United States patent documents, European Patent Office patent documents; WIPO patent application documents—as well as the Chinese Intellectual Property Office from 1970-2010 to identify patents associated with or containing biological sequences, associated biomolecules, genetic traits, biological resources, and living modified organisms. Additional queries were run to expand the dataset to include all associated patent records worldwide.

Step 1: Three different highly targeted queries were conducted on full text patent collections in order to identify all patent documents that contain reference to biological sequences:

[5] The InSTePP Global Genetics Patent Database utilizes Thomson Innovation's proprietary Derwent World Patent Index (DWPI) Manual Code classifications to assign each patent to one or more of eight high level industries: (1) pharmaceuticals, (2) chemicals, (3) veterinary, (4) agriculture, (5) energy, (6) environment and natural resources, (7) food and beverage, and (8) pulp and paper.

i. First, the full text collections of Thomson Innovation (TI) for the US, EP, WO, and CN from 1970 to 2010 were searched for the text string "SEQ ID", which is a standard term used ubiquitously in the text of patent documents to make reference to tables that diagrammatically illustrate exact nucleotide or peptide (amino acid) sequences claimed or involved in the invention.

ii. The second query strategy involved identifying all US, EP, JP, and KR patent publications documented as the source of one of the nucleotide or peptide sequence accessions listed in one of the major biological sequence databases, including GenBank, hosted by the National Center for Biotechnology Information (NCBI) in the U.S., the DNA Database of Japan (DDBJ), and the European Molecular Biology Laboratory (EMBL).

iii. The heximer queries were a set of queries which systematically queried for text strings which represent nucleotide sequences in the text of the patents. They are called heximers because the search employed six letter strings of nucleotide sequences, such as "GCTGCA". The search utilized all possible combinations of the five possible nucleotide characters.

Step 2: A broader query was conducted using 1,315 International Patent Classifications (IPC) codes found to be the most common IPCs among the results of the sequence based queries in step 1 above, and judged by our team of investigators to be relevant and sufficiently specific to biological subject matters, based on the IPC description.

Step 3: "Patent family expansion" queries were run to expand the dataset to include all patent family members of the records identified by the IPC queries in Step 2 above.

Step 4: After combining the results of steps 1, 2, and 3 above, cleaning and structuring, the main dataset consisted of a total 8,511,345 patent application and issued patent records.

Step 5: Often all the records were identified and the relevant data were downloaded, the patent documents were assembled into two more highly aggregated levels. First, records were aggregated into unique attempted patents (combining records at a given patent office level with the same patent application number). Second, records were aggregated into patent families (at the national and/or international level, as approproate).

As explained above, the InSTePP Global Genetics Patent Database provides a comprehensive collection of inventions involving biological subject matter. At the core of the collection are patents involving molecular biology, as indicated by the presence in the patents of biological sequences—whether nucleic acids or peptides. But these constitute only about 20% of the inventions in the database. Most of the inventions cover microbial cell lines, as well as modified plant and animal cells, tissues, and whole organisms, biological extracts and other biomaterials, as well as biological research tools, breeding methods, diagnostic methods, and bioinformatics inventions (Graff, Philips, and Pardey, 2015). Therefore, we prefer to use broader terms, such as "biological inventions" to refer to the technical scope of these patent data, rather than "biotechnology" and "genetics." These latter terms are often interpreted, at least in industry and the press, to refer only to applications of molecular biology or genetic engineering. If the term "biotechnology" is interpreted broadly, as its etymology suggests, to mean "technology" related to or using "biology" then that is an accurate representation of these patent data.

3.2. The patent family and its advantages for analysis

Our patent data are organized into patent families. A patent "family" is a set of one or more records, from one or more patent offices around the world, that all "relate to" (or "descend from") the same initial invention (Martínez, 2012). When an invention is made, a first (or "priority") patent application is filed at an initial patent office. This creates the initial or "priority"

record in the patent data for that invention. However, that initial or priority record may then be followed by related records in the patent data, such as the grant or publication of a patent based on the priority application, as well as new applications (and, potentially, subsequently issued patents) in other (foreign) patent offices.

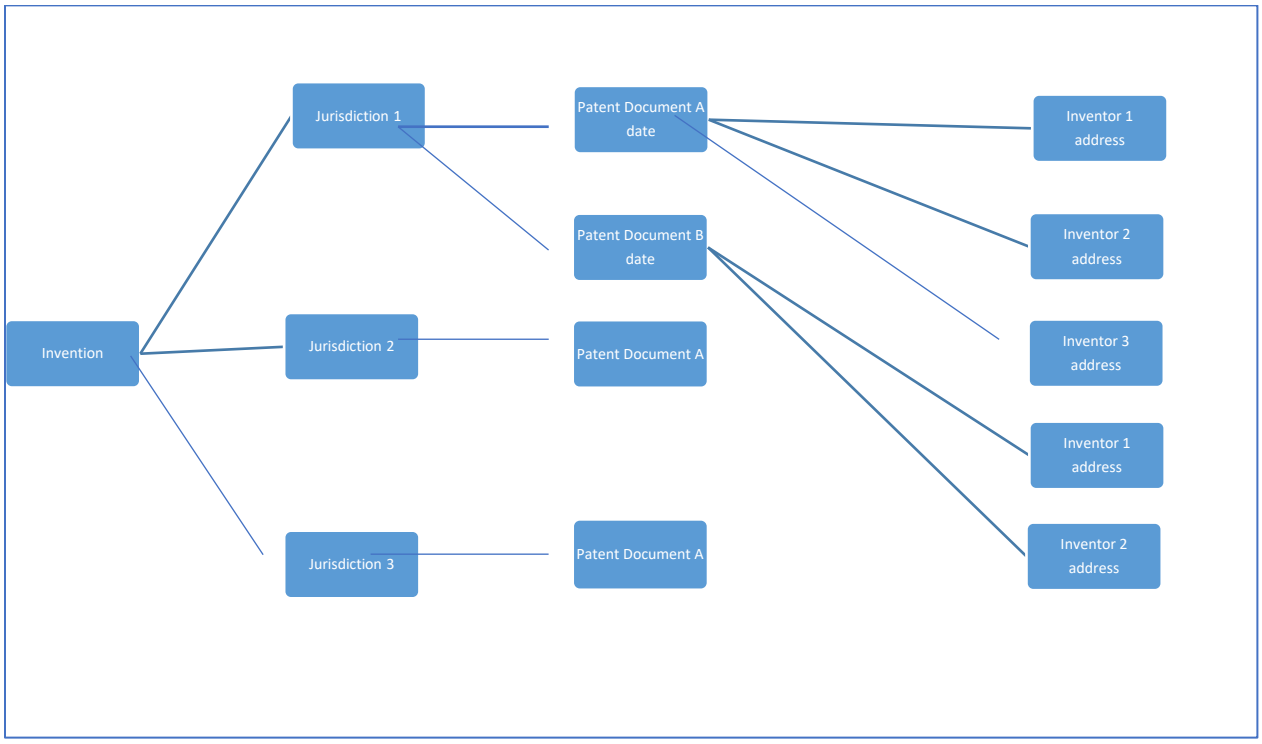


Figure 1. Example of a patent family

Unfortunately, the definition of a patent family is established by patent data providers for the ease of searching patents, not by law. Therefore, there are variations in patent family definitions that result in different patent data aggregations (Martinez, 2011). Generally, patent families fall into one of a few main types.

- Domestic Patent Family: A domestic patent family includes only patent documents filed in one country or patent office.
- International Patent Family: International patent families are have patent documents published in more than one jurisdiction.

- Singleton Patent Family: A family which contains only one patent publication of record. Necessarily it is filed in only one jurisdiction and it is the only member of that patent family.

There are several benefits of working with patent families. Martinez mentions five uses of patent family data: (i) It helps to present the double counting of single inventions when adding information from different patent offices (ii) It also helps to neutralize home advantage. When patent data are taken from single patent office, countries prefer home country first than the international jurisdiction. This will overestimate the patent propensity. Patent family data resolve overestimation of patent propensity (iii) Patent families are useful to forecast the patent applications counts (iv) To analyze and understand internationalization of technologies, and (v) to estimate the patent value.

Table 1. Distribution of patent family sizes in term of publication records, for the main set of 127,410 inventions for which location of lead inventor was identified

Number of patent families	Number of publication records per patent family	Percent of patent families of this size	Cumulative percent
43,850	1	34.41%	34.41%
20,420	2	16.00%	50.43%
19,087	3	14.98%	65.41%
16,077	4	12.61%	78.41%
10,194	5	8.00%	86.02%
5,734	6	4.50%	90.52%
3,090	7	2.42%	92.94%
1,978	8	1.55%	94.49%
1,315	9	1.03%	95.52%
910	10	0.71%	96.23%
	>10	3.77%	100%
Sum: 127,410	Minimum = 1		

Table 1 reports the distribution of patent family size for the main set of inventions for which location of lead inventor was identified and which were therefore used in this analysis. Patent family size is somewhat skewed: patent families in the data are made up of an average of 5.9 patent

records each; yet, over half of the patent families consist of just a single patent record (mean size is 5.9; median size is 1).

3.3 DWPI Manual Code classifications: Determining industry of application for inventions

The InSTePP Global Genetics Patent Database utilizes Thomson Innovation's proprietary Derwent World Patent Index (DWPI) Manual Code classifications to assign each invention's patent family to one or more of eight high level industries: (1) pharmaceuticals, (2) chemicals, (3) veterinary, (4) agriculture, (5) energy, (6) environment, (7) food and beverage, and (8) pulp and paper. Close to 10,000 DWPI Manual Code numbers were assigned by the analysts at InSTePP to one or more of these eight high-level industries, according to the scheme in Table 2.

Table 2. DWPI Manual Code Classifications and definitions

DWPI Manual Codes	Industry of application, assigned by InSTePP
Section A	Chemistry
Section B	Pharmaceuticals
Section C	Agriculture
Sections D01-D03, D05-06, D10	Food & Beverage
Section D04 plus selections from B12-K04 (Environmental testing)	Environment
Section D05 (Waste fermentation)	Environment
E11-Q (waste recovery, purification, treatment)	Environment
Section F	Pulp & Paper
Section H, plus selections from D04-D05 (Fermentation)	Energy
Section D10 (Animal and vegetable oils)	Food & Beverage
E10-E11 (alcohols, hydrocarbons), L03, and X16	Energy
Selections from A12-V, B04-P, B11-C, C04, C12-K, C14, D03-G, D05-H16 (transgenic animals)	Veterinary

Based on DWPI manual code classification associated with the industry applications in Section C (agriculture), Sections D04, D05, E11-Q (environment), Section D04-D05, E10-E11, H, L03, and X16 (energy), those patents associated with agriculture, energy, and environment were selected. These industries have a vast scope of utilization in production. Innovation in agriculture,

energy, and other resource intensive industries contributes enormously to productivity and sustainability gains.

This selection by industry of application resulted in an initial set of 210,057 patent families (inventions), consisting of 1,241,911 patent publications across 94 different patent offices, and representing just over 20 percent of the total inventions in the InSTePP Global Genetics Patent Database. A given invention may be categorized in one, two, or even all three of the selected industries of application. Table 3 shows individual and overlapping shares of inventions in the final main dataset across the three selected industries of application.

Table 3. Cross table of counts of inventions categorized by industry of application based on DWPI Manual Codes, for the main set of 127,410 inventions for which location of lead inventor was identified, including inventions assigned to multiple categories

Agriculture	60,468 (47.5%)		
Energy	3,476 (2.7%)	20,426 (16.0%)	
Environment	5,226 (4.1%)	6,296 (5.0%)	30,007 (23.6%)
	Agriculture	Energy	Environment

* 1,511 (1.2%) patent families are categorized in all three industries N= 127,410

3.4 Cleaning and characterization of inventor address data

Next, we seek to identify the geographic location where each invention was made by exploiting the inventor address information from the patent data. Much of the literature uses the address listed for the patent applicant or "assignee," which typically is the address of the head office of the company that employs the inventors. In contrast, the address listed for an inventor is typically their home residence. Since, we are analyzing the geographic distribution of actual innovation activities, it is more helpful to know the physical location of the individual inventors, which may be employed at a branch office, R&D facility, or field station different than the location of their employer's headquarters, which is most likely urban. The locations of inventors'

residences are expected to provide a better measure of this distribution than the locations of the applicants' head offices.

Furthermore, to resolve the question of how to allocate those inventions that were made by multiple inventors with different addresses, we utilize just the address of the lead inventor on the priority patent record for that patent family. This choice, in essence, designates a single primary geographic location for each invention, corresponding to that of the lead inventor. The main justification lies in the tendency that a first-listed inventor often has made the largest contribution to the invention, and/or to has led the team of inventors. Therefore, that individual's physical location may be shared by other inventors on the patent, or if not, may merely be considered more important to the creation of the invention. Other researchers have employed plurality (Jaffe, Trajterberg and Henderson 1993) and random selection (Thompson and Fox-Kean, 2005) criteria to avoid multi-counting of locations associated with multiple inventors. The choice of criteria involves tradeoffs between complexity of the measure and missing out on important contributor(s), and thus introducing different selection biases or location errors.

For those records that did report information on inventor address, extensive cleaning was required. Address formats varied significantly and can consist of any combination of inventor name, city, zip code, state, and country. Typically, the inventor address data consist of a text string that starts off with the name of the inventor, followed by city, zip code, state, and country. Unfortunately, the string format is not uniform, for example:

- Nielsen Flemming Skovgaard,3000 Helsingør,DK
- CHEIKY Michael C.,Thousand Oaks,CA,US
- Tao Bernard Y.,Lafayette,IN,US
- Lee Joung Phil,Incheon,KR

- Yin Hongmei,Dalian,CN
- OSTERHOUT ROBIN E.,US

However, some older records, mostly among U.S. filings, only included domestic address information, such as:

- Venkataraman Mahesh,Houston,TX
- Behrouzian Behnaz,Sunnyvale,CA

In the vast majority of inventor-address data strings, the city was listed just before the two-letter country or state code. Thus, the data string was split at separating commas, converted to lower case, punctuation and spaces were removed, and special characters were converted to corresponding plain roman characters (i.e. ñ to n, ä to a, etc).

Address data generally contain two-letter country and state codes. In many cases their meaning is not obvious from context, requiring strategies to resolve ambiguities, and thereby correctly allocate inventions. Due to an inversion between the state and the country code fields, confusions arose from several hundred records for “Buenos Aires, AR” as well as a number of other very Spanish sounding names of cities purportedly in Arkansas (like Bahia Blanca, AR). It is therefore important in this data to check for known code matches for locations, including the following:

- AR: Arkansas vs Argentina
- CA: California vs Canada
- CO: Colorado vs Colombia
- DE: Delaware vs Germany
- ID: Idaho vs Indonesia
- IL: Illinois vs Israel

- IN: Indiana vs India
- KY: Kentucky vs Cayman Islands
- MT: Montana vs Malta (not a big one)
- PA: Pennsylvania vs Panama
- WA: Washington vs Samoa.

For those inventors where the last two characters of the inventor address field was an ambiguous code, attempt was made to resolve by identifying the inventor's city. To do so, we undertook a review of all potentially ambiguous state/country codes (AR, CA, DE, IL, IN, etc) against detailed lists of city names for the corresponding state and/or country. If the inventor address string ended in the two-character code "CA," and if the preceding data segment matched any of the listed California cities, then INVENTOR COUNTRY was designated as "US", otherwise, INVENTOR COUNTRY was designated as "CA" (Canada). In a few cases, it was necessary to resolve ambiguous cities. There are number of cities that have similar names, located inside and outside the U.S. respectively. For example, Georgetown exists in California (CA) and in Canada (CA).

Potential problems were identified and weighed. If it seemed to be an important ambiguity, it would be resolved. Disambiguation exercises were first carried out for California and Canada. Similar exercises were extended for IL, IN, and others. and For those that were minor, we tried to minimize spending time. Overlooking small cities was not going to be an issue in terms of significance of results. For major cities, lack of uniformity of spelling or standardization of names was another issue that we resolved. for example, "New York" and "N.Y." needed to become, "NY." Similarly, standardization of zip codes, when available, was conducted. For U.S. inventors occasionally the data have a 4-digit suffix to US zip codes. We applied different algorithms to fix

such discrepancies. Finally, we observed large number of inventor data consisted of just inventor names.

3.5 Issues in assigning geographical coordinates

Unfortunately, in the data set the names of the cities are often misspelled in which case geocoding software does not recognize a city to assign coordinates. In such cases it was necessary to clean the names of cities or assign coordinates manually. The additional state information for U.S. inventors helps to assign coordinates to a city when it is recognized to be in a state. Furthermore, there were major discontinuities in available zip codes, and harmonization of zip code data involved significant manual data cleaning.

3.6 Summary of the patent dataset

The resulting "Main Collection" file contains 1,657,651 patent observations across the three industry categories of agriculture, energy, and environment, that collapse into 210,057 patent families. Of these 210,057 patent families in our main collection, only 127,410 contain some form of information on inventor address. Out of the total 127,410 patent families (inventions) almost half, or 62,378, only indicate the country of the lead inventor. Another, 43,697 indicate the city (and, if relevant, state) of the lead inventor. Just, 21,332 contain city and the additional detail of a postal or zip code. Table 4 summarizes the lead inventor address information for U.S. inventors on priority filing by patent family.

Table 4. Characterizing availability of lead inventor address data types on priority filings, by patent family, globally

Geographic Level: Globally	
Lead inventor address data based on priority filing	Patent family (invention) counts
Country only	62,378
Country + city/state	43,697
Country + city/state + zip	21,332
Total	127,407

Figure 2 maps the global pattern of these inventions. We can see the inventions are distributed across most of the United States, Europe, Japan, and major emerging economies like China, Brazil, South Africa, India. This image represents all countries with lead inventors in the global dataset.

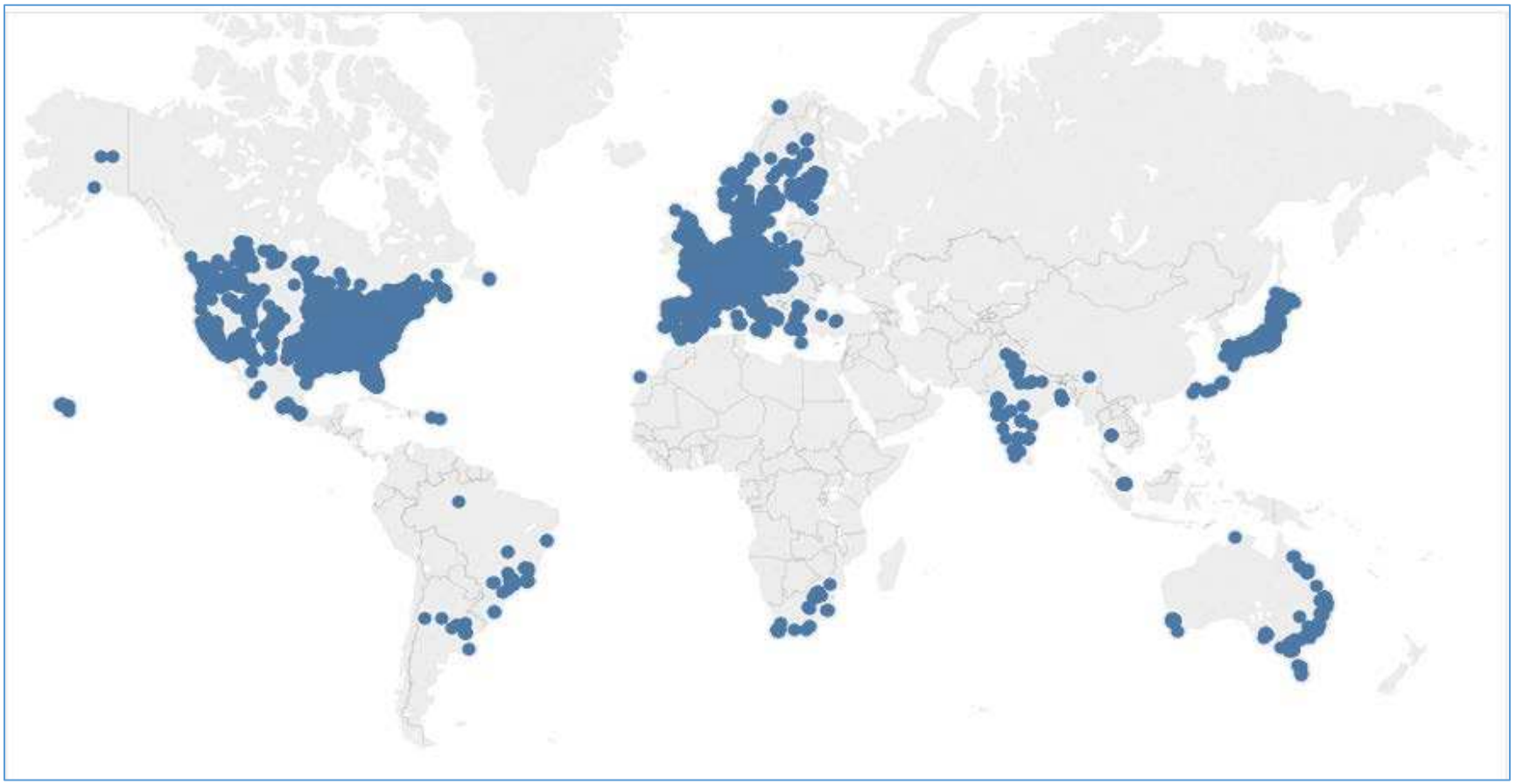


Figure 2. Geographic pattern of where biological inventions, for use in agriculture, energy, and environment, are made, globally

Figure 3 represents the proportion of inventions by country for which data of a lead inventor is available at the city or postal code level. The United States has 58.6% of the global inventions. The other countries with relatively large shares are Japan 6.7 %, France 4.6%, Germany 4.4%, and Great Britain 4.2%. Those countries whose share are less than 0.5 % are categorized in "Others." The combined share of all those countries that are part of "Others" are 8.4%.

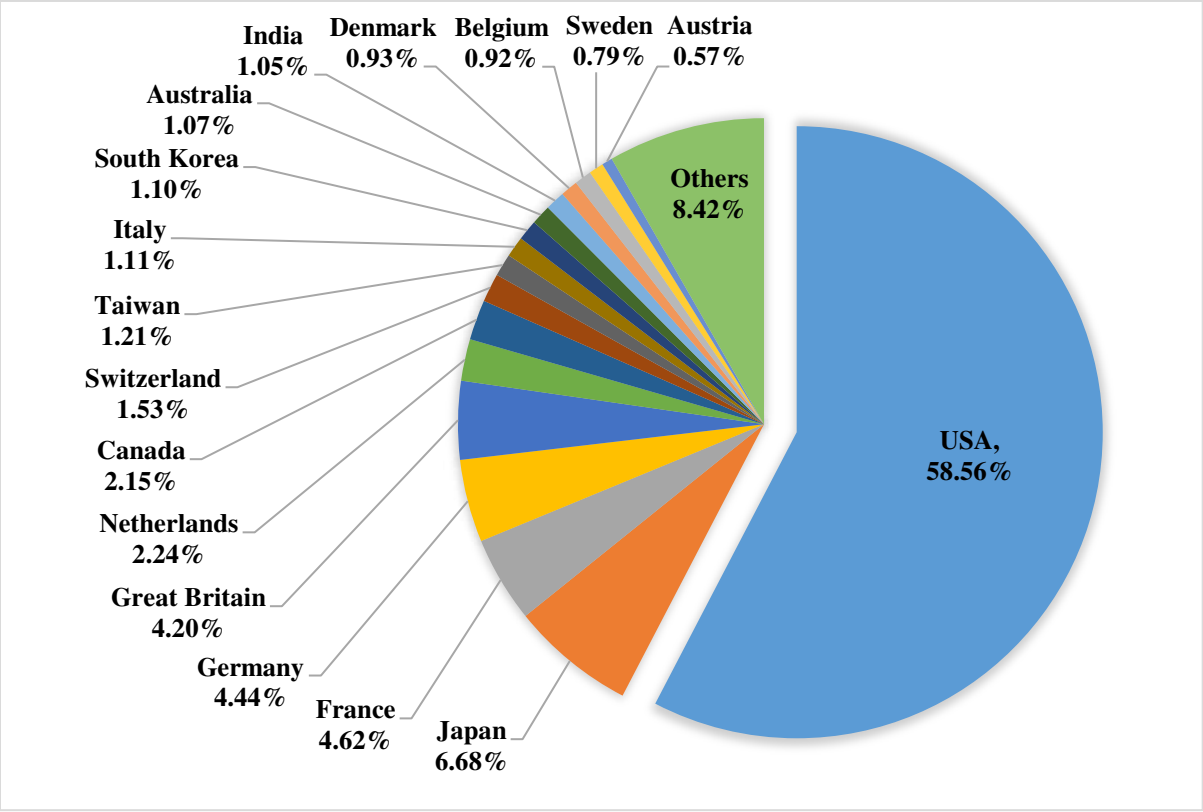
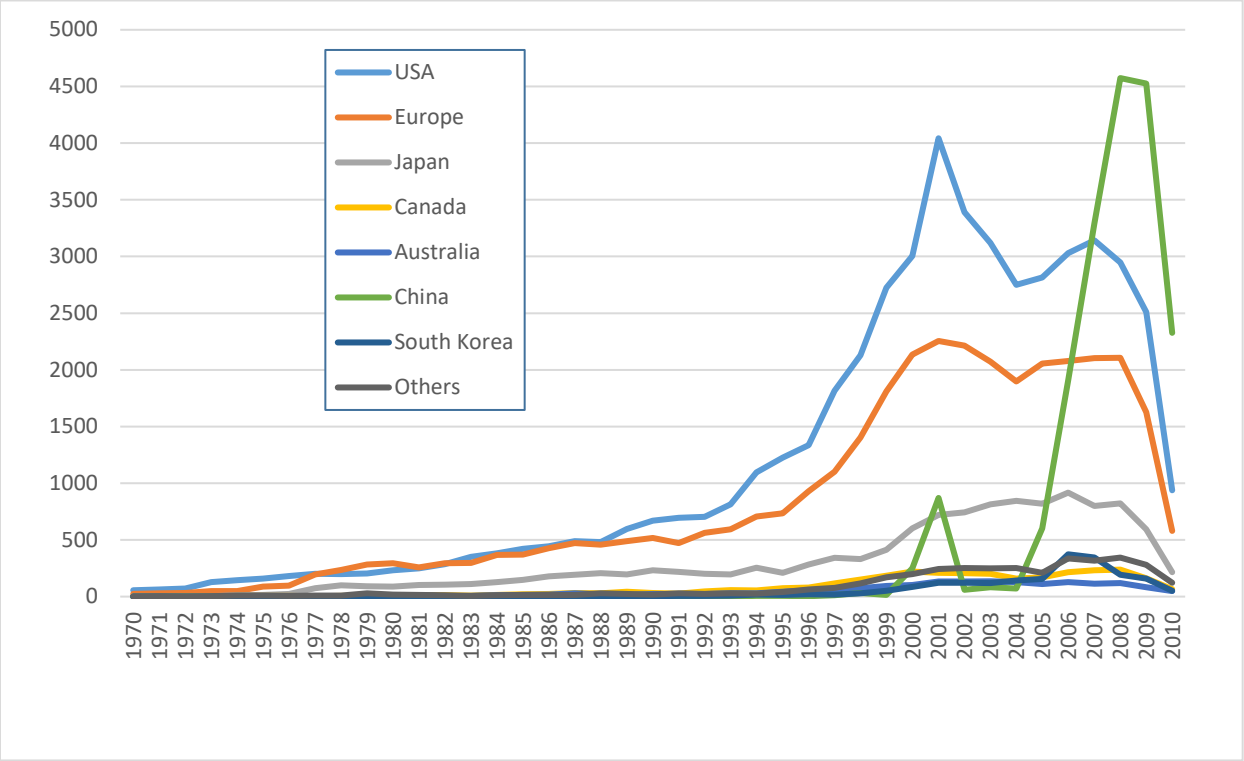


Figure 3. Share of inventions in the dataset, measured by patent families from 1970-2010, for which city or regional level inventor address is available



N= 127,282

Figure 4. Growth in global biological inventions with applications in agriculture, energy, and environmental management

The annual growth of the 127,282 biological inventions for which at least the country of the lead inventor is known is shown in Figure 4. For the mentioned countries we can clearly see the increase in biological inventions from 1980 onwards. The Green Revolution of the 1970s, the Diamond v. Chakrabarty decision on patenting of living organisms in 1980 (particularly in the U.S), the Bayh-Dole Act of 1980 (particularly in the U.S) are some of the primary reasons for this increase. Introduction of the Trade Related Aspects of Intellectual Property Rights (TRIPS) agreement also contributed in the rise in inventions. The bursting of the tech bubble in 2001 negatively affected the trend, and inventions gradually decreased. After the economic recovery in 2004, inventions again began to increase. The financial crisis of 2007-08 affected the rate, however the steep decline after 2008 is largely due to data truncation. The patent application filing and its

examiner's first substitutive review take about 21 months. In some cases substantive examination takes 3-4 years. Grant or refusal takes place after this examination. The InSTePP data set was compiled in 2011 and 2012, therefore, inventions that had been filed in 2008-2010 but were still not published or otherwise not included in the electronic patent records by 2011 are not part of this data set.

4. THE REGIONAL CLUSTERING OF BIOLOGICAL INVENTION FOR AGRICULTURE, ENERGY, AND ENVIRONMENT IN THE UNITED STATES

4.1 Introduction

The United States is one of the largest agricultural and energy producers in the world. Key factors that contributed to this success story are investment in agricultural research and stimulatory agricultural policies (Alston et al 2010). The United States has invested significantly in overall science spending. In 1980, United States accounted for 31% of the world's science spending, and 33% in 2006 (Alston et al 2010). The United States has also invested extensively in agricultural R&D, and devised policies to encourage public and private agriculture R&D results to increase agricultural productivity and sustainability. In the United States, the Land Grant university system was developed in the 19th century to address the needs for innovation in geographically diffused and relatively rural industries. Successive waves of mechanical, chemical, biological, and information technology innovations have transformed U.S. agricultural and resource sectors into the high technology industries that they are today.

Most of the economic analysis of research spending and technology policy in agriculture and related resource and environmental fields has taken a decidedly national and even internationalist perspective. The presumption appears to be that, given the right mix of spending and policy incentives, new knowledge and technologies will arise from national innovation systems and disseminate to industry broadly, often even as global public goods. Analysis has given less regard to the internal, regional dynamics of invention and commercial innovation. Yet, a burgeoning literature on innovation and entrepreneurship has pointed to the crucial role of economies of agglomeration or "clustering" in driving commercial innovation.

If clustering of innovation activities are crucial for driving commercial innovation and if such clustering tends to follow from agglomeration of other factors such as production and transportation infrastructure, it stands to reason that such clustering may be less prevalent for industries such as agriculture and natural resources, for which production activities tend to be widely-dispersed geographically and are predominantly rural. To what extent do innovation activities for these industries tend to cluster? And where? This section addresses three interrelated research questions. (1) How have biological inventions for use in primary resource-intensive industries—such as agriculture, energy, and natural resources—been spatially distributed across the United States, and, in particular, to what degree have they been geographically concentrated. And, if so, do they tend to occur in rural regions where the main production activities are located? Or, are they in urban areas associated with upstream input manufacturing or downstream output-processing industries? (2) What are the time-space dynamics of biological inventions for these industries? To what extent does the concentration of previous inventions effect where new inventions arise? And, (3) based on these insights, can we identify primary innovation clusters in the U.S. for these industries? What implications can be drawn for U.S. R&D policies?

This chapter is organized as follows. The next section briefly reviews the data and methodology, presents descriptive summary statistics and analyses to describe geographic distribution of biological inventions across the U.S. Section 4 brings in discussion and conclusions.

4.2 Data on biological inventions in the United States

Of the 127,401 inventions for which we have inventor address data, the lead inventor on 48,693 was in the United States. For these 48,693 U.S. inventions, available inventor address data was mixed, especially for inventions prior to the 1990s. For 14,497 of these, inventor address data simply indicates the inventor's country of residence as the United States, with no other

information. For another 29,217, the inventor's city and state are provided. Another 4,979 also contain inventor zip codes. Table 5 summarizes the available address data types for U.S. lead inventions.

Table 5. Characterizing availability of lead inventor address data types on priority filings, by patent family, for U.S. inventions

Geographic Level: US	
Lead inventor address data based on priority filing	Patent family (invention) counts
Country only	14,497
Country + city/state	29,217
Country + city/state + zip	4,979
Total	48,693

Geographic coordinates were assigned to each invention by batch algorithm, using information on city, state, and/or zip code of the lead inventor. For those that were not recognized, further cleaning, correction of misspellings, and assignment of geographic coordinates was undertaken by hand, as needed, in an effort to ensure that minor cities and towns (more typically in rural areas) were not underrepresented the final dataset. Complete city names and geographic coordinates were thus assigned for 34,196 inventions.

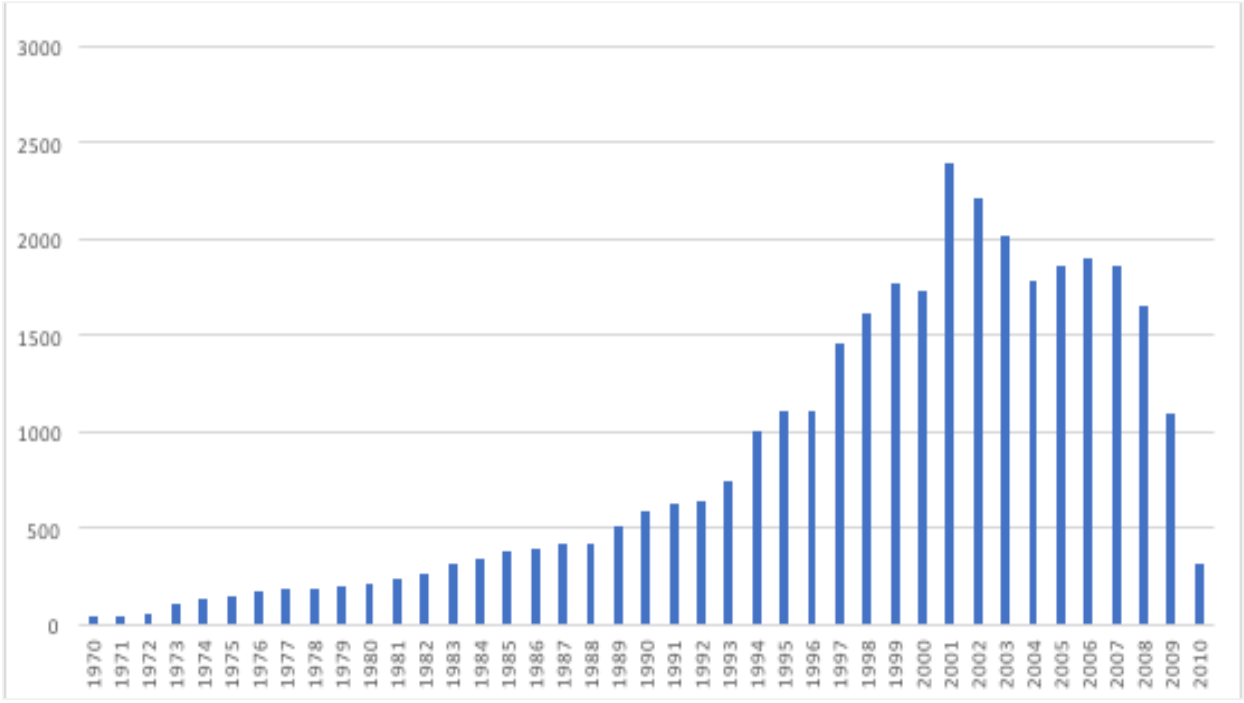
Table 6. Cross table of counts of U.S. biotech inventions categorized by industry of application based on DWPI Manual Codes, for the 34,196 inventions with a U.S. lead inventor, including inventions assigned to multiple categories

Agriculture	17,145 (50.7%)		
Energy	840 (2.5%)	5,174 (15.3%)	
Environment	1,434 (4.2%)	1,547 (4.6%)	7,658 (22.7%)
	Agriculture	Energy	Environment

* 398 (1.2%) patent families are categorized in all three industries N=34,196

These 34,196 patent families (inventions), with geocoded location of a U.S. lead inventor, identified at the city or the zip-code level, makes up our sample of inventions. Table 6 shows how

many belong to one (or more) of the three industries for which we selected, as based on DWPI Manual Code assignments.



N=34,196 patent families with a lead inventor address in United States and identified at the city or zip-code level

Figure 5. Growth in U.S. biological inventions with applications in agriculture, energy, and environmental management

The annual count of inventions grew at an increasing rate from 1970 through 2000. After peaking in 2001, the annual number of inventions stabilized at between 1,500 and 2,000 per year. After 2006, truncation begins to affect these data (Figure 5). Factors which drive the exponential growth phase of biotech inventions in the United States include the emergence of strong intellectual property (IP) rights in biological inventions following the Supreme Court decision in Diamond v. Chakrabarty in 1980, the Bayh-Dole Act of 1980s, and the role of public-private partnership (Graff et al, 2013). These trends are consistent with earlier studies of patenting trends in agbiotechnology (Graff et al, 2003). They are also consistent with studies of invention in

biofuels, which in the U.S. grew most rapidly between 2005 and 2009, but never amounted to more than about 300 patented inventions per year (Albers et al, 2016).

4.3 Analyzing spatial distribution of inventions: Three approaches

The literature has no consensus on an appropriate analytical framework to study the dynamic nature of clusters. Most contributions to the literature discuss the contemporary (mature) functioning of clusters, but not the cumulative/developmental stages of clusters. Finding an appropriate analytical technique to study the dynamic nature of clusters remains a challenge. Yet it is important, as policies need to consider the structure as well as the dynamic nature of clusters in order to support and encourage innovation.

While scholars have long discussed the spatial concentration of innovative activities, fewer provide mechanical details of the analysis of such clusters. The literature (Lim, 2003; Usai, 2011; Wang et al, 2006; Tan et al, 2017) has described Gini Coefficient and Local Indicators of Spatial Association (LISA) techniques to determine the spatial concentration of a variable. These techniques have limitations. For example, the Gini Coefficient can estimate the degree of geographic concentration but is unable to provide information about spatial structure between local and neighboring regions (Lim, 2003). LISA depends on conceptualization of spatial regions. Optimized hotspot analysis has been a preferred solution to take care of these limitations.

Lim (2003) uses Moran I and Moran scatter plot to determine spatial concentration at the metropolitan level. Unfortunately, that study does not disclose the technical details that support the autocorrelation analysis for such a expanse as the entire the U.S. Seeing that the Midwest is almost empty, using a global autocorrelation does not make sense unless utilizing a join and relates command, but even that is still not necessarily a viable solution. Tan et al (2017) provide similar analysis for regions of China, and Moreno et al (2005), for European regions.

Following from the suggestive analyses in these studies, three approaches—mapping, Moran I, and regression analysis—are used in this chapter to explore the spatial distribution of inventions across the U.S. and its emergence over time.

4.3.1 Mapping

Arc-GIS is used to display the distribution and concentration of these inventions across the U.S. Arc-GIS requires data coordinates, shape files of cities and state to show the distribution and concentration of inventions. The regional economics, economic geography, and urban studies literatures largely use Arc-GIS to exhibit the distribution of data because of its nested commands instead of programming an algorithm. Figure 6 shows inventions from 1970 to 2010 by decade, and figure 7 shows cumulative inventions for 1970-2010. The geographic distribution of inventions during each of the four decades separately from 1970 to 2010 (Figure 6, panels a, b, c, and d) visually suggests spatial cumulativeness (Breschi, 2010), with an increasing number of new inventions in later decades where there was a concentration of previous inventions in earlier decades. The spatial distribution of all inventions over the entire period of 1970-2010 (Figure 7) suggests that invention activity was largely concentrated in more populated areas. We also observe, in contrast, we observe less intense or barren areas with few inventions, corresponding to less populated areas. This is the first time through mapping we observe of a rural-urban division of inventions.

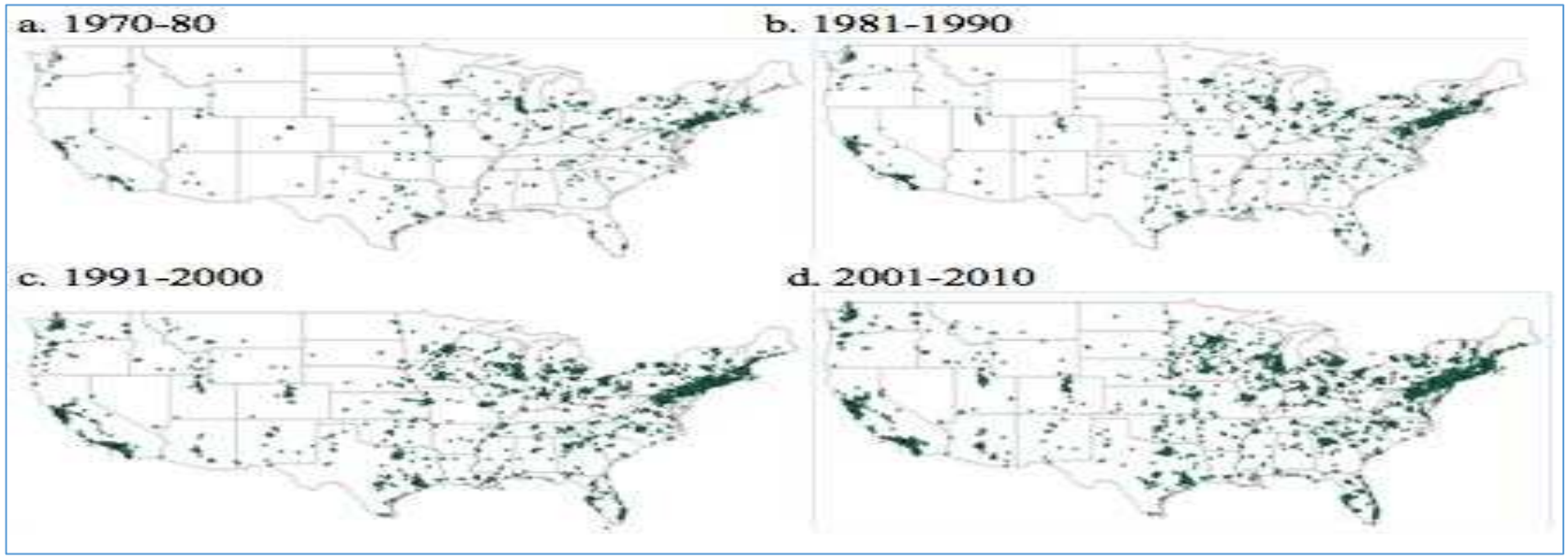


Figure 6. The spatial distribution of inventions, by decade, of the 34,196 patent families (inventions) by address of lead inventor identified at the city or the zip-code level

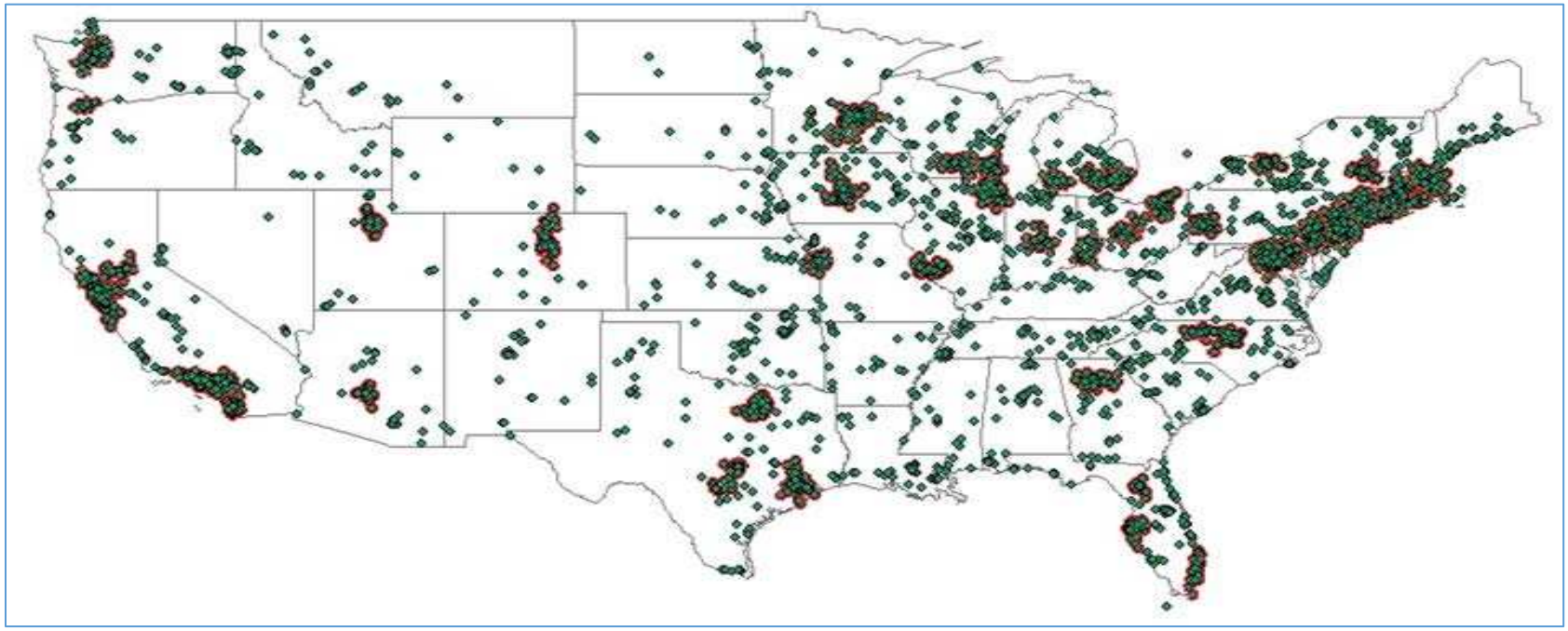


Figure 7. Spatial distribution of all 34,196 inventions, 1970-2010, by address of lead inventor identified at the city or the zip-code level, with the thirty largest clusters outlined

To preserve the integrity of the primary clusters of inventions as our units of analysis, we traced polygons in Arc-GIS around the highest density regions of mapped inventions (outlined in Figure 7). The 30 largest (and largely urban) clusters account for 58 percent of total inventions in the dataset (see Table 7). Moreover, this share has remained remarkably stable since the early 1980s, varying within just a few percentage points of this average for 30 years. The five largest clusters are the San Francisco Bay Area (incl. Silicon Valley, San Francisco, and Oakland), New York-Newark, Washington-Baltimore, San Diego, and Boston. This largely aligns with other lists of the major biotechnology clusters identified in the literature (Audretsch and Stephan, 1996; Zucker and Darby, 1996) and in industry analyses (DeVol et al, 2004). The San Francisco Bay Area is an outlier, with more than twice as many inventions as the second largest cluster, the New York city metro area.

However, there are clusters high on the list that consist of significantly smaller cities, such as Des Moines, Iowa, which ranks between Houston and Philadelphia, and Madison, Wisconsin, which ranks between Chicago and Seattle. Within our customized MSA for Des Moines is Ames, Iowa—the location of Iowa State University, the Land Grant institution for the state of Iowa—and Johnstown, Iowa—with the headquarters and an R&D center for Pioneer-DuPont, the largest hybrid corn seed company in the world and one of the most prolific applicants for gene patents overall (Graff et al, 2013). Madison, Wisconsin, is the location of University of Wisconsin, the Land Grant institution for the state of Wisconsin and one of the largest agricultural research universities in the United States. In fact, of the 30 clusters on the list, half of the regions host a Land Grant university with significant agricultural research capacities.

Table 7. The 30 largest clusters of biological inventions in agriculture, energy, and environment in the U.S., based on cumulative count of inventions 1970-2010

	Cluster name	State(s)	Count of inventions	% of total inventions	Cumulative % of inventions	Moran I Optimized Hot Spot
1	San Francisco Bay Area (incl. Silicon Valley)	CA	3,020	9.62	9.67	*
2	New York-Newark	NY, NJ	1,211	3.86	13.53	*
3	Washington-Baltimore	DC, MD, VA	1,187	3.78	17.31	*
4	San Diego	CA	1,135	3.62	20.92	*
5	Boston	MA	1,017	3.24	24.16	*
6	Los Angeles	CA	939	2.99	27.15	*
7	Houston	TX	924	2.94	30.10	*
8	Des Moines-Ames	IA	882	2.81	32.91	*
9	Philadelphia-Camden-Wilmington-Trenton	PA, NJ, DE	811	2.58	35.49	*
10	Chicago	IL	713	2.27	37.76	*
11	Madison	WI	544	1.73	39.50	*
12	Seattle	WA	516	1.64	41.14	*
13	Raleigh-Durham-Chapel Hill	NC	449	1.43	42.57	*
14	Cleveland-Akron-Canton	OH	420	1.34	43.91	
15	Detroit-Ann Arbor-Lansing	MI	420	1.34	45.24	
16	St Louis-Columbia	MO, IL	415	1.32	46.57	
17	Minneapolis-St Paul	MN	409	1.30	47.87	*
18	Denver-Boulder-Ft Collins	CO	376	1.20	49.07	
19	Sacramento-Davis-Woodland	CA	344	1.10	50.16	
20	Atlanta-Athens	GA	296	0.94	51.11	
21	Cincinnati	OH	270	0.86	51.97	
22	Indianapolis	IN	255	0.81	52.78	*
23	Dallas	TX	233	0.74	53.52	
24	Salt Lake City	UT	227	0.72	54.24	
25	Portland region-Corvallis-Eugene	OR, WA	183	0.58	55.51	
26	Rochester	NY	172	0.55	56.06	
27	Princeton-New Brunswick	NJ	170	0.54	56.60	
28	Kalamazoo	MI	167	0.53	57.13	
29	Omaha-Lincoln	NE	149	0.47	57.61	
30	New Orleans-Baton Rouge	LA	145	0.45	58.06	

4.3.2 Moran I and optimized hotspots

The Moran I measures spatial autocorrelation between the observed values in a specific location and spatially weighted average values of another location (Lim 2003). To understand the spatial patterns in the data, the Moran I index evaluates whether the data are clustered, dispersed, or random (ArcGIs Pro). The Moran I index ranges from +1 which indicates positive autocorrelation to -1 which indicates negative autocorrelation. The P values is used to accept or reject the hypotheses.

$$\text{Moran I} = \frac{N}{\sum_i \sum_j Wij} \frac{\sum_i \sum_j Wij\left[Xi - \overline{X}\right]\left[Xi - \overline{X}\right]}{\sum i\left[Xi - \overline{X}\right]^2}$$

where N is the total number of observations, w_{ij} represents elements of spatial weight matrix between the two regions "i" and "j." Similarly, $[X_j - \text{Mean}(X_j)]$ and $[X_j - \text{Mean}(X_j)]$ represent deviation from the mean at "i" & "j" and time "t" respectively. When a row standardization is applied, the first part of the equation becomes equal to 1.

Moran I values (see table 8) accept the presence of a spatial trend in the inventions. The P-values show the probability of a spatial pattern, and Z-values represent the standard deviations. For high significance, positive Z-values are associated with low P-values. We calculate Moran I for each ten years following a similar sequence with that in the four panels of Figure 6. We also calculate Moran I over the entire period 1970-2010 corresponding to Figure 7. Moran I of 1981-1990 is not significant even though its Moran I value is positive. The rest of the data show significantly positive Moran I values. Positive Moran I values represent positive autocorrelation, meaning areas (i.e. cities) with high levels of inventions cluster together.

Optimized clusters analysis uses the Moran I statistics which has been widely used to measure spatial autocorrelation (Fischer, 2006; Wang et al, 2016; Tan et al, 2017). The concept of "optimized" clusters refers to regions that have a higher observed concentration of inventions

compared to a random distribution of inventions. In other words, optimized clusters compare invention density in a specific region with a complete spatial randomness. We analyze optimized clusters in this dataset in order to validate the previous identification of clusters.

Table 8. Moran I of inventions across U.S. regions

Years	Moran I	P-Value	Z-Value
1970-1980	0.1979	0.0000*	8.3377
1981-1990	0.0207	0.2305	1.1880
1991-2000	0.0423	0.0014*	3.1832
2001-2010	0.0599	0.0000*	5.5758
1970-2010	0.0574	0.0000*	5.5758

*P values significant at < 1%.

The main advantage of optimized cluster/hotspot analysis is to choose the scale of analysis, appropriate distance band, standardization based on questioning the data to yield optimal clusters/hotspots. Optimized clusters analysis help to understand whether the spatial patterns are significant or not. It not only shows a more concentrated region, less concentrated, and no significance among the polygons. Similarly, optimized cluster analysis adjusts automatically for multiple testing, spatial dependence running False Discovery Rate (FDR) correction method[6].

[6] http://pro.arcgis.com/en/pro-app/tool-reference/spatial-statistics/how-optimized-hot-spot-analysis-works.htm

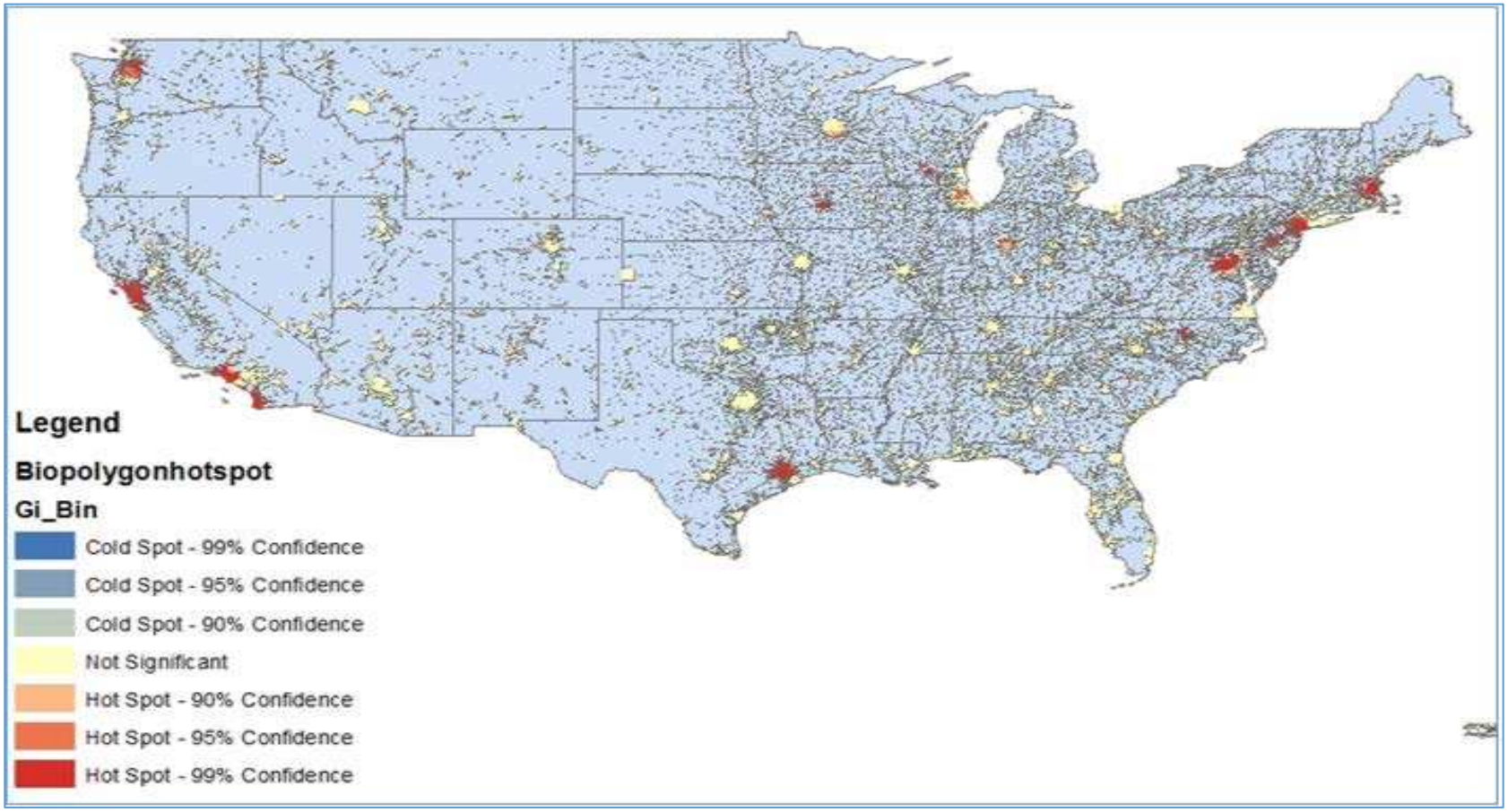


Figure 8. Optimized Clusters or Hotspots

Figure 8 shows the optimized clusters identified in this analysis. We see spatial concentration of inventions in major metropolitan areas. Hotspots are represented with a significance level of 99%, which identify highly innovative areas that are contiguous with other relatively innovative neighboring areas. Cold spots are defined as lone highly innovative areas not surrounded by other highly innovative areas. In figure 8 we do not observe cold spots. This indicates that most of the highly innovative regions are surrounded by other relatively innovative regions. Similarly, many regions are labelled as "not significant" which indicates neighboring areas with similar innovation levels, including some of the emerging areas for these industries, as described in Table 6 of the 30 largest clusters. In Table 6 the "*" in the last column marks those regions in the U.S. that are determined to be optimized hotspots in this analysis. Not surprisingly the regions identified as Moran I optimized hotspots congregate high on the list of largest clusters as identified by the mapping.

One important distinction is that it is not necessary that hotspots are those areas with the highest values. It is possible that an individual polygon (city/metro area) may have a high level of innovation but may not be calculated as a hotspot if it does not have relatively high levels of innovation surrounding it. In this case the neighborhood (city/metro area) is not different from the study area (city/metro area). It also possible to have low value e.g. "0" and the neighborhood has sufficiently high values to bring the index value up enough to get a hotspot in a polygon where there is a low level of innovation. Generally, hotspot analysis helps to understand whether spatial patterns are significant or not. The basic purpose of hotspot analysis is to determine patterns, to identify causes, and to predict future trends (ESRI, 2014).

Optimized clusters/hotspot analysis constructs contiguity (neighbor) and weights matrices to identify and assign weights to the neighborhoods of the regions. First the contiguity matrix is

developed, and then the weight matrix is constructed based on the contiguity matrix. Construction of these two matrices is necessary to choose an appropriate distance band and conceptualization of spatial relationships. The following three considerations are important while running optimized hotspot routines.

Conceptualization of spatial relationships: The features (polygons) have shared commonalities, and the interactions among the features are important to understand. Therefore, the analysis must choose among different conceptualizations of spatial relationships: inverse distance, inverse distance squared, fixed distance band, zone of indifference, polygon contiguity, or assignment of spatial weights. Each of these conceptualizations has pros and cons. The selection can become problematic especially when there is a high degree of heterogeneity of polygon sizes.

How to choose an appropriate distance band: The appropriate distance band is based on the study area and scale of analysis (country, state, Metropolitan Statistical Area, Metropolitan Statistical Area, Combined Statistical Area). There is no right or wrong distance band, but selection depends on the several processes. Different distance bands match different scales of analysis. The literature has proposed rules to choose an appropriate distance band. For example, the "Z" score is derived from spatial autocorrelation to select an appropriate distance band. Ultimately, the underlying research question and the scale of analysis inform the decision.

Standardization: It is also important to consider whether to assign equal weights to neighbors, or not. If not standardized, the analysis assigns different weights to the neighbors. This means that neighbors with larger values will affect those with smaller values. The literature generally prefers row standardization in order to assign equal weights.

4.3.3: Regression analysis

4.3.3.1 Identification of regions for statistical analysis

Acs, Anselin, and Varga (2002) raise the question regarding the proper unit of analysis for innovation systems. To explore the urban-versus-rural nature of biological innovation for agricultural and natural resource industries, ideally we wish to identify each relevant geographic region that serves as a contiguous home to such innovations, or what many describe more loosely as a "cluster." We then want to compare those regions with clusters with similarly sized geographic regions that exhibit variation in degrees of innovation, including regions even that have no evidence of having produced patented inventions.

All the regions included in the analysis must have available data on associated explanatory factors or "covariates" to the observed innovation. While, our patent family data are consistently denominated at the geographical level of the city, we find that availability of data for associated explanatory factors or "covariates" is available for the entire period of 1970 to 2010 for the 929 metropolitan and micropolitan statistical areas (MSAs and μSAs) in the United States, as delineated by the United States Office of Management and Budget (OMB). MSAs primarily represent urban areas, and μSAs, relatively rural areas. MSAs consist of a core county or set of adjacent counties in which lies an urban area having a population of at least 50,000. MicroSAs consist of a core county, having a population of 10,000 to 50,000, with possible adjacent counties. The adjacent counties and core counties have a high degree of social and economic integration through economic flows and commuting ties (United States Bureau of Economic Analysis, 2018).

To preserve the integrity of the primary clusters of inventions as our units of analysis, we traced polygons in Arc-GIS around the highest density regions of mapped inventions (outlined in Figure 7). We then compared these traced polygons to the boundaries of MSAs and found that 20

of these invention clusters spanned more than one MSA. For each of these, we combined the two or more MSAs that encompassed, as closely as possible, the high-density portions of the observed invention clusters to create a custom statistical area. These combinations reduced the 929 official MSA and μSAs to 897 statistical areas, consisting of our 20 custom statistical areas together with 877 remaining unmodified MSAs and μSAs.

Source: United States Census Bureau

Figure 9. The geographic coverage of Metropolitan Statistical Areas (MSAs) and Micropolitan Statistical Areas (μSAs)

Two general approaches are taken to analyze the patterns of invention as well as factors associated those patterns. First, we seek to test hypotheses of cumulativeness in inventions within our identified regions. We have visually noted an apparent tendency for inventions to accumulate in specific regions based upon our preliminary mapping. We now seek stronger systemic evidence that inventions are indeed concentrated in those areas. Second, we seek systematic evidence of regional urban characteristics versus rural or agricultural characteristics being associated with higher levels of invention in these technologies.

We can analyze the cumulative nature of inventions by region with equations

$$\text{Inventions}_{it} = \beta_i + \beta_1 \text{Invenitonlag1}_{it} + \beta_2 \text{Invenitonlag2}_{it} + \cdots + \beta_6 \text{Inventionlag6}_{it} + \mu_{it}$$

where the Invenitonlag variables represent the counts of inventions at different respective lags, and

$$\text{Inventions}_{it} = \beta_i + \beta_1 \text{CumulativeInv}_{it}$$

where CumulativeInv is the cumulative sum of prior inventions at time t in statistical region i.

4.3.3.2 Cluster growth

To analyze spatial cumulativeness of biotechnology inventions for agriculture and resource applications, we test how the presence (or absence) of inventions within a given region affect the probability of subsequent inventions arising in that region. We regress invention counts on lags of previous invention counts for each region in each year (Table 9). The time series optimal lag length criteria (AIC/BIC) is not appropriate for these panel estimation techniques, but an optimal lag length in panel data can be determined manually by starting from a lag of 1 year, then 2 years, and so on, stopping when the coefficient of the lagged explanatory variable becomes negative.

Table 9. Fixed effects regression of lagged invention counts

Variables	Coef.	St. Err	t	p>|t|
Inventions 1st lag	0.7218	0.0053	135.82	0.0000
Inventions 2nd lag	0.1946	0.0064	29.98	0.0000
Inventions 3rd lag	0.0508	0.0065	7.74	0.0000
Inventions 4th lag	0.0942	0.0065	14.37	0.0000
Inventions 5th lag	-0.0807	0.0064	-19.77	0.0000
Inventions 6th lag	-0.0807	0.0052	-15.46	0.0000
Constant	0.1079	0.0106	10.09	0.0000
F(896,35868) = 1.62 Prob > F = 0.0000				

In addition, to validate the overall significance of a region's previous invention activity on current inventions, we construct a cumulative prior invention count variable, defined as the sum

of inventions from year 0 to year t-1. We regress current year invention counts on the cumulative sum of prior inventions for each region for each year (Table 10).

Table 10. Fixed effects regression of cumulative invention counts

Variables	Coef.	St. Err	t	p>|t|
Cumulative Inventions	0.5232	0.0002	178.71	0.0000
Constant	-848.2288	4.7504	-178.56	0.0000
F(896,35879) = 103.74 Prob > F = 0.0000				

We find that lagged invention counts and the cumulative sum of prior inventions show positive and significant effects on current invention counts. Lagged invention counts have a significant relationship for up to four years (i.e., inventions in year t are positively related to inventions in years t-1, t-2, t-3, and t-4). More recent past activity has greater ability to explain current rate of inventions: as the lag increases beyond four years, the effect disappears. Yet, the relationship between the cumulative sum of past inventions and current inventions is also positive and significant, confirming that these biological inventions for agriculture and natural resource applications exhibit spatial cumulativeness and therefore remain relatively concentrated spatially.

4.4 Discussion and Conclusion

The analysis in this chapter has used a unique dataset of biological inventions, identified by inventor addresses in patent data, to answer questions regarding how biological inventions for use in primary resource-intensive industries—such as agriculture, energy, and natural resources—have been spatially distributed across the United States. Our preliminary mapping indicates relative concentration as well as spatial cumulativeness of innovation for these industries in urban areas, with new inventions in later decades occurring where there was already a concentration of inventions in previous decades. And the visible concentrations align with both major and minor metropolitan areas across the map of the United States.

By tracing polygons around the geographic footprint in ArcGIS of inventions in the 30 largest clusters and then ascertaining the counts of inventions contained within each, we confirm that 56 percent of the inventions in the dataset were made in just these 30 largely urban regions, which correspond to the primary biotech clusters identified in other studies, including most of the very largest cities in the United States, including San Francisco Bay Area, New York, Boston, and San Diego, along with Los Angeles, Chicago, and Houston. However, we also secondary urban areas located near areas of high agricultural production, and we note that half of the 30 largest clusters include a Land Grant university with significant agricultural research capacity.

We also explore the space-time dynamics of biological inventions for these industries, such as cumulativeness, the extent to which previous inventions in a given region increase the probability of new inventions arising in that region. From negative binomial panel regressions of lagged invention counts and the cumulative sum of prior inventions, we find positive and significant relationship between past numbers of inventions and the numbers of new inventions by region. These biological inventions for agriculture and natural resource applications exhibit spatial cumulativeness, remaining relatively concentrated spatially over time.

Based on these insights, this study contributes to the literature by showing how biological inventions intended for predominately rural industries are distributed for the period 1970-2010. It finds them spatially clustered largely within metropolitan areas. This study shows how the concentration of innovative activities within regions is spatially correlated with the concentration of innovative activities of neighboring regions, using an appropriate regional cluster analysis technique called optimized hotspots. It also examines how these clusters develop over of time.

To summarize, the main results are:

1. The spatial distribution of biological inventions spans much of the U.S, but a spatial clustering pattern clearly exists.
2. In terms of concentration of biological inventions, a rural-urban division exists. However, the inventions are not concentrated in rural areas near agricultural or natural resource production but rather in urban regions.
3. The number of inventions in an area in prior years has significant explanatory power for the number of inventions in any current year. This relationship represents a localized spillover phenomenon.
4. The neighborhood innovation potential over time affects a given region's innovation. This means we can see an increasingly clustered space-time relationship.
5. The non-significant areas e.g. the areas not surrounded by highly innovative neighbors also are found to persist across the U.S.
6. While we do see some inventions in rural areas, rural areas do not appear to be the hotspots of innovation in agricultural, energy, or environmental biotechnologies.

These findings identify and quantify the most significant regional clusters of biological invention in the United States. Yet, this chapter is mainly exploratory and does not address several important empirical questions. The following chapter further develops the empirical analysis to better understand what drives the distribution of agricultural and resource industry inventions within metropolitan regions.

5. EXPLORING THE COVARIATES OF REGIONAL CONCENTRATION OF BIOLOGICAL INVENTIONS FOR AGRICULTURE, ENERGY, AND ENVIRONMENT IN THE UNITED STATES

5.1 Introduction

Extensive theoretical and empirical work has explained the general mechanisms of agglomeration in production, innovation, and entrepreneurial processes. While some significant differences in views persist and major issues remain, what emerges is that several related factors are important in driving the clustering of innovative activity: exchange of ideas, availability of a skilled labor pool, input-output linkages, population density, and a decentralized and cooperative culture. Moreover, these factors seem to complement each other. The cluster is the socio-economic/geographic unit within which these factors interact and have a conducive impact on economic growth. Input-output linkages facilitate the discovery of and interaction among suppliers and customers. These interactions help exchange of ideas or technology spillovers. Population density encourages knowledge spillovers. If population is more educated and skillful then the flow of information and ideas is more fruitful. What results is a virtuous cycling, circumscribed in space.

To the extent that the causes and virtues of agglomeration hold true, agriculture and other geographically diffused industries are stuck in something of a dilemma. Innovation activities that arise from production activities, whether described as learning-by-doing (Arrow, 1971) or user-led innovation (Von Hippel, 1988), are necessarily linked to a resource base and thereby a skilled labor pool which, for these industries, is geographically diffused. Innovators in the field, as it were, cannot easily benefit from the virtuous cycling of knowledge spillovers that occurs within a cluster, which naturally gravitates to the high population density of urban centers. Conversely, when

innovations that are potentially useful for agriculture do arise within the vortex of an urban innovation cluster, they are handicapped by virtue of being distant from the community of producer-practitioners of skilled labor that otherwise would contribute to development, iteration, and refinement. The urban-based innovators for agriculture are also less connected through input-output linkages and thus less routinely engaged with suppliers and buyers in idea exchange, in the fortuitous recombination of existing ideas, and discovery though experimentation.

While we cannot resolve this dilemma in a single analysis, we can begin to shed some light on it through empirical analysis of innovation patterns in one key area of technology, one that seems, in fact, to accentuate the features of this apparent dilemma. Innovations in biotechnology and genetic resources for agriculture, biotechnology and biorefining for energy, and biological methods for environmental monitoring and remediation all have vast geographic scope of potential utilization. Yet, the innovation clusters of the biotechnology industry in general have, historically, been highly concentrated. In the previous chapter we establish that spatial proximity does appear to play a role in such inventions and that prior inventions in a region increase the probability of new inventions. What other factors appear to be associated with the growth of such innovation clusters?

This study departs from most of the literature on innovation systems which quantify innovation and its factors at the national or even the state level. In this analysis, distribution and concentration of innovation is measured at the more granular zip code and city level. In other studies, when authors compare the spatial distribution of inventions between states, such as California and Colorado, the result may be displayed as a dark color for the whole state of California to show a high number of inventions, and light color for the whole state of Colorado to show a lower number of inventions. In contrast, this study considers the heterogeneity of actual

inventions within and across the regions of California and Colorado. Thus, it becomes evident that while the metro region around Denver, Colorado, has fewer inventions than the regions around San Francisco, San Diego, or Los Angeles, it has more inventions than the metro region around Sacramento as well as all other regions in California.

This chapter uses the inventor address data from patent filings in the InSTePP Global Genetics Patent Database to count inventions annually for 1970-2010 in each the 897 custom statistical areas in the United States created in the previous chapter. We draw upon data for associated explanatory factors or "covariates" for the entire period of 1970 to 2010 for the 929 metropolitan and micropolitan statistical areas (MSAs and μSAs) from the U.S. Bureau of Economic Analysis, such as population, earnings by place of work, and farm proprietor income, after checking for multi-collinearity to assure their relative independence.

This chapter is organized as follows. The next section reviews the literature on factors that explain the clustering of innovation activity. Section 3 describes the data we compile, and section 4 outlines the empirical methodology for analysis of covariates of invention counts for the set of 897 clusters identified in the preceding chapter. Section 5 presents results of a panel estimation, and section 6 extends this to a spatial panel estimation. Discussion and conclusions are offered in section 7.

5.2 Literature: What causes the clustering of innovation?

Clusters arise as firms decide to locate within proximity of one another. In a comparative analysis of high technology clusters, Saxenian (1996) seeks to figure out what factors are responsible to form a flourishing cluster. She analyses why is it that high technology business in California's Silicon Valley flourished while along Route 128 in Massachusetts they declined in the 1990s. Saxenian explains that, despite similar histories and technologies, Silicon Valley

developed a decentralized but cooperative industrial system while Route 128 came to be dominated by hierarchical, self-sufficient corporations.

Empirical studies of firm location behavior have considered a complex mix of factors, including transport cost, local factor prices, production and substitution possibilities, market structure, and competition (McCann, 2013). Scholars have investigated which of these factors is the dominant influencer under certain situations. Yet, without being able to control for industry and technology, systematic conclusions cannot be drawn about optimal firm location behavior resulting in industrial clustering or, conversely, industrial dispersion (McCann 2013). Beyond these typical cost and industrial organizational factors, other less tangible factors may be at least as important in determining firm behavior, including knowledge or information spillovers, local non-traded inputs, and the local skilled labor pool.

When firms of the same industry get close to each other, the employees of one particular firm have easy access to the employees of another firm. This easy access can be through face-to-face meetings, sharing lunch time, and other social activities. In these meeting, the employees share the tacit information, discuss new technology development and market trends. Information sharing among the employees give them an edge to compete in the market. McCann elaborates with the financial industry examples of Wall Street, New York, The City of London, and the Marunouchi district of Tokyo. These are financial hubs, where the flow of information can change in a minute. To keep track of changing information within minutes and take decisions on that basis, immediate access to the other market participants is essential.

Another important factor that encourages firms to cluster together is sharing non-traded inputs including access to specialized local infrastructure. The City of London and Wall Street have many specialist legal and software firms that provide services only to financial institutions.

These services are very expensive, but when there are many firms in the same locality, the average cost decreases. Similarly, in the City of London, there is a dedicated broadband fiber optic cable system. The firms who want to access this facility must be in the designated locality, otherwise, the firms cannot avail this opportunity. As the number of the firms that do avail this opportunity increases, the average cost decreases (McCann 2013).

The other source of attraction is the availability of a skilled labor pool. The easy availability of a skilled labor pool reduces the firm's acquisition cost. Glaeser and Resseger (2010) examined the complementarity between cities and skills. To prove their hypotheses, first they divide agglomeration theories into two groups: Those that consider the spread of knowledge in cities, and those do not. Those that do not support the importance of knowledge flow are of the view that good governance in cities, easy access to ports or harbors, and the possibility of easy capital access are the main drivers of agglomeration. Glaeser and Resseger show a strong connection between per worker productivity and metropolitan area population. The result is stronger for cities with higher levels of skill and almost non-existent in less-skilled smaller metropolitan areas. This suggests that urban density is important because proximity spreads knowledge, that makes a worker more productive or an entrepreneur more successful. Their results suggest a strong complementarity between skills, city size, and learning. Agglomeration effects are stronger for cities with more skills.

Delgado et al. (2010) analyze the role of regional clusters in entrepreneurship. The presence of complementary economic activity creates externalities that enhance incentives and reduce barriers for new business creation. Clusters are identified as the best mechanism through which location-based complementarities are realized and entrepreneurship as the best channel through which cluster-driven agglomeration operates. The authors consider entrepreneurship an important

factor for innovation. The paper concludes that clusters have a significantly positive impact on entrepreneurship.

Audretsch et al (1996) explore the mechanisms through which clusters influence entrepreneurship, and entrepreneurship influences economic growth. They develop what they call the Knowledge Spillover Theory of Entrepreneurship (KSTE) in which knowledge spillovers serve as a source of entrepreneurial opportunities, generating additional innovative outputs from any given amount of investment in knowledge-generating inputs, such as R&D expenditures. Their KSTE theory seeks to explain endogenous entrepreneurship, growth, localization, entrepreneurial performance, and entrepreneurial access. The theory draws upon economic growth models, namely Solow and Romer. The Solow model considers capital and labor as the main factors of production and assumes technical change accounts for the unexplained residual in growth accounting regressions. The Romer model considers knowledge as an important input in the basic neoclassical production function. The entrepreneurial growth model focuses on knowledge spillovers and their commercialization. As such, the entrepreneurial economy exemplifies creative construction rather than "creative destruction" (Schumpeter, 1942).

Entrepreneurship works as a conduit for knowledge spillovers. It links the investment in knowledge and economic growth. Commercialization of new ideas provides an international comparative advantage to the developed countries. The localization hypothesis posits that when knowledge spillovers involve tacit knowledge or less codified knowledge, geographic proximity matters and that shapes the location of entrepreneurial firms. Geographic proximity provides comparative advantage.

Still, knowledge spillovers need to be effectively accessed and absorbed. The authors highlight two important factors by which entrepreneurs access and absorb external knowledge: a

spillover conduit (such as members of a board of directors, managers) and close geographic proximity. Finally, for the better entrepreneurial performance the KSTE concludes in favor of venture capital rather traditional bank-based financing.

To make the complementary factors productive in terms of entrepreneurship, policy makers should understand the role of the regional innovation system within the context of the national innovation system. Regional and national innovation systems should be in line with federal trade policies and the nation patent system. The cooperative and decentralized culture that exists in places like Silicon Valley can lead to increasing returns and economic growth.

To outline a complete mechanism that explains how these complementarities exhibit path dependency and the growth of a cluster is complex. We can start by examining just a handful of leading factors to have a deeper understanding of that path dependency (Figure 10). For example, Rauch (2014) studies only cities as spatial clusters. Glaeser and Resseger only study measures of skill within cities. Bauernschuster et al. (2010) study social capital access. Out of six fundamental factors (see Figure 10), we can discuss pairwise, for example, population density and skilled labor force, skilled labor force and exchanges of ideas, and so on. The order of these relationships does not matter.

The simplest mechanism would involve skilled labor that meet at a particular place and exchanging ideas. If the skilled labor pool were not present, then exchanges of the ideas would not take place. We should not ignore the significance of population density, which is an important facilitator of those meetings. Once the exchange of ideas starts within the skilled labor pool, input-out linkages become more important. Input-output linkages explain the interconnectedness of industries and technologies, as customers and suppliers. Finally, exchange of ideas, skilled labor pool, input-output linkages, and population density come to characterize a decentralized and

cooperative system. Figure 10 illustrates such path dependence, but while the events in the path matter, their order may not.

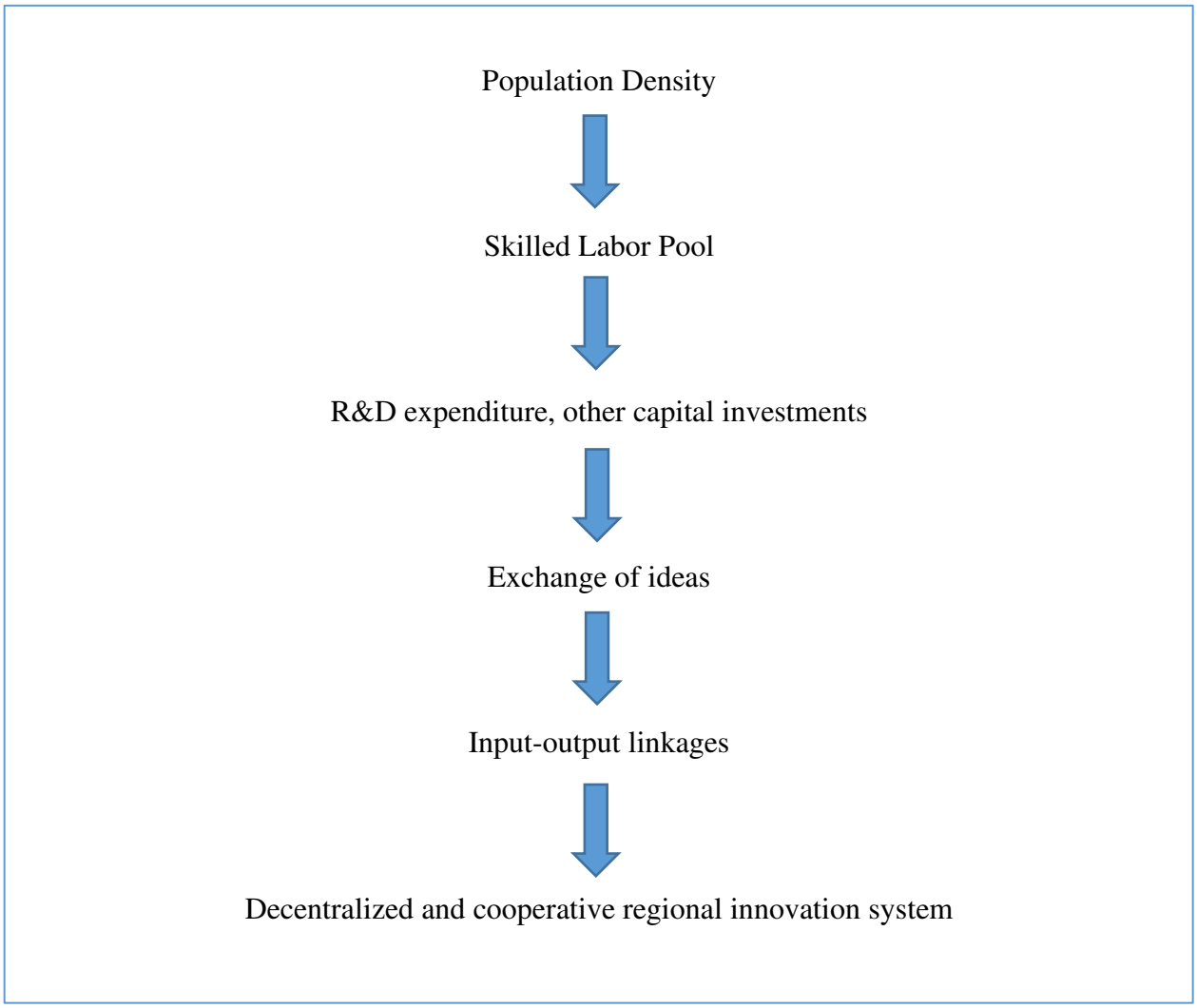


Figure 10. Path dependency in the growth of an innovation cluster

Urban and regional economists have mostly discussed the positive side of these complementarities. Rarely we can investigate how these complementarities affect labor productivity, the wage structure, and increase in the prices of the amenities, which causes increase in spatial inequality. Labor markets are less flexible and have limited mobility. Acs (2001) discusses how the skilled labor of Route 128 was affected by the flourishing of Silicon Valley. The effect of labor migration was not only felt in the computer and IT industries, but even more

so in other related industries. Those local firms that were dependent on serving the large computer and IT firms on Route 128 were severely affected. On one hand the complementary factors developed Silicon Valley, but, on the other hand, they created spatial inequality vis-à-vis Route 128. The influx of the large corporation to Silicon Valley also affected those industries that were following the Route 128 model of centralized hierarchies and non-cooperativeness. We can see the industrial clusters in the east and west coast of the USA. The Midwest has fewer clusters and the spatial inequality of the Midwest in the USA is apparent. Industrial clusters in Germany are in the northwest and southwest. The spatial inequality compared to eastern Germany is obvious. In China, the industrial clusters are dominated in the east, and the spatial inequality with the west of China is discernible.

In summary, this literature builds upon the key theories of Marshall (1920) and Krugman (1991). These theories and the literature consider the following factors are most important in deriving the clustering of economic activity: Exchange of ideas, availability of the skilled labor pool, input output linkages, population density, R&D expenditure decentralize and cooperative system. These factors complement each other. If the labor pool is less skilled or is unskilled, then exchanges of ideas may not complement the labor pool, as is the case of a more skilled labor pool. Input-output linkages facilitate the suppliers understanding and meeting consumer demand. Exchanges of ideas or spillovers of information help input-output linkages to flourish. Population density and knowledge spillovers have a positive relationship. If population is more education and skilled then the flow of information is fruitful, otherwise, the dense population might not benefit. These complementarities on one hand can lead to increasing returns and economic growth, but on the other hand can lead to spatial inequality.

5.3 Data

To address the questions of what factors are most important in the formation of the observed innovation clusters in the life sciences being applied in agricultural and natural resource industries, we assemble two types of data—data on inventions and data that provide at least proxies for factors associated with clustering—organized into relevant geographic units for analysis of regional innovation. Table 11 provides summary statistics of inventions, population, earnings, and farm income for 897 statistical areas[7] across the United States for the entire time period of 1970-2010, for a total of 36,777 observations for each measure. For this empirical analysis, we draw upon data from the Bureau of Economic Analysis on MSAs and μSAs. After reviewing a range of possible measures, we select the variables of population, earnings by place of work, and farm proprietor income. Evaluation of multi-collinearity assures their relative independence, and each is available for all statistical areas for the entire time span. The geographic coverage of the 897 statistical areas included in the analysis is illustrated in Figure 9. The remaining rural areas are not included in this analysis, as the data did not extend to the remote regions outside of the MSAs and μSAs.

- Population is fundamentally a size variable. The number of people relates to both the overall level of economic activity (and is thus highly correlated with regional gross product) as well as the size of the labor pool, including the skilled. We have seen from the literature that the size of highly skilled human capital pool is highly correlated with population size. Because the geographic size of the regions is, if anything smaller in urban areas (see Figure 9) higher population also indicates higher population density, another factor implicated in theories on

[7] There are 929 official MSA and μSAs for the United States. In the previous chapter, we found that 20 of the observed invention clusters spanned more than one MSA. In these cases, we combined two or more MSAs to create a custom statistical area. These combinations reduced our number to 897 statistical areas, consisting of our 20 custom statistical areas together with 877 remaining unmodified MSAs and μSAs.

innovation clustering (Glaeser and Resseger, 2010). Our hypothesis is that inventions are positively related to population.

- Earning by place of work includes wages and salaries together with supplements to wages and salaries. We include it as our measure of relative level of economic development or economic activity as well as the quality of the workforce, as highly trained scientists and engineers will be expected to earn more than low skilled labor. We expect rates of invention to be positively related to regional earnings.
- Farm proprietor income counts the net income (receipts net of expenses) for sole proprietor and partnership farms, which make up over 90 percent of agricultural operations in the United States. We observe that farm income is, not surprisingly, very small in major metropolitan areas. The MSAs surrounding smaller urban centers as well as many of the more rural μSAs have significant farm incomes. On the one hand, we expect that, as a measure of relative rural regions, farm income is likely to be negatively related to the number of inventions. However, among similarly sized regions that do have some inventive activity, we expect that those with higher farm income will have more linkages and spillovers within the industry and therefore more opportunity for inventions.

Table 11. Summary statistics of inventions and selected clustering covariates, 1970-2010

Variables	Obs.	Groups	Mean	Std. Dev	Min	Max
Inventions (count)	36,777	897	0.7316	5.8479	0	279
Population (million)	36,777	897	0.2618	1.0162	0	20.52
Earnings by Place of Work (million)	36,777	897	4.4463	24.4164	0	851.03
Farm Proprietor Income (million)	36,777	897	0.0206	0.0472	-0.12	0.98

Table 12 shows correlation coefficients for the underlying variables. The correlation values show positive relationship of inventions and explanatory variables (earning by place of work, farm proprietary income, and population). Earning by place of work is relatively much more related to invention followed by population, and farm proprietary income.

Table 12. Correlation matrix

	Inventions	Earnings by Place of Work	Farm Proprietary Income	Population
Inventions	1.00			
Earnings by Place of Work	0.52	1.00		
Farm Proprietary Income	0.20	0.30	1.00	
Population	0.43	0.88	0.36	1.00

Correlations of the explanatory variables are relatively low related to the dependent variable (invention), and they are low relative to each other, except between population and earnings by place of work (0.88). Increase in population size is related to earnings by place of work. The correlation analysis helps to understand the regression coefficients' magnitudes (values). For example, a large coefficient value may be due to its high correlation. Similarly, the correlation values show us the relationships among the explanatory variables. If two explanatory variables are highly correlated (multicollinear), then one should be dropped from the analysis, if it is not justified intuitively. We cannot drop population, which is highly correlated with earnings by place of work, because of its importance in measuring size and how it shows the urban concentration mechanism.

5.4 Methods

The dependent variable (inventions) is count data, consisting of non-negative integers values like {0,1,2,3,…279}. The literature has widely discussed the use of Poisson and negative binomial maximum likelihood regression models for this type of data (Hausman, Hall, and Griliches, 1984). The difference between the two is in their distribution functions, where the Poisson distribution assumes the mean and variance for the data are equal (equal dispersion property), and the negative binomial distribution allows for over-dispersion in the data. The negative binomial distribution has one more parameter than the Poison distribution, a dispersion parameter to adjust variance to mean. The Poisson distribution a special case of the negative binomial distribution, where the mean and variance of is constrained to be equal. Intuitively the nature of invention data support Poisson model. We have many instances of the number "0" in the invention data, and it is because of its nature: for many of the Metropolitan and Micropolitan Statistical Areas there are no inventions in a given year. We can see in the descriptive statistics (Figure 5) that until 1990 there were very low numbers of inventions overall. In later years we can see the rise in inventions.

The Poisson probability function is given as

$$f\left(Y_i \middle| \theta\right) = \begin{cases} \dfrac{\theta^{Y_i} e^{-\theta}}{Y_i!}; & Y = 0,1,2,\ldots \\ 0 \quad\quad ; & Otherwise \end{cases}$$

The Poisson equal dispersion property means $E\left(Y_i\right) = \theta$ and var (Y_i) = θ. The mean and the variance of the data must remain equal. If this is the case, then the Poisson model can be used.

If the mean and standard deviation of the data are not equal, then the negative bionomial model can be used to allow for the over-dispersion of the data. The general form of the negative bionomial probability distribution is

$$f\left(Y_i|\theta\right)=\begin{cases}\left(\dfrac{r}{r+\theta}\right)^r\dfrac{\Gamma\left(r+Y_i\right)}{Y_i!\Gamma\left(r\right)}\left(\dfrac{\theta}{r+\theta}\right)^{Y_i} & ;\ Y_i=0,1,2,\ldots \\ 0 & ;\ Otherwise\end{cases}$$

For the negative bionomial probability function $E\left(Y_i\right)=\theta$ and var (Y_i) = $\theta + \frac{\theta^2}{r}$, where r is the dispersion parameter.

The log-likelihood equation based on the Poisson distribution for probability of invention (Y_i) at a particular point in space and time is

$$l(\lambda;U)=\sum_{i=1}^{n}\log\lambda\,(s_it_i)-\iint_{A\,0}^{T}\lambda\,(s,t)dtds-\log(n!)$$

Where U is a 987 x 41 matrix with rows containing regions s_i and years t_i (1970, 1971, … ,2010). A is the two-dimensional study area, and 0-T is the time period for these observations. The above equation becomes

$$l(\lambda;U)=\sum_{i=1}^{34196}\log\lambda\,(SAi,1970)-\iint_{U.S.1970}^{2010}\lambda\,(SA)dtds-\log(34196!)$$

The integrated intensity function $\lambda\,(s,t)$ indicates we are only interested in areas in the USA from 1970 to 2010:

$$\overline{\lambda}=\iint_{USA\ 1970}^{2010}\lambda\,(s,t)dtds$$

This equation describes points that occur at a particular area at a particular time. We can derive the integrated intensity function from equation 3. In this model SA is the two-dimensional statistical area. This sub-sampling is represented by β in equation 2, where s_i and t_i represent the space and time dimensions. The s_i shows the invention distribution at a given SA, and t_i shows the

invention distribution at a specific time. The total number of statistical areas are 897 and total inventions are 34,196.

We also assume

$$\lambda(s,t) = x(s,t)'\beta$$

where $x(s,t)'$ is a P x 1 vector having covariates at a specific location at a specific time within the study area, and β is P x 1 vector of regression coefficients.

The right-hand side of this equation is $x(s,t)'\beta$ is given at equation (5). Before estimating, the general form of fixed and random effect equations are:

The general form of fixed effect model is given as

$$Y_{it} = \beta_{1i} + \beta_2 X_{2it} + \ldots\ldots\ldots + \alpha_i + \delta_t + \mu_{it}$$

Where α_i = fixed or individual effect

δ_t = time specific intercept

μ_{it} = error term

And the general form of a random effect model is given as

$$Y_{it} = \beta_{1i} + \beta_2 X_{2it} + \ldots\ldots\ldots + \mu_{it}$$

$$\beta_{1i} = \beta_1 + \epsilon_{it}$$

$$Y_{it} = \beta_1 + \beta_2 X_{2it} + \ldots\ldots\ldots + \mu_{it} + \epsilon_{it}$$

The specific fixed effect model that we are going to estimate is

$$\text{Inventions}_{it} = \text{Population}_{it} + \text{FarmPIncome}_{it} + \text{EarningPlce}_{it} + \alpha_i + \delta_t + \mu_{it}$$

where α_i = fixed or individual effect

δ_t = time specific intercept

μ_{it} = error term.

5.5 Factors associated with cluster growth

While we are not trying to explore all of cluster formation theory, we are seeking to test indications of whether greater invention is observed in urban areas relative to rural and agriculturally intensive areas. As such, we regress only a handful of independent variables on our counts of inventions, by region.

The most common panel estimation techniques are fixed and random effect models. Results of these two models are shown in Table 13. In the fixed effect model any unobservable factors left out of the set of explanatory variables are considered time invariant, and, thereby, the fixed effects model help to remove bias in the estimator created by omitted variables. They are captured in α_i (e.g., the individual effect). In contrast, the random effects model assumes that unobservable variables are correlated with the variables in the model; while there may be smaller standard errors, coefficients may be biased due to the omitted unobservable variables. There is no straightforward answer to selecting between the fixed effects or random effect model. Although, theory suggests the Hausman test to choose between these two techniques. Selection between these two techniques depends on choice of variables in the model, the nature of omitted variables, whether those omitted variables are correlated with variables included in the model, and their variation across time.

There are indeed several factors—such as R&D expenditures, or invention and trade policy—which the literature has established are important for formation and growth of innovation clusters. However, without R&D expenditure data available at the regional level and without appropriate indicators of innovation and trade policy that would be meaningful at the regional level, these and other such factors are inevitably excluded from the equation. Therefore, it is important to control for unobservable factors to have unbiased estimators.

Table 13. Combined panel regression on counts of inventions by U.S. region and year, 1970-2010

Variables	Fixed Effects	Random Effects
Population	3.5768*	0.3346*
Farm Proprietor Income	2.2132*	4.0914*
Earnings by Place of Work	0.1103*	0.1348*
Constant	-0.7414	-0.7414
F -Statistic	F(896,35877) = 29.56 Prob > F = 0.0000	Wald chi^2 (3) = 9819.59 Prob > chi2 = 0.0000
R^2	within = 0.2120 between = 0.3349 overall = 0.2340	within = 0.2073 between = 0.3519 overall = 0.2707

* significant at 1%

The Hausman test in Table 14 also suggests the fixed effects model. The significant P-value recommends rejecting the null that the unobservable variables affecting the inventions are uncorrelated with the observable variables and to accept the alternative hypothesis that such unobservable variables affecting the inventions are correlated with the observable variables. Together, this suggests that the fixed effects model is the more appropriate for interpretation of results.

All the parameter estimates of covariates are positive and significant in the fixed effects model in Table 13. Highly significant coefficient value of population indicates that inventions in a particular region are highly dependent on the size of its population. Therefore, we confirm our expectations that the preponderance of biological inventions for agricultural and natural resource industries have been made in more urban areas. The significant positive coefficient on earnings, while not as large as the coefficient on population, shows that it is also correlated with number of inventions indicating that the level of economic activity as well as the quality of human capital are related to invention activity. Interestingly, the strongly positive coefficient on farm proprietor

earnings indicates that, all else being equal, those regions with more agricultural production also have greater rates of innovation for the industry.

Table 14. Hausman test of null hypothesis that unobservable variables affecting the inventions are uncorrelated with the observable variables

Variables	Coef.		St. Err	Sqrt (diag(V_b-V_B)	
	Fixed(b)	Random(B)	Difference (b-B)	S.E.	
Population	3.5768	0.3346	0.1134	0.0030	0.0030
Farm proprietor income	2.2132	4.0914	0.7490	0.0000	0.0000
Earning by place of work	0.1103	0.1348	0.0019	0.0000	0.1310
Chi(2) = (b-B)' [V_b-V_B)^(-1)](b-B) = 278.29					
Prob > chi^2 = 0.0000					

5.6 Spatial models

The significance of spatial analysis (i.e., introducing weights) through different techniques is debated in the scholarly literature. For this reason, we give both sets of results, with and without introducing spatial weights. Scholars give their justification both for and against using spatial weights. We are of the view that use of spatial weights in the analysis should be based on intuition. If it makes sense intuitively, then a researcher can use it; if not, then it should not be considered.

Four spatial models are discussed in the literature, namely, the spatial lagged model, the spatial autoregressive model, the spatial Durbin model, and the spatial error model. The spatial lagged model is also called local spatial model, where the weight matrix "W" is multiplied by the explanatory variables. It captures the spatial effect across all explanatory variables. The spatial auto regressive model introduces the weight matrix "W" in the dependent variables only. This model is also called global spatial model. When the weight matrix "W" is introduced to the

residuals, it is called a spatial error model. Intuitively, in our case, it is important to introduce the weight matrix "W" to both the dependent and explanatory variables, therefore, we are using the spatial Durbin model.

The general form of the spatial Durbin model is:

$$Y_{it} = \delta \sum_{j=1}^{N} W_{it} Y_{jt} + \beta X_{i,t} + \sum_{i=1}^{N} \sum_{j=1}^{N} W_{it} X_{i,j,t} + a_i + \delta t + \epsilon_{it},$$

where W_{it} is an 896 x 896 spatial weight matrix and

Y_{it} = count of inventions

X_{it} = explanatory variables

α_i = fixed or individual effect

δ_t = time specific intercept

μ_{it} = error term.

5.6.1 Construction of the weight matrix "W"

GeoDa software is utilized to construct 896 x 896 contiguity weight matrix. We have seen that around 210 polygons (Metropolitan and Micropolitan Statistical areas) have 3 neighbors each, and 205 have 4 neighbors each. The maxim neighbors of a polygon are 15. It seems important to consider the neighborhood spillover phenomenon because almost 46% of our data have 3 or 4 neighbors. After constructing the weight matrix, we export it to STATA for the econometric analysis.

5.6.2 Spatial panel estimation

By replacing "Y" with dependent variable (count of inventions), and "X" with explanatory variables (population, farm proprietor income, earnings by place of work) in the general form of the spatial panel equation. We estimated spatial panel equation as discussed above. The results (Table 15) on farm proprietor income are surprising. We got a substantially high value of farm

proprietor income by introducing the spatial weights. This may be called the "Des Moines effect". This shows that agriculturally dominated regions like Des Moines have a substantial effect on inventions. Conversely, the coefficient value of population is decreased by introducing the weights.

Table 15. Spatial panel regression on counts of inventions by U.S. region and year, 1970-2010

Variables	Fixed Effects	Random Effects
Population	1.9200*	1.9600*
Farm Proprietor Income	7.6000*	8.0200*
Earning by place of work	2.2100*	2.5100*
Spatial (rho)	0.0535	0.0528
R^2	within = 0.2008 between = 0.3387 overall = 0.2391	within = 0.1973 between = 0.4136 overall = 0.2977

5.7 Discussion and conclusion

We investigate the relationship of invention counts with other broad characteristics of metropolitan or rural regions in the United States. We show that numbers of inventions are positively related with population, and thus that these inventions tend to be made in more urban areas. Inventions are also positively related to workplace earnings, an indicator of the level of economic activity as well as the quality of human capital. Finally, all else being equal, those regions with more agricultural production also have greater rates of invention.

7. SPATIAL DISTRIBUTION AND COVARIATES OF REGIONAL CLUSTERING OF BIOLOGICAL INVENTIONS ACROSS OECD COUNTRIES

6.1 Introduction

Geographic concentration and clustering of invention activities, even across a single country, is a complex and dynamic process. In the context of the larger global economy, it is even more so. To extend our investigation of how biological inventions for use in primary resource-intensive industries—such as agriculture, energy, and natural resources—have been spatially distributed, we explore invention patterns observed across multiple member countries of the Organization of Economic Cooperation and Development (OECD), an international organization consisting of 37 mostly high-income countries around the world, including the United States which was the focus of analysis in the preceding two chapters. By looking more broadly at innovation across multiple countries, this chapter will seek more general answers to the same kinds of questions posed in earlier chapters: How have biological inventions been spatially distributed across the OECD member countries? And, to what degree have they been geographically concentrated at the sub-national level? (2) What are the space-time dynamics of biological inventions across the OECD? Where and when have they occurred? To what extent does the concentration of previous inventions effect where new inventions arise? (3) What are the determinants of inventiveness at the level of regions across OECD member countries?

The literature has documented trends in overall inventive activities at the national level across the OECD. Prior research also highlights the dynamic nature of clustering of inventive activities. But, the specification, classification, and categorization of inventions are important before discussing their spatial distribution, and underlying cluster growth factors. Breschi (2000)

stresses understanding the technological regime and the nature of technologies that introduce empirical regularities before analyzing the dynamic nature of invention clustering. Greunz (2003) emphasizes that the order of the geographical proximity matters: A region having closer proximity will have a larger role in the spatial pattern. Moreno et al (2005) confirms regional interdependencies matter in spatial distribution and concentration. Usai (2011) posits the inventive activities might be negatively affected if the technology is rural or service oriented.

To explore the dynamic nature of clustering across the OECD, this study considers only biological inventions having application in agriculture, energy, and the environment—variously framed as agbiotech, green biotechnology, or the bioeconomy. While this focus provides important technology and industry controls, it also allows an investigation of an important theoretical dilemma specific to these industries. Because they are highly dependent on land and other natural capital, production activities are necessarily geographically dispersed. Yet, economies of agglomeration and the driver of innovation based upon knowledge spillovers are dependent on co-location. Thus, it is not obvious whether innovation activities for these industries should be more dispersed, approximating the geography of production and the skilled labor pools of the industries, or if they should arise in dense clusters, as has been observed for biotechnology and other high technology industries. First, this study characterizes the spatial distribution of such biotech inventions in OECD countries to identify the degree of clustering. It then analyzes the emergence of those distribution patterns or clusters. Finally, it explores what factors or covariates are associated with biological inventions made a the level of sub-national regions across the OECD. This study contributes to the literature by exploiting inventor address data from patents to analyze the geography of invention and to explore covariates of the creation of knowledge.

The InSTePP Global Genetics Patent Database—which encompasses all patent filings made from 1970-2010 globally on biological subject matters—is used to map the spatial distribution of bioinventions based on lead inventor address, at the TL2 geographic region level for OECD member countries for which sufficient data is available. Leveraging the Derwent World Patent Index (DWPI) Manual Code classification scheme, biotechnologies with agricultural, energy, and environmental applications are selected. Analysis is done at the TL2 regional level—and not smaller—because data for most of the covariates (explanatory variables) are consistently available from the OECD only at the TL2 level for most member states. A number of covariates, including population, R&D expenditure, gross value addition of agriculture, and an intellectual property (IP) strength index are explored for their relationship with biotech cluster growth across the OECD. Based upon availability of data, different scenarios are run for this analysis, varying time period and the scope of countries included.

This chapter is organized as follows. Literature specific to clustering in the OECD is discussed in Section 2. Section 3 describes the data and methodology and presents descriptive summary statistics. Section 4 presents three different types of spatial analyses of biological inventions across the OECD. Section 5 closes with discussion and conclusions.

6.2 Literature and background

The literature that explores spatial distribution, and concentration of invention across OECD countries is more limited than the literature that focuses on spatial concentration within individual countries. North America and Europe are the two largest subsets of OECD an represent a large share of OECD economic activity. We already have discussed U.S specific literature in the preceding chapters. Therefore, to shift the focus to the OECD, we now explore in more detail some of the European literature and a few OECD-specific studies.

Breschi (2000) highlights empirical regularities in the spatial distribution of innovative activities across sectors in Europe. To understand the regularities he advances the importance of technological regime. "Technological regimes" are broadly defined by the level and type of opportunity and appropriability conditions and cumulativeness of technical knowledge and by the nature of knowledge and the means of knowledge transmission and communication. Breschi concludes that spatial distribution of inventive activities differs across technologies and the distribution pattern across technologies differs within and among countries.

Moreno et al (2005) analyze the spatial distribution of innovative activity measured by patent application across 175 regions of 17 countries in Europe. The time period they consider is 1978-2001. These authors use the Centro Ricerche Economiche Nord Sud (CRENoS: Centre for North-South Economic Research) database on regional patenting at yjr European Patent Office, classified by ISIC sectors. Two important findings are drawn from this study: (1) The mapping of innovation across Europe shows activities are concentrated in northern and central Europe, and only modest activities in southern Europe, and (2) with the passage of time the concentration of innovation is spreading to some Scandinavian and southern European regions. The authors also concluded that interregional interdependences exist. Internal regional factors, R&D expenditures, and agglomeration economies are found to influence innovative activity.

Usai (2011) explores the geography of inventive activities across Organization for Economic Cooperation and Development (OECD) regions. Usai also utilizes the CRENoS database, for two periods (1998-2000) and (2002-2004). The results show that OECD inventions are concentrated in Europe, North America, and Japan. The results also show distinct patterns are found in which highly inventive regions cluster together. R&D expenditures, human capital, and local agglomerations are important covariates expected to influence innovation activities across

the OECD. Usai also concludes that if a region is rural or when a technology is service oriented, these conditions may negatively influence inventive activity.

Greunz (2003) analyze inter regional knowledge spillovers of 153 European sub-national regions over the period of 1989-1996. Greunz concludes that geographical proximity of 1st, 2nd, and 3rd order matters. 1st order proximity means a region has a significant impact on patenting activity in another, and the effect slowly and gradually decreases. 4th order proximity means that a region does not matter to another. The author concludes that private sector R&D expenditures play an important role in national patenting activity.

Guastella and Van Oort (2015) emphasize on the importance of spatial heterogeneity while studying innovation clustering. The authors analyze patent applications during 2007-08 for the 250 NUTS-II regions of 25 E.U. countries. They argue that failing to consider spatial heterogeneity results in biased estimation. Based on Moran I values, the authors conclude that spatial association exists. The variables with substantially high explanatory value exhibit high spatial association. In this paper, market potential (the potential for activities, and enterprises in the region to reach markets and activities in other regions) and market size (gross value added per employee) display high spatial heterogeneity.

Finally, Tappeiner et al (2008) investigate patent applications for 51 European NUTS-1 regions and find the spatial distribution of inputs drive auto-correlation in patenting. They basically contradict the conventional empirical findings of spatial autocorrelation of inventive activities as evidence of knowledge spillovers. The main explanatory variables used in this analysis are R&D expenditures and human capital.

6.3 Data

Out of the total 210,057 patent families for biotech in agriculture, energy, and the environment, only 127,410 contain information on inventor address. Of these 127,410 inventions, the lead inventors of 91,794 is in one of the OECD member countries (see Figure 4). Of these, 48,693 were in the United States, as analyzed in Chapters 4 and 5. The remaining 43,101 were in all other OECD countries combined. Approximately, 10,000 European inventions are not considered because their address data were not readily able to be cleaned due to the wide variation in national standards of reporting European place names. The details of the geographic data cleaning process follow the steps discussed above regarding the global data set, in Chapter 3.

Table 16. Characterizing availability of lead inventor address data in primary filing records, by patent family

Lead inventor address data in priority filing	Patent family counts
Country only	41,618
Country + city/state	45,197
Country + city/state + zip	4,979
Total	91,794

Countries outside the OECD are not considered in this analysis. Of these, China is the largest inventor, with around 18,000 inventions. Out of the total 91,794 inventions from OECD countries, 41,618 indicate only the inventor's country of residence, 45,197 indicate the inventor's country and city, and 4,979 indicate country, city/state, and postal codes. Table 16 summarizes the availability of lead inventor address data in priority filling records, by patent family. Out of these 91,794 OECD inventions, 50,176 yielded a geocoded location at the city or postal code level and make up our sample of inventions.

For the 50,176 OECD inventions identified by the location of lead inventor, Table 17 shows how these inventions are categorized by industry of application. Agriculture has accounts

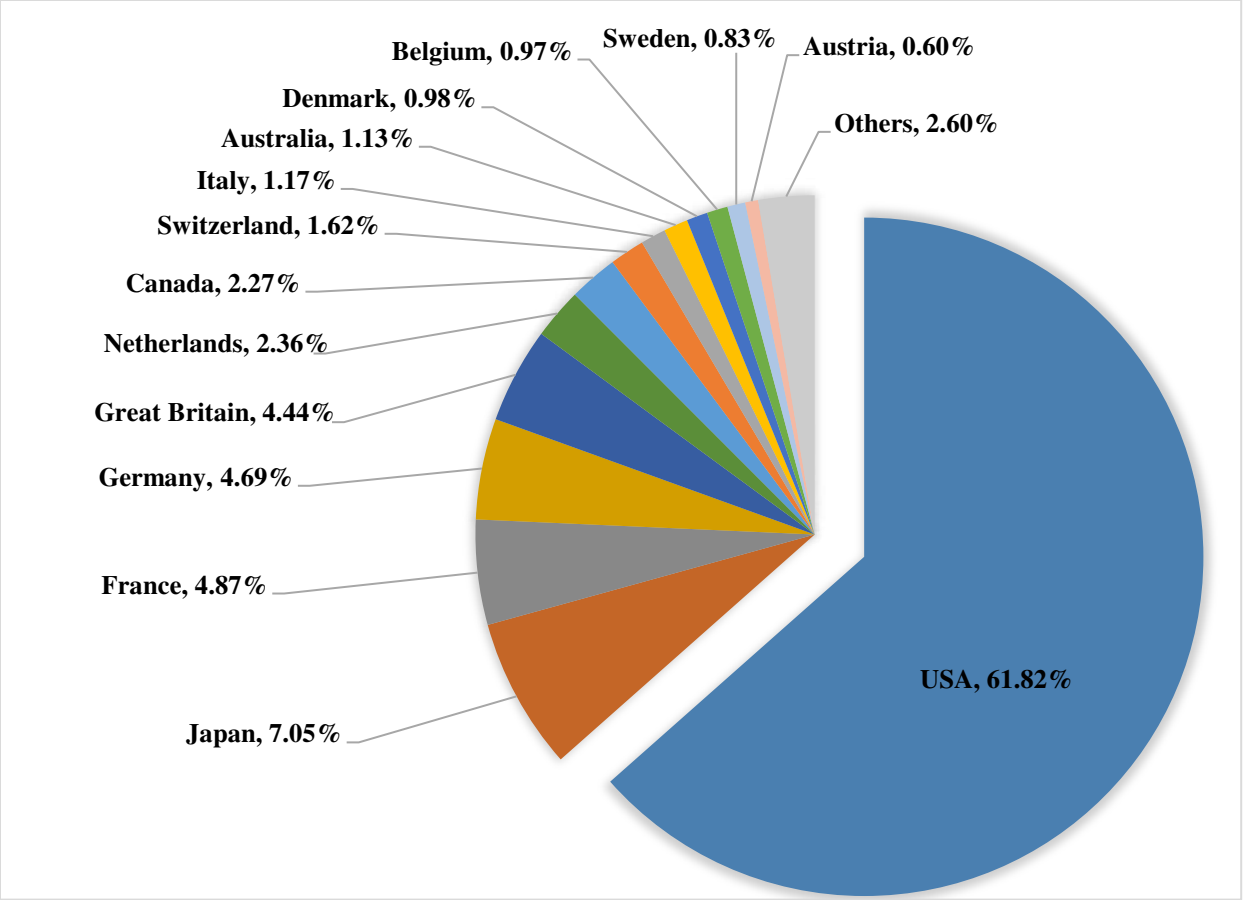
for the largest share, of 51.9%, followed by Environment 19.0%, and Energy 15.9%. The rest of the share shows how these technologies overlap across industries.

Table 17. Cross table of counts of inventions categorized by industry of application based on DWPI Manual Codes, for the 50,176 inventions with a lead inventor in an OECD country, including inventions assigned to multiple categories

Agriculture	26,032 (51.9%)		
Energy	1,211 (2.4%)	7,960 (15.9%)	
Environment	1,754 (3.5%)	2,515 (5.0%)	9,544 (19.0%)
	Agriculture	Energy	Environment

* 1,130 (2.3%) patent families are categorized in all three industries N= 50,176

Figure 10 represents the share of invention data of a lead inventor of a country available at the city and zip level for OECD region. The United States has the largest share, at 61.8% of the OECD inventions. The other countries with relatively large shares are Japan 7.1%, France 4.9%, Germany 4.7%, and Great Britain 4.4%. Figure 10 list the rest of the OECD countries' shares but combines those whose share are less than 0.5% each together in "Others" make up only 2.6%.

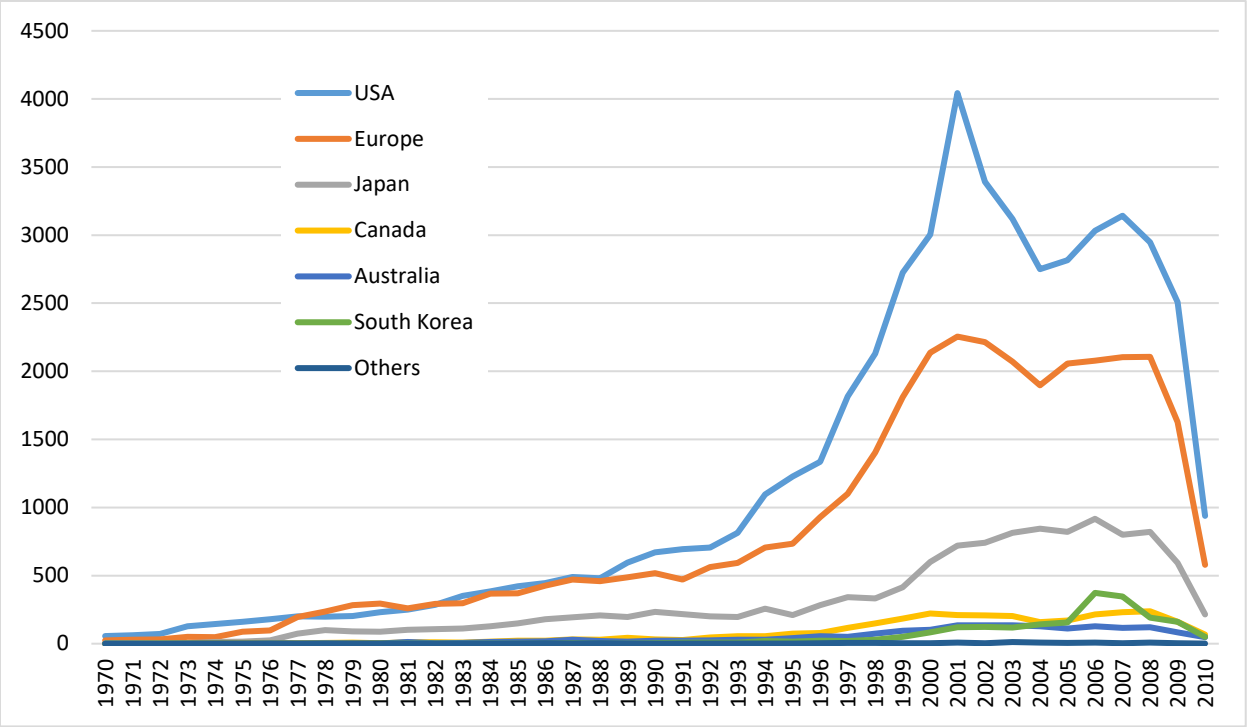


N= 50,176

Figure 11. Inventions measured by patent families from 1970-2010, for which lead inventor address with city level data is available

The annual counts of biological inventions are shown in Figure 12 for the OECD countries. The count of inventions are increasing from 1970 through 2000. Due to the largest share of the U.S within the OECD data set, the U.S specific invention policies like the 1980 Diamond v. Chakrabarty decision of the U.S. Supreme Court and the Bayh-Dole Act 1980 are important reasons for this increase. Other global factors that affect these trends are the Trade Related Aspects of Intellectual Property Rights (TRIPS) agreement in 1995. The bursting of the tech bubble in 2001 negatively affected the rise, and inventions gradually decreased thereafter. A slight increase can be seen after the economy started recovering in 2004. The financial crisis of 2007-08 is expected to negatively affect this pattern, but is confounded by the steep decline after 2007 due to data

truncation. Patent application filing and the patent examiner's first substitutive review takes about 21 months. In some cases, substantive examination takes 3-4 years. Grant or refusal takes place after this examination. The InSTePP data set was compiled in 2011, therefore, inventions that were in the examination process, and were not published or granted approval could not be observed at that time and are not part of this data set.



N= 50,176

Figure 12. Growth in inventions by country/region of lead inventor

6.3.1 OECD territorial level classifications and regional data

For the 37 OECD countries, geographic regions are classified on two territorial levels (TLs) for the purpose of reporting data, reflecting the administrative organization of countries: (1) the 398 larger territorial level (TL2) regions are roughly equivalent to states or provinces; and (2) the 2241 smaller territorial level (TL3) regions are roughly equivalent to metropolitan regions.

The European Union maintain a similar system of classifications designated Nomenclature of territorial units for statistics, abbreviated NUTS (from the French Nomenclature des Unités

Territoriales Statistiques), levels 1,2 and 3. NUTS-1 are major socio-economic regions. NUTS-2 are the basic regions for application of regional policies. NUTS-3 are small regions for specific diagnoses. Generally, TL2 is equivalent to NUTS2, and TL3 with NUTS3; however, this relationship is not consistent for some countries (OECD 2009).

Analysis at the TL3 or NUTS-3 level would most closely correspond to our analysis at the level of MSAs and microSAs in the United States in Chapters 4 and 5. Unfortunately, regional data for the TL3 regions of the OECD countries are not consistently reported. And, therefore, we must proceed in our exploration of factors explaining variation in inventions only at the higher level of aggregation of the TL2 regions, which in the United States corresponds to the state level.

Table 18 provides summary statistics for several factors with potential for explaining variation in inventions for the 193 TL2 regions of 17 OECD member countries for which complete data are available for the time period of 2000-2010 in the OECD regional data catalogue, resulting in a total of 2123 observations for each variable. To capture the agglomeration effects within a state or province (TL2 region), we consider its "Population" as a size measure. Higher population also indicates higher population density. Our hypothesis is that inventions are positively related to population. To explore whether the rural-urban division exists across OECD countries, we consider Gross Value Added (GVA) from agriculture by TL2 region. We see negative values of GVA in the data summary, indicating economic shocks to the agricultural sector in some regions. Other important explanatory variables expected to positively affect the level of inventions are Human Capital (tertiary education), R&D expenditure, and total regional income.

Table 18. Summary statistics of inventions and selected covariates, for years 2000-2010, for 193 TL2 regions of 17 OECD countries

Variables, by TL2 region, by year	Obs.	Mean	Std. Dev.	Min	Max
Invention counts	2123	9.04	29.53	0	514
Population (1,000s)	2123	3533.98	4281.71	25	37332
Income (US$ per head)	2123	19361.00	8693.88	4612	54339
GVA from Agriculture (millions of US$)	2123	1610.90	2287.05	-180	31485
R&D Expenditures (% GDP)	2123	1.61	1.32	0.06	10.24
Human Capital Tertiary Ed (per 1000)	2123	27.14	11.60	6.9	63.6
IP Index (national) (scale from 0 to 5)	2123	4.51	0.33	2.96	4.88

Data Sources: OECD Regional Data Catalogue, Ginarte and Park (2015)

In addition to these economic variables, an intellectual property (IP) index developed by Ginarte and Park (2015) is used to measure the strength and enforcement of IP across countries and years, which may be expected to impact the creation and patenting of inventions. The IP index ranges from 0 to 5, with 5 as the highest possible level of IP legal strength and enforcement in a country. Most of the high-income countries of the OECD are in the range of 4.88, which means these countries generally have strict IP regimes. The middle-income countries of the OECD have values in the range below the mean of 4.51.

6.3.2 Correlation analysis

Correlation analysis reveals some of the relationships among the factors we expect to account for the TL2 regional (state/province) level of inventions (Table 19). In general, this analysis shows that two of the explanatory variables (GVA from agriculture and population) are highly correlated with the dependent variable (inventions). The rest of the other explanatory

variables (human capita, R&D expenditures, and regional income) are low related with the dependent variables. We therefore expect a high regression coefficient of GVA and population, given the high correlation value of these two variables with invention. Interrelationships among the explanatory variables are low, except between these two population with GVA and income with the IP index. We cannot drop population or GVA based on intuition, because this relationship is important to see the rural urban dilemma. Similarly, the large value of IP index shows high enforcement of IP laws, which we also cannot drop, based on intuition.

Table 19. Correlation matrix

	Inventions	GVA of Ag.	Human Capital	IP Index	Pop.	R&D Expend.	Income
Inventions	1.00						
GVA of Ag.	0.69	1.00					
Human Capital	0.23	0.14	1.00				
IP Index	0.27	0.23	0.49	1.00			
Pop.	0.66	0.75	0.22	0.28	1.00		
R&D Expend.	0.25	0.10	0.41	0.40	0.20	1.00	
Income	0.34	0.26	0.52	0.73	0.36	0.41	1.00

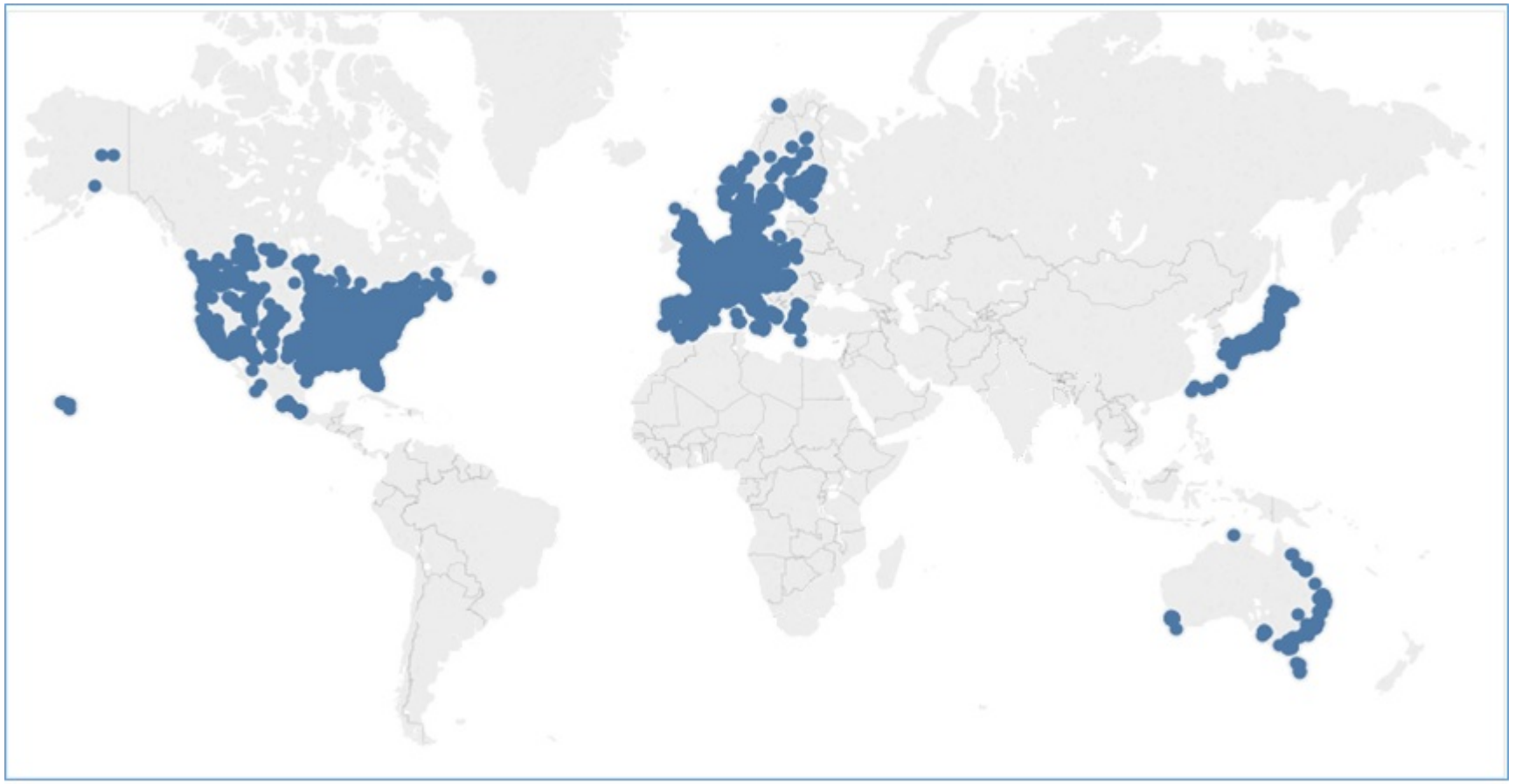
The high correlation observed among population, ag GVA, and our counts of biological inventions likely results directly from what we might call the "California and Ontario" effect. Since we are constrained in the analysis to TL2 regions, due to data availability, this means we are measuring large states and provinces, such as California in the United States or Ontario in Canada, which as regions have some of the largest populations, very large and rich agricultural industries, as well as large high-technology sectors (such as Silicon Valley and the San Diego Biotech Beach

in California or the high technology industry around Toronto and Guelph in Ontario.) This underscores the rationale, argued in Chapter 4 above, for pursuing analysis at the smaller geographic level of individual innovation clusters, which we showed to be more closely aligned with metropolitan regions and would invariably be better reflected we were able to access explanatory variable data from OECD denominated at the TL3 level.

6.4 Spatial analysis of inventions: Three techniques

6.4.1 Mapping of inventions

Arc GIS is used to map the spatial distribution and concentration of the 50,176 geocoded biological inventions across OECD countries (Figure 13). However, because of the small resolution of the global map, the spatial distribution and concentration is not readily visible at the regional level. We applied the same methodology discussed in Chapter 4 (for spatial distribution of inventions in the U.S.) to identify the 30 largest innovation clusters across all OECD member countries (Table 19). These 30 largest clusters account for 45% of the total inventions in the data set. The top four clusters are the same as those identified in Chapter 4 for the U.S., but the fifth cluster is Tokyo (Japan). The other top clusters outside the U.S are Paris (France), Osaka-Kyoto (Japan), London (Great Britain), Frankfurt-Heidelberg (Germany), Copenhagen (Denmark), Bonn-Koln (Germany), Toronto (Canada), Basel-Lorrach-Saint Louis (Switzerland), Delft-Leiden (Netherlands).



N= 50,176

Figure 13. Geographic pattern of biological inventions for use in agriculture, energy, and environment in OECD member countries

Table 20. The 30 largest clusters of biological inventions for agriculture, energy, and environment across OECD member countries, based on cumulative count of inventions 1970-2010

#	Top Identified Clusters	State/Region	Country	Inventions	%	Cumulative %
1	San Francisco	CA	US	3,020	6.37	6.37
2	New York City	NY	US	1,211	2.56	8.93
3	Washington-Baltimore	DC, MD, VA	US	1,187	2.51	11.43
4	San Diego	CA	US	1,135	2.40	13.83
5	Tokyo	Southern-Kanto	JP	1,067	2.25	16.08
6	Boston	MA	US	1,017	2.15	18.23
7	Los Angeles	CA	US	939	1.98	20.21
8	Houston	TX	US	924	1.95	22.16
9	Des Moines-Ames	IA	US	882	1.86	24.02
10	Paris	IIe-de-France	FR	879	1.86	25.88
11	Philadelphia region	PA, NJ, DE	US	811	1.71	27.59
12	Chicago	IL	US	713	1.51	29.09
13	Osaka-Kyoto	Kansai Region	JP	585	1.23	30.33
14	London	London-South East-East of England	UK	558	1.18	31.51
15	Madison	WI	US	544	1.15	32.66
16	Seattle	WA	US	516	1.09	33.74
17	Raleigh-Durham-Chapel Hill	NC	US	449	0.95	34.69
18	Frankfurt-Heidelberg	Hassen, Baden-Wurttemberg, Rheinland-Pfalz	DE	430	0.91	35.60
19	Cleveland	MI	US	420	0.89	36.49
20	Detroit-Ann Arbor-Lansing	MI	US	420	0.89	37.37
21	St Louis	MO, IL	US	415	0.88	38.25
22	Minneapolis-St Paul	MN	US	409	0.86	39.11
23	Denver-Boulder-Ft Collins	CO	US	376	0.79	39.91
24	Davis-Sacramento	CA	US	344	0.73	40.63
25	Copenhagen	Hovedstaden	DK	335	0.71	41.34
26	Bonn-Koln	Nordrhein-Westfalen	DE	334	0.71	42.04
27	Toronto	Ontario	CA	316	0.67	42.71
28	Basel-Lorrach-Saint Louis	Nordwestschweiz,Baden-Wurttemberg	CH	300	0.63	43.34
29	Delft-Leiden	Zuid-Holland, Noord-Holland	NL	297	0.63	43.97
30	Atlanta-Athens	GA	US	296	0.62	44.60

6.4.2 Spatial distribution

After analyzing visually, now we empirically measure the extent to which these biological inventions for agricultural, natural resource, and environmental applications exhibit spatial cumulativeness, and therefore remain relatively concentrated spatially. To see whether prior inventions affect current inventions, we regress invention counts on lags of invention counts for each TL2 region in each year (Table 21). For panel estimations, time series (AIC/BIC) lag length criteria are not appropriate. Optimal lag length for panel data can be determined manually by starting from a lag of 1 year, then 2 years, and so on, stopping when the coefficient of the lagged explanatory variable becomes negative. We find that lagged invention counts have a positive and significant effects on current inventions counts within the TL2 regions. The lagged invention counts are positively related to current invention counts for the previous two years: t-1 and t-2. This means that the recent past has more effect on current inventions. The lagged effect after two years disappears.

To check the for the overall concentration of inventive activities, we constructed a cumulative prior invention count variable defined as the sum of inventions from year 0 to year t-1. Table 22 shows the results of the cumulative sum of prior inventions on current year invention counts for each region for each year. Similarly, the cumulative sum of inventions has a positive and significant effect on current inventions. This means that biological innovation exhibits spatial cumulativeness and therefore remains relatively concentrated spatially.

Table 21. Fixed effects regression of lagged invention counts, 2000-2010, for 193 TL2 regions of 17 OECD countries

Variables	Coef.	St. Err	t	p>|t|
Inventions 1st lag	0.8771	0.0126	69.41	0.0000
Inventions 2nd lag	0.1372	0.0163	8.40	0.0000
Constant	0.4624	0.0671	6.89	0.0000
	F(3,7212) = 12180 Prob > F = 0.0000			

Table 22. Fixed effects regression of cumulative invention counts, 2000-2010, for 193 TL2 regions of 17 OECD countries

Variables	Coef.	St. Err	t	p>|t|
Cumulative Inventions	0.0522	0.0005	88.68	0.0000
Constant	-461.3751	5.2526	-87.84	0.0000
	F(1,7799) = 7863.40 Prob > F = 0.0000			

6.4.3 Covariate analysis

We also seek to explore covariates of invention counts, and thus potential general explanatory factors for cluster formation across OECD countries. To do so, we select as the dependent variable the count of bioinventions by TL2 region by year, and as explanatory variables the population, R&D expenditures (as % of regional GDP), income level (as regional GDP per capita), gross value added (GVA) of agriculture, human capital (as tertiary education level) by TL2 region by year (from OECD) and an intellectual property (IP) index by country and by year (from Ginarte and Park, 2015).

The panel estimation methods used are identical to that explained above in section 5.5 on factors associated with cluster growth in the U.S. However, in this case, two different scenarios are developed to overcome data availability limitations for TL2 regions of OECD member countries. Out of the current 37 OECD member countries, 27 countries have biological inventions counts over a sufficient number of years. We do have some countries with only one- or two-years' of inventions data. Therefore, those countries are omitted. The OECD has data for member countries from 1990 onward, but it is not possible to obtained a balanced panel before 2000 for all 27 countries.

6.4.3.1 First Scenario: 17 countries, 2000-2010

To obtain a balance panel, we initially select 17 countries for which we have data for all explanatory variables at the TL2 level for 2000-2010. We then run both a fixed effects and a random effects model for that panel.

Fixed Effects: The results of the fixed effects model (Table 23) show negative coefficient values on the regional population, GVA of agriculture, and income variables. The other explanatory variables have significant positive coefficient values. A negative effect of population indicates that inventions are more likely to come from TL2 regions with smaller populations. This result seems to contradict the initial conclusion from the cluster mapping, which had indicated that the largest clusters of biological inventions were in urban areas. Similarly, all else being equal, these biological inventions are more likely to come from regions with lower per capita income. Together, these results seem to indicate that systematically higher counts of these biological inventions tend to come from smaller urban or rural areas that are not as high income. This would be consistent with the "Des Moines" effect described above in section 5.6. While even though

many capital cities populate the ranks of the top 30 innovation clusters (Table 20) at the TL2 level it may be the lower tier regions that invent in greater numbers across the OECD.

Table 23. Preliminary results of panel regression on invention counts, for 193 TL2 regions from 17 OECD countries for years 2000-2010

Variables	Fixed Effects	Random Effects
Population	-0.1011*	0.0055*
R&D Expenditure	0.8167*	5.0578*
GVA of Ag	-0.0346*	0.0046*
IP Index	3.5880*	24.0186*
Human Capital Tertiary Ed	1.8521*	0.0609*
Regional Income	-0.0034*	-0.0010
Cons	428.104	-0.109.1335
F -Statistic	F(6,1924) = 4.84 Prob > F = 0.0000	Wald chi^2 (6) = 285.16 Prob > chi^2 = 0.0000
R^2	within = 0.1938 between = 0.5899 overall = 0.1679	within = 0.0114 between = 0.5981 overall = 0.1791

The negative value of the coefficient on the GVA of ag indicates that inventions are more likely to arise in areas that are less dependent on agriculture, i.e. more urban areas. This negative coefficient on GVA of ag supports our hypothesis of an urban-rural divide affecting innovation clusters for these technologies. And, as expected per the literature, higher measures of human capital, R&D expenditures, and the IP index are positively related to the count of inventions.

Random Effects: Justification for using a random effects model needs to assume that the unobservables are correlated with the variables in the model. Many unobservable factors, like innovation policies, biotech regulations, and trade policies, may be captured in the IP index. Therefore, random effects may be an appropriate model. If this is the case, we find positive coefficients for almost all explanatory variables, except regional income.

6.4.3.2 Second Scenario: 22 countries, 2000-2010

Several OECD countries, including France, Japan, Switzerland, Netherlands, and Portugal, have data for R&D expenditures and human capital available only at the TL1 level. Therefore, a second scenario is developed to incorporate these additional countries into the analysis by using the average national (i.e. TL1) value for a given year for all the TL2 regions within that country. It is therefore not possible to see regional variations within country due to the use of these aggregate level data. This compromise brings the total number of countries to 22.

Table 24. Preliminary results of panel regression on invention counts, for 198 TL2 regions from 22 OECD countries for years 2000-2010

Variables	Fixed Effects	Random Effects
Population	-0.0201*	0.0033*
R&D Expenditure	-2.1570**	0.0023
GVA of Ag	-0.0099*	-0.0049*
IP Index	0.6854	11.9824*
Human Capital Tertiary Ed	0.2649***	0.0514
Regional Income	-0.0002	-0.0007*
Constant	428.104	-0.109.1335
F -Statistic	F(6,1723) = 125.94 Prob > F = 0.0000	Wald chi^2 (6) = 242.40 Prob > chi^2 = 0.0000
R-Square	within = 0.3049 between = 0.4212 overall = 0.3436	within = 0.1453 between = 0.1106 overall = 0.0990

The fixed effect model results for this second scenario (see Table 24) differs somewhat from those for the first scenario. In this second scenario we see a strong negative value of the estimated coefficient on R&D expenditures, which seems to contradict the established literature. The results of the random effects model for this second scenario are very similar to the results of the random effects model for the first scenario.

6.5 Discussion and Conclusion

We have seen limited empirical research on regional innovation clusters in the OECD. The first of this kind of paper was published by Usai (2011). The data limitations that Usai mentions are relevant for this study as well. Usai's selection of countries and time period is limited to data availability. Incomplete coverage of OECD countries for the empirical analysis limits the analysis that is possible. TL3 (metropolitan regional) level of analysis is not possible given the data availability. Therefore, we follow Usai in this study concentrate on the TL2 level.

We included in the analysis only those TL2 regions with at least one invention. A number of regions were omitted because they did not show any inventions for the entire time period. For example, of 22 TL2 states in Mexico, we observe inventions in only 6. The remaining 16 states were dropped from the sample. We have also seen limited numbers of inventions in Eastern Europe. Similarly, most of the TL2 regions in countries like Turkey, Greece, and New Zealand no or only a have a limited number of inventions.

Overall, the OECD has a very diversified economic structure. The OECD countries' innovation and trade policies vary. It may be difficult to justify empirically clustering of these countries together in a single analysis. Having diverse economic and geographic structures it is not appropriate to use a spatial weight matrix. Moreover, a contiguity weight matrix is not useful if the unit of analysis (TL2) boundaries are disconnected.

By the OECD definition, some of the TL2s, in fact, lie inside another TL2. Instead of combining/aggregating these two TL2s we treated them separately. This choice also slightly affects the regional invention variation. But, the total number of such "enclosed" TL2s are only 6 across the OECD. However, these TL2s are highly concentrated. For example, DE3-Berlin is embedded inside DE4-Brandenburg.

We did not customize by combining two or three TL2s if they have a single cluster of inventions that spans the TL2 boundaries, as we did in Chapter 4 for the U.S. statistical areas. We found that in most of highly concentrated TL2 regions, inventions tend to be spread across the whole TL2. Without TL3 level of data, it is not possible to select the more dense part of a TL2 region and separate it from the remaining part of TL2 which was more sparsely populated.

While these results are still preliminary, due to ongoing data limitations, several observations and comments can be offered in conclusion:

1. While inventions are distributed across the OECD, there do appear to be concentrated clusters of invention occurring in larger urban regions (based on identification of the top 30 clusters across the OECD).
2. Inventions made in prior years have a significant impact on a current year's inventions. This represents the localized spillover phenomenon and the cumulative nature of cluster formation.
3. Region size (as measured by population) and level of economic activity (as measured by income) do not appear to be related to regional count of invention activity for these industries. Likely the systematically higher numbers of inventions are coming not from the very largest and richest urban areas, but more likely from smaller urban areas with higher rates of R&D.
4. R&D expenditures (regional) and the IP index (national) that characterize a region are strongly related to the observed invention activity.
5. A rural-urban division does appear to exist. Invention counts appear to be negatively correlated with gross value added of agriculture by region.

6. CONCLUSIONS AND FUTURE DIRECTIONS

This study analyzes the spatial distribution and concentration of biotechnologies developed for application in agriculture, energy, and the environment. The overall findings of this study may be useful to inform policy makers about geographic patterns of knowledge creation and spillovers, which economic and policy factors drive invention activity at the regional scale, and, indirectly, the role of regional clustering in driving innovations for food security and sustainability. This analysis explores the dual dilemma that seems to affect the formation of innovation clusters for these geographically diffused industries. If innovators tend to aggregate, yet producers are constrained to be located in rural areas, they are necessarily distant from one another leading to the "dual dilemma" of innovation in agricultural and resource industries:

1. Urban cluster-born innovations are isolated from the rural community of skilled users
2. Rural user-led innovations are necessarily diffused, not well connected to urban clusters.

First, we analyze in the United States the clustering of invention of biotechnologies for applications in agriculture, energy, and environment. Using a novel patent data set and applying exploratory and empirical techniques, Chapter 4 concludes that, while biological inventions are distributed all across the U.S, highly concentrated clusters have emerged in largely urban regions. Clearly a spatial clustering pattern exists: inventions do not tend to be diffused as are the production activities in these industries but tend to concentrate in urban areas.

Regressions are developed to show that the number of inventions made in prior years within a metropolitan statistical area has a significant impact on the probability of inventions made in subsequent years. This relationship represents the localized spillover phenomenon and the cumulative nature of innovation clusters. Finally, while we do see inventions in rural areas in the

U.S., the rural areas simply do not appear to be the hotspots of innovation in agricultural, energy, or environmental biotechnologies.

Next, we look for covariates of the regional concentration of biological inventions for agriculture, energy, and the environment in the United States. Regression analysis finds that, indeed, inventions are positively related with population and economic activity. Moreover, regions with more agriculture production also have greater rates of innovation for the industry. Based on the results we can say these technologies belong to "Schumpter Mark II" camp of deepening and concentrated growth in innovation.

Finally, we expand the scope of analysis to examine the spatial distribution and covariates of biological inventive activity in regions across OECD countries. Summary statistics, mapping, and regression analysis are used to study the spatial distribution and covariates associated with regional invention activity across OECD countries. The exploratory analysis shows biological inventions are spread across OECD countries, but clusters are found in larger urban regions. Regression analysis shows, again, that previous inventions within a region seem to give rise to new inventions. Covariate regressions show that expected measures of human capital, R&D expenditures, and intellectual property strength are strongly related to invention activity at the regional level. However, region size (as measured by population) and level of economic activity (as measured by regional income) are inversely related to invention activity for these industries, but at the TL2 (state/province) level, this can have several interpretations. Across OECD countries, invention counts appear to be negatively correlated with gross value added of agriculture by region, implying a rural-urban division may in fact exist more universally: (urban) innovation is not co-located with (rural) production.

Based on these results we draw a few implications for agricultural innovation policies:

- First, policies that seek to encourage biotech innovation and its commercialization in agriculture and other resource-intensive rural industries need to recognize that the preponderance of inventions are being made in urban areas. This is consistent with economic theories on economies of agglomeration and innovation, and it appears to be a normal pattern.
- Commonly held policy objectives for economic development in rural and agricultural communities is not likely to succeed by seeking to supplant or compete with urban-based innovation clusters. But, rural economic development efforts may find some opportunities by seeking to nudge, shift, or complement existing innovation clusters on the margin.
- There is a dual market failure at play when spillovers are hindered due to geographic dispersion, or the lack of agglomeration. Not only is there private underinvestment in the underlying (R&D) activity that generates those positive externalities. But there is even failure of those positive externalities that are generated to have as much beneficial impact on third parties as they might have, due to high search costs, travel costs, and other transaction costs.
- The relevance of the Land Grant system in the United States and concomitant public investments in applied R&D at such intuitions continues to play important roles in facilitating innovation for these industries by creating new human capital, networking existing human capital, and facilitating knowledge spillovers across the clusters and the peripheries of these geographically dispersed industries, both from urban to rural and from rural to urban.

- Technology transfer and commercialization strategies should recognize that potential partners for further development of new biotechnologies are likely to be found in one of a handful of major clusters around the country.
- State and regional policymakers, economic development officials, agriculture officials, and strategic partners in industry need to consider collective action for fostering largely urban entrepreneurship for largely rural industries and creating linkages between them.
- Recognize and seek ways to address the extent to which (urban) innovators and (rural) producers in these industries are not intimately co-located.

The biotechnology industry has been a powerful force for innovation and economic development largely due to long-sighted R&D policies. Significant investments in basic research along with strong by transparent intellectual property and regulatory policies have been acknowledged as crucial elements in giving rise to the industry. What has not been acknowledged to the extent that it has likely make a difference is the strategic development of biotechnology clusters. The importance of co-location and economies of agglomeration has been significant for the growth of the industry. However, it is important to recognize that, outside of human therapeutics and manufacturing-based industrial applications, the virtues of these economies of agglomeration may begin to break down. The dual dilemma of agriculture appears to be the user led innovations are necessarily diffused, while cluster born innovations are isolated from the community of skilled users. Urban based innovators and rural users are distant from one another. One of the crucial and most promising interventions is the Land Grant system, holding the innovation system together and facilitation what spillovers do occur, even contributing to the formation and growth of many of the major innovation clusters that we observe today.

The importance of future innovations in these technologies are immense—in terms of assuring food security, economic development, and sustainability of agriculture, energy, and resources. An understanding of what are effectively the ecosystems that sustain and drive such innovation is essential to sustaining it for the challenges faced ahead.

REFERENCES

Acs, Zoltan J., (Eds.) (2001), Regional Innovation, Knowledge and Global Change. London: Pinter Publishers.

Acs, Zoltan J., Anselin, L., and Varga, A., (2002) "Patents and innovation counts as measures of regional production of new knowledge," Research Policy 31(7) 1069-1085

Albers, Stevan C., Berklund, Annabele M., and Graff, Gregory D., (2016) "The rise and fall of innovation in biofuels" Nature Biotechnology, Vol 34, No. 8:

Alston, Julian M., Matthew A. Andersen, Jennifer S. James, and Philip G. Pardey, (2010) Persistence Pays: U.S. Agricultural Productivity Growth and the Benefits from Public R&D Spending, New York: Springer.

Arrow, Kenneth J. (1971) "The economic implications of learning by doing" In Readings in the Theory of Growth. Palgrave Macmillan, London. 131-149.

Audretsch, David B., and Paula E. Stephan (1996) "Company-scientist locational links: The case of biotechnology," The American Economic Review, 86(3), 641-652.

Audretsch, David, B., and Maryann P. Feldman, (1996) "R&D spillovers and the geography of innovation and production," American Economic Review, 86(3), 630-640.

Bauernschuster, S., Falck, O., and Heblich, S., (2010), "Social capital access and entrepreneurship." Journal of Economic Behavior & Organization. 76, 821-833

Boschma, Ron., and Fornahi, Dirk., (2011), "Cluster Evolution and a Roadmap for future research." Regional Studies, 45:10. pp. 1295-1298

Breschi, S., (2000), "The geography of innovation: A cross-sector analysis." Regional Studies, 34:3, pp. 213-229

Breschi, Stefano (2010) "The Geography of Innovation: A Cross-sector Analysis," Regional Studies, 34(3), 213-229

Breschi, Stefano, Franco Malerba, and Luigi Orsenigo (2000) "Technological regimes and Schumpeterian patterns of innovation," The Economic Journal, 110(463), 388-410

Cohen, Wesley M., and Daniel A. Levinthal (1989) "Innovation and learning: the two faces of R&D." The Economic Journal, 99(397), 569-596.

Delgado, M., Michael, P., and Scott, S., (2010), "Clusters and entrepreneurship," Journal of Economic Geography, 10, 495-518

DeVol, Ross, Perry Wong, Junghoon Ki, Armen Bedroussian, and Rob Koepp (2004) America's Biotech and Life Science Clusters, Milken Institute.

Ellison, Glenn, Glaeser, Edward, L., and Kerr, William, R. (2010) "What causes industry agglomeration? Evidence from coagglomeration patterns," American Economic Review, 100(3), 1195-1213

Fischer, Manfred, M., (Eds), (2006), Innovation, Network, and Knowledge Spillovers. Berlin Heidelberg: Springer.

Foley, J.A., Ramankutty, N., Brauman, K.A., Cassidy, E.S., Gerber, J.S., Johnston, M., Mueller, N.D., O'Connell, C., Ray, D.K., West, P.C. and Balzer, C. (2011) "Solutions for a cultivated planet," Nature, 478, 337-342.

Glaeser, E, L., and Resseger, M, G. (2010) "The complementarity between cities and skills," Journal of Regional Science, 50(1), 221-244

Graff, Gregory D., Susan E. Cullen, Kent J. Bradford, David Zilberman, and Alan B. Bennett (2003) "The public–private structure of intellectual property ownership in agricultural biotechnology," Nature Biotechnology, 21(9), 989-993

Graff, Gregory D., Devon Phillips, Zhen Lei, Sooyoung Oh, Carol Nottenburg, and Philip G. Pardey (2013) "Not quite a myriad of gene patents," Nature Biotechnology, 31(5), 404-410.

Graff, Gregory D., Devon Phillips, and Philip G. Pardey (2015) The InSTePP Global Genetics Patent Database: Data Documentation, International Science and Technology Policy and Practice (InSTePP), University of Minnesota.

Greunz, L,. (2003) "Geographically and Technologically Mediated Knowledge Spillovers Between European Regions." The Annals of Regional Sciences, 37(4), pp. 657-680.

Griliches, Zvi (1990) "Patent statistics as economic indicators: A survey." Journal of Economic Literature, 28(4), 1661-1707.

Guastella, G., and Oort F V., (2015) "Regional Heterogeneity and Inter Regional Research Spillovers in European Innovation: Modelling and Policy Implications." Regional Studies 49 (11), 1-16.

Hall, Bronwyn H., Adam B. Jaffe, and Manuel Trajtenberg (2001) The NBER patent citation data file: Lessons, insights and methodological tools, No. W8498, National Bureau of Economic Research (NBER).

Hausman, Jerry, Bronwy H. Hall, and Zvi Griliches. (1984) "Econometric Models for Count Data and with Application to the Patents-R&D Relationship," Econometrica, 52(4), 909-938.

Hefley, T.J. and Hooten, M.B., (2016) "Hierarchical species distribution models." Current Landscape Ecology Reports, 1(2), 87-97.

Huffman, Wallace E., and Robert E. Evenson (2006) Science for Agriculture: A Long-Term Perspective, 2nd Edition, Ames, IA: Blackwell Publishing.

Jafee, A., M. Trajternberg, and R. Hendrson., (1993), "Geographic localization of knowledge spillovers as evidenced by patent citation." Quarterly Journal of Economics, 108:3, pp.577-598

Krugman, Paul (1991) Geography and Trade, Cambridge, MA: MIT Press.

Kwon, Hyuk-Soo., Lee, Jihong., Lee, Sokbae., and Oh, Ryungha., (2017), "Knowledge spillovers and patetn citations:trends in geographic localization, 1976-2015." Cemmap working paper CWP55/2017

Lim, Up., 2003, "The spatial distribution of innovative activity in U.S. Metropolitan Areas: Evidence from patent data." The Journal of Regional Analysis & Policy, 33:2.

McCann, P., (2013), Modern urban and regional economics. Oxford: Oxford University Press

Malerba, Franco, and Luigi Orsenigo (1990) "Technological regimes and patterns of innovation: a theoretical and empirical investigation of the Italian case," in Arnold Heertje and Mark Perlman, Eds., Evolving technology and market structure, (University of Michigan Press, Ann Arbor, MI), 283-305.

Malerba, Franco, and Luigi Orsenigo (1996) "Schumpeterian patterns of innovation are technology-specific," Research Policy, 25(3), 451-478.

Marshall, Alfred (1890) Principles of Economics, Macmillan, London.

Marshall, Alfred (1920) Principles of Economics, MacMillan, London.

Martínez, Catalina (2010) "Patent families: When do different definitions really matter?" Scientometrics, 86(1), 39-63

Moreno, R., Paci, R., and Usai, S., (2005) "Geographical and sectoral clusters of innovation in Europe." The Annals of Regional Science, 39(4), 715-739

Nelson, Richard, and Sidney Winter, (1982) An evolutionary theory of economic change, Bellknap, Cambridge, MA.

OECD (2009). OECD Regions at a Glance, OECD Publishing

Rauch, F., (2014) "Cities as spatial clusters." Journal of Economic Geography, 14, 759-773

Ryan, Camille D., and Peter W. Phillips (2004) "Knowledge management in advanced technology industries: an examination of international agricultural biotechnology clusters," Environment and Planning C: Government and Policy, 22(2), 217-232.

Samad, Ghulam and Graff, Gregory D. (2020) “The Urban Concentration of Innovation and Entrepreneurship in Agricultural and Natural Resource Industries” Chapter 6 in N. Iftikhar, Jonathan Justice, David Audretsch, Eds., Urban Studies and Entrepreneurship, (Springer)

Saxenian, Annalee (1996) Regional Advantage: Culture and competition in Silicon Valley, and Route 128. Harvard University Press, Cambridge, MA.

Schmookler, Jacob (1954) "The level of inventive activity," Review of Economics and Statistics, 183-190.

Sunding, David, and David Zilberman. (2001) “The agricultural innovation process: Research and technology adoption in a changing agricultural sector.” Handbook of Agricultural Economics: 207-261.

Sun, Y., (2000), “Spatial distribution of patents in China.” Regional Studies, 34:5, pp. 441-454

Tan, Duoduo., Cheng, C., Lei, M., and Zhao, Y., (2017), “Spatial distributions and determinants of regional innovation in China: Evidence from Chinese Metropolitan Data.” Emerging Markets Finance & Trade, 53. pp. 1442-1454

Tappeiner, Gottfried & Hauser, Christoph & Walde, Janette, (2008). “Regional Knowledge Spillovers: Facts or artifact?” Research Policy, Vol 37 (5), pp. 861-874

Ter Wal, Anne L.J., (2013) “Cluster emergence and network evolution: A longitudinal analysis of the inventor network in Sophia-Antipolis.” Regional Studies, 47(5), 651-668.

Thompson, P., and M. Fox-Kean., (2005), “Patent citations and the geography of knowledge spillovers: A reassessment.” American Economic Review, 95:1. pp. 450-460

USDA Economic Research Service (2018), Farm finance indicators, state ranking, 2016, United States Department of Agriculture, Washington DC.

Usai, S., (2011) “The geography of inventive activity in OECD Regions.” Regional Studies, 45:6, pp. 711-731

Von Hippel, Eric (1988) The Sources of Innovation, Oxford: Oxford University Press.

Wang, Z., Cheng, Y., Ye, Zinyue., and Wei, Y.H. Dennis, (2016), Analysing the space-time dynamics of innovation in China: ESDA and Spatial Panel Approaches. Growth and Change, 47(1) 111-129

Zucker, Lynne G., and Michael R. Darby (1996) “Star scientists and institutional transformation: Patterns of invention and innovation in the formation of the biotechnology industry,” Proceedings of the National Academy of Sciences, 93, 12709-12716.