

Mukim, Megha

Conference Paper

Does agglomeration boost innovation? An econometric evaluation

50th Congress of the European Regional Science Association: "Sustainable Regional Growth and Development in the Creative Knowledge Economy", 19-23 August 2010, Jönköping, Sweden

Provided in Cooperation with:

European Regional Science Association (ERSA)

Suggested Citation: Mukim, Megha (2010) : Does agglomeration boost innovation? An econometric evaluation, 50th Congress of the European Regional Science Association: "Sustainable Regional Growth and Development in the Creative Knowledge Economy", 19-23 August 2010, Jönköping, Sweden, European Regional Science Association (ERSA), Louvain-la-Neuve

This Version is available at:

<https://hdl.handle.net/10419/119220>

Standard-Nutzungsbedingungen:

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

Terms of use:

Documents in EconStor may be saved and copied for your personal and scholarly purposes.

You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.

If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.

Does agglomeration boost innovation? An econometric evaluation

(preliminary draft – please do not circulate or quote without permission from the author)

Megha Mukim¹
London School of Economics
May 2010

Abstract:

Innovation is crucial to regional economic competitiveness and to productivity growth. A salient feature of the Indian economy is the geographic concentration of both, economic activity and innovation; the latter measured by patent activity. Theoretical models argue that the clustering of economic activity within a geographic region results in knowledge spillovers, which in turn drives innovation. The literature also posits that the presence of human capital is critical to the generation of new knowledge. This paper studies how and why economic geography and factor endowments matter for innovative activity – in other words, what is the relationship between human capital and patent generation, and crucially, how is this affected by the spatial distribution of economic activity?

The paper analyses patent activity (applications) between 1999 and 2007 across districts in India. By using an econometric model, it estimates the effects of R&D expenditures, economic clustering and the distribution of human capital endowments in generating innovation. I find that apart from R&D spending, agglomeration economies and the skill-set of workers have a significant effect on patent activity in a district. The estimates are robust to simultaneity bias, to different model specifications and to the type of applicant. Understanding the magnitude of the effects of economic geography and factor endowments is vital for policy formulation aimed at encouraging innovative activity.

JEL Classification: R12, O3

I. Introduction	2
II. Related Literature	3
III. Patents in India	5
IV. Empirical Model	8
V. Results and Discussion	13
VI. Endogeneity Issues and Robustness Checks	17
VII. Conclusion	21

¹ Email: m.mukim@lse.ac.uk

Website: <http://personal.lse.ac.uk/mukim>

I. Introduction

Economists and urban geographers have long noted the co-location of clusters of economic activity and of innovative activity. Theory dictates that people and organisations choose to live close to one another mainly because it is expensive to transport goods, people and ideas across space. Historically, transportation costs have been falling and the advent of new information and communication technologies led many to predict the ‘death of distance’. On the contrary, economic activity continues to be concentrated and agglomerations continue to exist, and grow in size. If it is no longer as costly to transport goods, people and ideas from one place to the next, why do people choose to agglomerate?

It was Marshall (1919) who was the first to formally identify the benefits of agglomeration. Clustering of economic activity allows for the sharing of inputs, skilled labour and knowledge – producers can access both buyers and suppliers more easily and at lower costs, and can exploit the presence of a thick labour market with multiple specialisations. Agglomeration also facilitates face-to-face contact by shortening interaction distances, which results in knowledge spilling over, leading to innovative activity.

Romer (1986, 1990) developed a theory to explain economic growth by making endogenous the effect of the accumulation and spillover of technological knowledge. This seminal contribution provided an explanation of economic growth that went beyond a simple focus on the role of investment in physical capital and increases in the labour supply. Changes in technology constitute an important source of new knowledge.

And lastly, Feldman and Florida (1994) suggest that innovation is ‘increasingly dependent on a geographically defined infrastructure that is capable of mobilising technical resources, knowledge, and other inputs essential to the innovation process’. Examples of such innovation-supporting advantages include the presence of networks of industry, concentration of R&D activity, be it in the private or in the public sector, and the presence of a large pool of service activities with multiple specialisations.

In this paper, I employ a regional knowledge production framework to empirically identify the effect of agglomeration economies, knowledge spillovers and infrastructure on innovative output. Section II outlines the literature and discusses the main focus of the most relevant papers. Section III provides a descriptive overview of patent activity in India. Section IV describes the empirical model and the sources and description of the data. This is followed by a summary of the study results and a discussion of the key findings in Section V. Section VI addresses possible endogeneity concerns and carries out robustness checks, and the paper concludes with a discussion of key contributions and limitations.

II. Related Literature

The idea that clustering creates externalities that could lead to an increase in innovation and which in turn could drive economic growth, is not a new one (Krugman 1991, Kelly and Hageman 1999, Paci and Usai 1999 and Hanson 2001). The literature has jointly dealt with the concepts of agglomeration and innovation under the general ‘research umbrella’ of economic geography. Different studies have defined their clusters of observations differently – as industrial districts, technological clusters, learning regions, and innovation milieus (for an overview of these studies see Ibrahim 2009).

Griliches (1979) proposed such a knowledge function approach at the level of the firm – a Cobb-Douglas utility function measured a firm’s innovative output to physical R&D inputs and some indicators of the knowledge stock. Jaffe (1986) was the first to use such a knowledge production function framework, within which he measured the effect of knowledge spillovers on innovative output. The knowledge production function approach has since then been widely used in the empirical literature – see Table 1 for an overview of some of these studies.

The empirical literature has many findings regarding the effect of regional differences on innovative output. For instance, some studies find evidence of a positive impact of localised knowledge spillovers resulting mainly from the presence of academic research (Anselin et al 1997, Fischer and Varga 2003, Fritsch and Slavtchev 2007, Kantor and Whalley 2009, Ponds 2010). Others concern themselves with the effects of neighbours, i.e. the economic geography impact of R&D investments in surrounding regions (Bottazzi and Peri 2003, Bode 2004). Some authors, those who are able to access superior data, use patent citations to estimate the extent of knowledge spillovers (Jaffe et al 1993, Co 2005). A number of studies also draw out the difference between diversity and specialisation within a location and the associated impact on innovation (Feldman and Audretsch 1999, Carlino et al 2007).

Most studies have made use of data from the United States (Audretsch and Feldman 1996, Anselin et al 2000, Rodriguez-Pose and Crescenzi 2008, Crescenzi et al 2007²) or from the European region (Autant-Bernard 2001, Fischer and Varga 2003, Castro and Quevedo 2005, Fritsch and Slavtchev 2007). In short, there are very few studies that deal with developing countries, and none that analyse India. This paper attempts to fill in this gap in the literature.

² Free availability of US patent data from the NBER website has helped to spur multiple empirical studies based on the data.

Table 1: Summary of Empirical Literature

Study	Country	Sample	Unit of Observation
Jaffe (1986)	USA	432 firms; 1973, 1979	Firm
Jaffe et al (1993)	USA	950 and 1,450 patents; 1975, 1980	State
Feldman and Florida (1994)	USA	4,200 innovations; 1982	State
Audretsch and Feldman (1996)	USA	8,074 innovations, 1982	State
Anselin et al (1997)	USA	4,200 innovations; 1982	MSA
Feldman and Audretsch (1999)	USA	3,696 innovations; 1982	MSA
Paci and Usai (1999)	Italy	1978-1995	Local Labour System
Anselin et al (2000)	USA	1982	MSA
Piergiovanni and Santarelli (2001)	France	20,700 patents; 1991-1992	State
Autant-Bernard (2001)	France	1994-1996	Department
Acs et al (2002)	USA	4,476 patents; 1982	MSA
Bottazzi and Peri (2003)	Europe	3,010 patents; 1977-1995	NUTS 1/NUTS 2
Fischer and Varga (2003)	Austria	1993	District
Bode (2004)	Germany	27,361 patents; 1998	NUTS 3
Co (2005)	New England, USA	1975-1999	MSA
Del Barrio-Castro and Garcia-Quevedo (2005)	Spain	1996-2000	NUTS 2
Barkley et al (2006)	Southern counties, USA	125,589 patents; 1990-1999	MSA
Carlino et al (2007)	USA	581,000 patents; 1990-1999	MSA
Girma et al (2007)	China	239,085 firms; 1999-2005	Firm
Fritsch and Slavtchev (2007)	Germany	31,434 patents; 1995-2000	NUTS 3
Knudsen et al (2008)	USA	1999	MSA
Rodriguez-Pose and Crescenzi (2008)	Europe	1995-2003	NUTS 1/NUTS 2
Ibrahim et al (2009)	USA	5,353 patents; 2000-2003	Telecom clusters
Kantor and Whalley (2009)	USA	1981-1996	County
Kerr (2009)	USA	1975-1984	City
Menon (2009)	USA	1,161,650 patents; 1975-1999	MSA
Ponds et al (2010)	The Netherlands	3,332 patents; 1999-2001	NUTS 3

Notes:

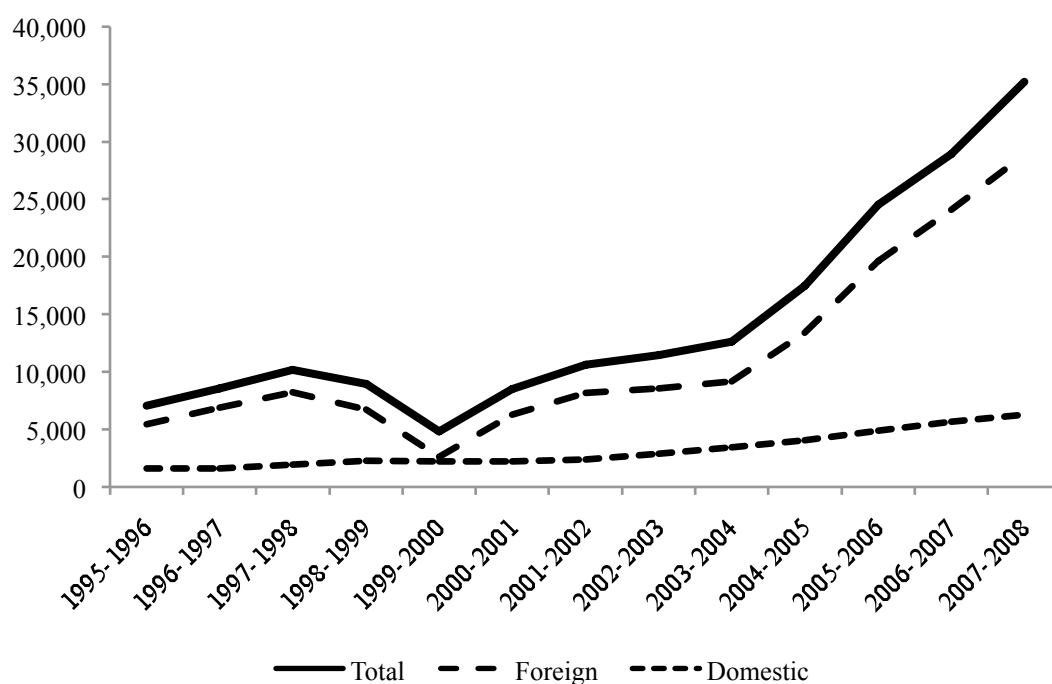
MSA: Metropolitan Statistical Area

NUTS: European Nomenclature for Territorial Units for Statistics

III. Patents in India

Patenting activity is considered a yardstick for innovative activity within a country, and in keeping with the literature, this paper uses patents as a proxy for the outcome of the inventive process. Data on patents is taken from the Indian Patent Office (IPO). The IPO registers patent applications at its four branches located in Kolkata (West Bengal), Chennai (Tamil Nadu), Mumbai (Maharashtra) and Delhi (Delhi). Although it is easier to aggregate data on the basis of which office an application was made to, the analysis within this paper requires data on the address of the applicant, or the location of the organisation or firm, at the level of the district. Data disaggregated at the level of the state is made available in the annual reports of the IPO, however information on the exact location of the applicant is only contained within the weekly journals published by the IPO³. Each patent file contains information on the following variables of interest: (1) date of application (2) the name of the inventor(s) (3) the address of the inventor(s) and, (4) the International Patent Classification (IPC) code.

Figure 1: Patent Applications to the IPO (1995-2008)



Source: IPO Annual Reports

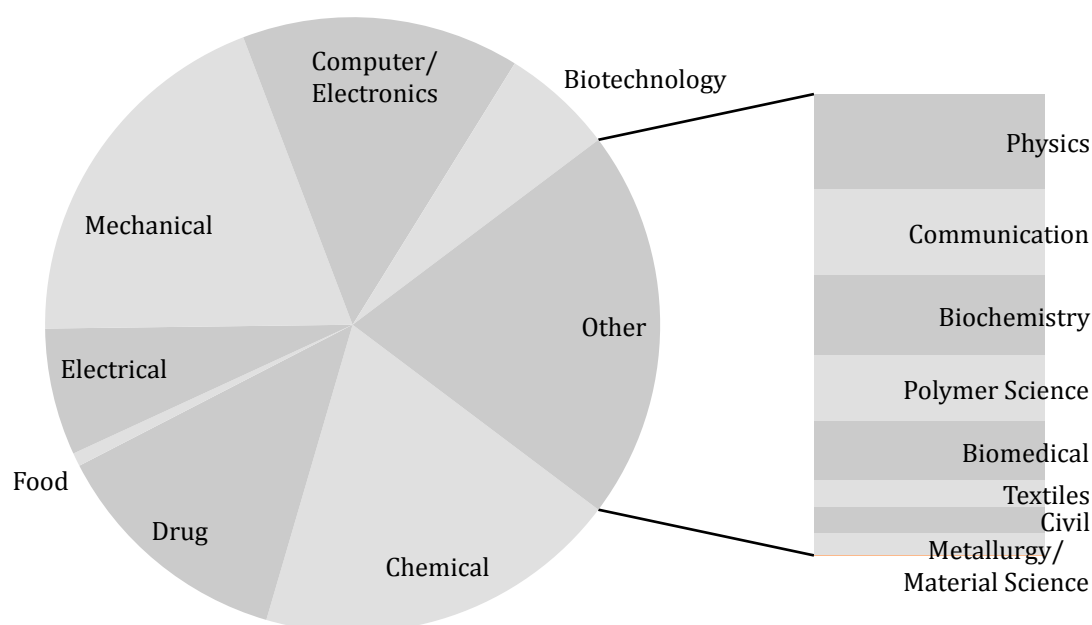
Most patent jurisdictions publish a patent application 18 months after its original filing date, after which the application is considered to be in the public domain. Before this publication, the content and the very act of filing are considered proprietary information that is closely guarded by most filers. According to the IPO's 2007-2008 Annual Report, the total patents in force, as of 31 March 2008, were 29,688, of which 7,966 comprised of patents granted to applicants residing in India. Although applications from residents within the country have increased steadily over the last 10 years or so, applications made by foreigners have accounted for the bulk of

³ These weekly journals can be accessed from: <http://www.patentoffice.nic.in/>

increase in total applications (see Figure 1). For a descriptive overview of pre-1995 patenting trends in India, see Rajeswari (1996).

Patents are classified according to the International Patent Classification system, which is technology based rather than product based (see Figure 2). This is a limitation on the economic usefulness of the data, since it does not automatically allow an analysis of innovation by industry or sector. However, I match IPC classifications with 2-digit industry level classifications, and am able to explore the impact of industry agglomeration on innovative output. The one-to-one matching is provided in Table A.1 in the Appendix.

Figure 2: Patent applications by sector (2007-2008)



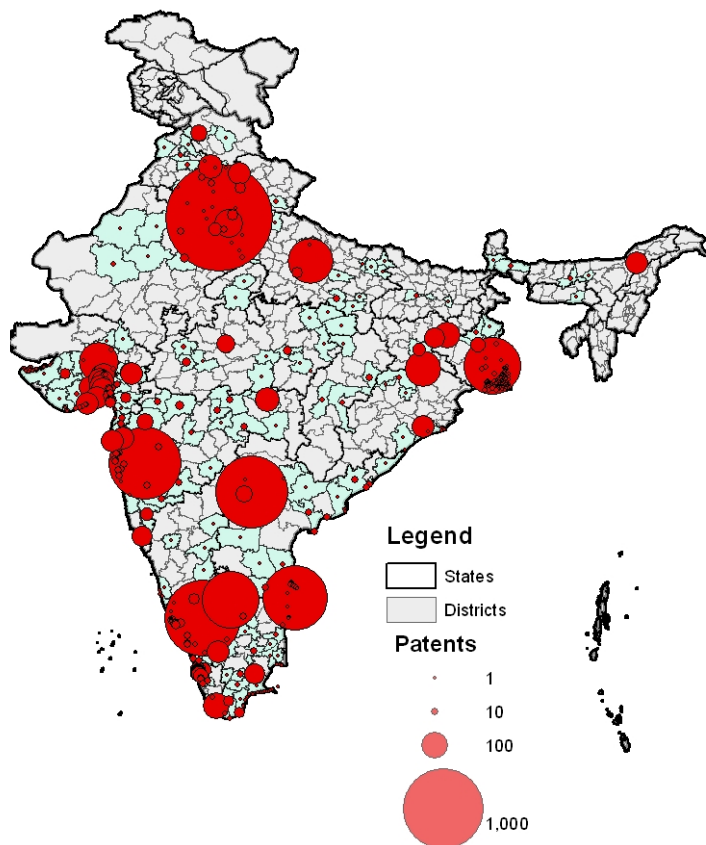
Source: IPO Annual Report (2007-2008)

In addition, I also differentiate between different types of applicants - from public and private funded research organisations, universities, and from private firms, industry and individuals. I use these applicant classes as a further robustness check later in the paper. One of the shortcomings of the IPO data is that it does not differentiate patent applications made by individual laboratories of the Council for Scientific Research (CSIR) and lists the organisation's headquarters in Delhi as the default location for all such applications. I am thus forced to drop these observations, and then re-instate them using data from the CSIR which disaggregates applications made on the basis of the location of the laboratory⁴.

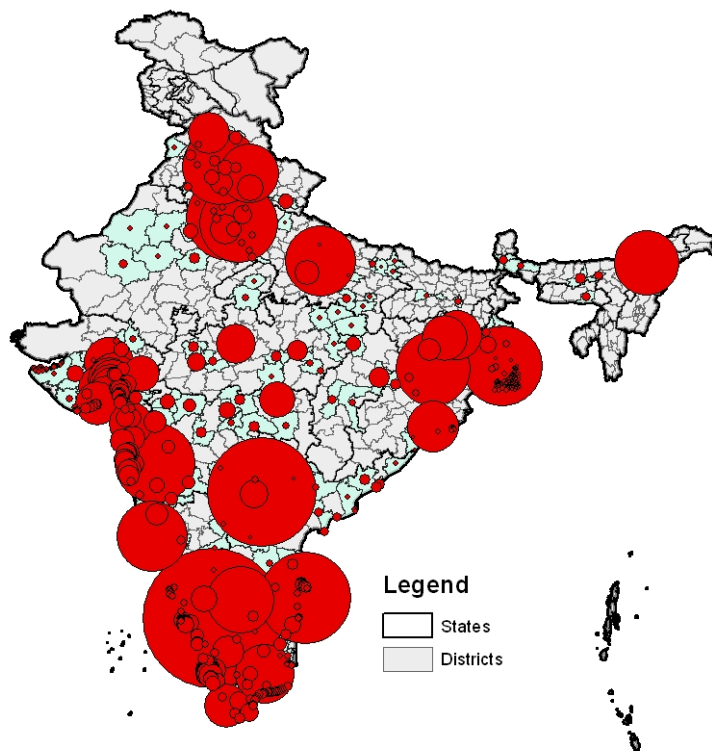
⁴ The full list of CSIR laboratories can be found on the website of the Department of Scientific and Industrial Research: http://www.dsir.gov.in/a_report/english/2005-06E/Annexures-8.pdf.

Figure 3: Spatial distribution of patents

Total



Controlling for district size



Source: FAO and IPO

Because patents contain geographic information about their inventors, I make use of this data to study the factors affecting the generation of knowledge in these locations. Patent activity in India shows signs of high levels of spatial concentration (see Figure 3). The maps provide an actual representation of the density of patent applications for the country – the size of the circle is proportional to the number of applications within the district. The left-hand-side map illustrates that whilst some districts in the country host a lot of patent activity, others are virtually empty. Of the possible 604 districts, patent activity is present in only 190 districts. The right-hand-side map carries out the same exercise, but after controlling for district population. Patent activity now shows an increase in districts in the southern and eastern parts of the country, and patent activity in districts around Delhi and Mumbai is better highlighted. However, in general, the results show that, keeping in mind the simplest no-clustering (uniform distribution) benchmark, there is evidence of concentration of patent activity in the country.

IV. Empirical Model

I test the importance of geography and human capital for knowledge spillovers by implementing an extended version of the Griliches-Jaffe regional knowledge production function at the district level in India.

At the level of the firm, such a production function assumes that there exists a stable relationship between R&D investments by a firm and the production of economically useful knowledge. However, a number of empirical studies have found an insignificant or even a negative relationship between the R&D expenditures of a firm and its innovative output. This anomaly is explained by the fact that a firm-level study is unable to take into account the spillovers received by a firm from its geographical region and that investments made in one sector can often spillover into other sectors, such that the total exceeds the sum of the individual components. The empirical link between knowledge inputs and innovative output becomes stronger with higher aggregation of the unit of observation, from the firm to the industry to the region.

What constitutes the right level of aggregation? A number of authors have pointed out that the effect of knowledge spillovers remains relatively localised and begins to diminish with increasing distance (see Jaffe et al 1993, Varga 1998, Acs et al 2002 for the United States, and Autant-Bernard 2001, Fischer and Varga 2003 for Europe). Perhaps while the marginal cost of transmitting information across geographic space has been rendered invariant by the telecommunications revolution, the marginal cost of transmitting knowledge, and especially tacit knowledge, rises with distance (Audretsch 1998). This argues for as small a geographical unit of analysis as possible. Previous studies have used 'states' (Jaffe, 1989; Smith, 1999) or Metropolitan Statistical Areas (Anselin et al., 1997, 2000), or 'districts' (Keeble and Wilkinson 1999; Piergiovanni and Santarelli 2001) as their boundaries. Some studies approached comparison and analysis at the national level (UK Department of Trade and Industry 2001), where 'cluster' refers to every technological concentration within the country (Ibrahim 2009).

If I choose the state as the unit of observation, I would be able to expand my sources of data to patent applications made to the US and the European Patent Offices from applicants based in India. Krugman (1991) has emphasised that 'states aren't really the right geographical units', and although he was referring to states within the United States, it does not make much sense to compare states like Uttar Pradesh (population: 190 million, area: 243 thousand km²) with states such as Chandigarh (population: 900 thousand, area: 114 km²) and city states such as Delhi (population: 12 million, area: 1,484 km²). Thus, a better, albeit imperfect, unit of observation in this case would be the district. This assumes that knowledge flows are bounded within a relatively narrow geographical range.

In the typical regional production function approach, the innovative output of a region depends upon the level of R&D (measured either by employment or expenditures) within the region and in neighbouring regions. Agglomeration within a location is also theorised to be a catalyst to innovation, and I include economic geography

variables to capture agglomeration at different levels. A vector of local economic characteristics that could contribute to the generation of new knowledge is also often included. These include the level of human capital, the presence of scientists and engineers, industrial diversity, the quality of technological infrastructure within the location etc. Following other empirical studies (Anselin et al 1997, Feldman and Audrestch 1999, Del Barrio-Castro and Garcia-Quevedo 2005, and Ponds 2010), this paper will use a count data model within the knowledge production framework.

Based on these specifications, the following pooled cross-section spatial model is estimated:

$$(I) \quad P_j = \beta_0 + \beta_1 RD_j + \beta_2 Access_j + \beta_3 \sigma_k + \beta_4 HI_j + \beta_5 Edu_j + \beta_6 X_j + \varepsilon_j$$

Where j indexes the districts, and k indexes the IPC-industry classification. RD_j refers to private research and development expenditures within the district, $Access_j$ refers to R&D expenditures in neighbouring districts, σ_k is a measure of agglomeration by industrial sector and district, HI_j is a measure of industrial diversity in the district, Edu_j is a measure of human capital, X_j refers to other district-level characteristics and ε is a stochastic error term.

The dependent variable is the count of new patent applications within a district. As mentioned before, these patent counts are taken from weekly journals of the Indian Patent Office. Only domestic applicants are included in the analysis, specifically when the address of the applicant is within the territory of India. The location of the patent refers to the location of the patent applicant, and not to the location of the firm – this is to avoid biases resulting from centralised patent applications from companies with multiple branches in different locations. Patents are assigned to the address of the first inventor named on the patent. The total patent applications gleaned from weekly journals total 15,782 applications made in 190 districts belonging to 35 states and union territories from 1999-2008. The choice of years of study is dictated by data limitations. Whilst annual patent application data exists for the period 1999-2008, I have data on district-level location variables for only a few years. Economic geography and infrastructure data is mostly taken from household-level surveys conducted by the NSSO. I use data from the Employment and Unemployment surveys: Round 55.10 (July 1999 – June 2000) and Round 61.10 (July 2004 – June 2005). This effectively restricts the sample to two cross sections: 1999 and 2004. The sources of data used to construct the predictor variables are summarised in Table 2.

Table 2: Predictor Variables

	Variable	Indicator	Source(s)	Availability	
				1999-2000	2004-2005
<i>R&D</i>	RD	Private R&D expenditures	Prowess, CMIE	✓	✓
<i>Economic Geography</i>	Access	Neighbouring R&D	Prowess, Orthodromic distance calculations	✓	✓
	HI	Economic Diversity	NSSO	✓	✓
	σ	Localisation Economies	NSSO	✓	✓
<i>Infrastructure</i>	Education	Proportion of population with a High-School education	NSSO	✓	✓
	Technical Education	Proportion of population with a technical diploma/degree	NSSO	✓	✓
	Telephone	Proportion of households with a telephone connection	NSSO	✓	✓
	Electricity	Proportion of population with access to electricity	NSSO	✓	✓
	Wages	Non-agricultural hourly wages	NSSO	✓	✓
	Wealth	Proportion of high-income households	NSSO	✓	✓

Notes:

NSSO: National Sample Survey Organisation

CMIE: Centre for Monitoring of the Indian Economy

Industrial R&D is measured as the expenditures of firm-level expenditures from the Prowess database, across different districts in India. Prowess is a corporate database that contains normalised data built on a sound understanding of disclosures of over 18,000 companies in India. The database provides financial statements, ratio analysis, fund flows, product profiles, returns and risks on the stock market etc. Unfortunately this does not include information on R&D contracted to universities and to research institutions.

The accessibility indicator is defined as the potential for opportunities for R&D expenditures in neighbouring regions to spill over to the location:

$$Access_j = \sum_m \frac{RD_m}{d_{j-m}^\beta}$$

Where, $Access_j$ is the accessibility indicator estimated for location j , RD_m is the R&D expenditures at destination m ⁵, d_{jm} is a measure of distance between origin j and destination m , and b describes how increasing distance reduces the expected level of interaction⁶. The exponent value is an indicator of how distance is a restrictive factor, and in the simple model, accessibility is estimated without exponent values. The accessibility measure is constructed by allowing transport to occur along the straight line connecting any two districts. Instead of calculating the distance between any pair of districts across the country, I restrict the links to districts within a 500-kilometre radius.

Localisation economies (σ_k) can be measured by own industry employment in the region, own industry establishments in the region, or an index of concentration, which reflects disproportionately high concentration of the industry in the region in comparison to the nation. I measure localisation economies as the proportion of sector k 's employment in location j as a share of all of sector k 's total employment in the country. The higher this value, the higher the expectation of intra-industry concentration benefits in the region.

$$\sigma_k = \frac{E_{k,j}}{E_k}$$

Since the data provided by the IPO lists the IPC classification, which is mainly a technological based categorisation, it was necessary to relate this to the Indian National Industrial Classification (NIC) system to enable the computation of industrial agglomeration. A broad overview of the matching is presented in Table A. 1 in the Appendix.

A higher level of industrial diversity may also translate into the presence of a wider selection of producer services essential to innovation, such as information technology, legal, marketing services etc. I use the Herfindal measure to examine the degree of economic diversity, as a measure of urbanisation in each region. The Herfindal index for region j (HI_j) is the sum of squares of employment shares of all industries in region j :

$$HI_j = \sum_k \left(\frac{E_{jk}}{E_j} \right)^2$$

Unlike measures of specialisation, which focus on one industry, the diversity index considers the industry mix of the entire regional economy. The largest value for HI_j is one when the entire regional economy is dominated by a single industry. Thus a higher value signifies lower level of economic diversity.

⁵ I am grateful to Eckhardt Bode for providing me with the syntax for computing the great-circle (orthodromic) distance calculations.

⁶ In the original model proposed by Hanson (1959), b is an exponent describing the effect of the travel time between the zones.

It is especially important to control for local inputs into the R&D process. Skilled workers endowed with a high level of human capital are a mechanism by which economic knowledge is created and transmitted. I include the share of the population with a higher education (defined as a high school degree or more) as a proxy for the general quality of human capital. However, more specifically, I also include the proportion of the population that possess a degree in a scientific subject – defined as agricultural sciences, engineering, medicine etc. Skilled workers endowed with a high level of human capital are a mechanism through which knowledge externalities materialise, and I would expect these to have a positive effect on the generation of patents.

Non-agricultural hourly wage rates are used as an indicator of labour input costs. The expected effect of this variable is hard to pin down theoretically. On the one hand, one would expect innovative activity to be inversely related to labour costs, since high costs within a location could drive down productivity. On the other hand, since I am unable, at this stage of the analysis, to differentiate wages on the basis of the skill set of workers, it is possible that there would be a positive effect since high wages are in effect accounting for the presence for highly skilled labour in the workforce.

Other district-level characteristics include variables which proxy for quality of infrastructure, in particular access to electricity and telephones. A summary of the variables used is provided in Table 3 below. Please note that in the empirical analysis, all variables have been standardised (i.e. rescaled to have a mean of zero and a standard deviation of one). Additionally, all covariates in the empirical model are lagged by one period in order to mitigate potential endogeneity concerns.

Table 3: Some Descriptive Statistics

Variable	Expected sign	#		Mean	
		1999-2000	2004-2005	1999-2000	2004-2005
Patents					
R&D	+	589	579	3.7064	10.0654
Access	+	588	578	1.8398	18.4365
Localisation	+	485	575	0.0043	0.0036
HI	-	589	579	0.3377	0.4185
Education (High-School)	+	489	579	0.0566	0.0742
Technical Education	+	487	577	0.0024	0.0112
Electricity	+	489	579	0.5608	0.6331
Telephone	+	489	579	0.0849	0.3678
Wages	-/+	486	575	93.9052	101.0757
Wealth	+	489	579	0.0551	0.0508

Notes: # refers to the number of districts for which data is available. There are a total of 604 districts in the country.

V. Results and Discussion

I illustrate the key characteristics of the data and the subsequent modelling choices, by using the 1999 cross-section as an example. One of the key characteristics of the data is that it is over-dispersed. In Table 4 below, the mean number of patent applications per district is around 6.8, the standard deviation is over 75, i.e. over 10 times the mean. A Poisson model implies that the expected count, or mean value, is equal to the variance. This is a strong assumption, and does not hold for my data.

Table 4: Characteristics of the Data (1999 cross-section)

Variable	#	Mean	Std. Dev.	Min.	Max.
count	588	6.77	75.22	0	1353
count>0	126	31.61	160.55	1	1353
obs	588	0.7857	0.41	0	1
Fit_p	485 ⁷	0.4797	0.29	0	0.94
Fit_nb	485	0.7590	0.15	0.06	0.97
Fit_zip	485	0.7636	0.20	0.00	0.98

A frequent occurrence with count data is an excess of zeroes compared to what would be expected under a Poisson model. This is indeed a problem faced by this data – the mean number of non-zero patent counts is around 31 and the standard deviation is 161, i.e. around 5.2 times the mean. Also note that of 588⁸ districts, a total of 462 districts have zero patent applications. This implies that one would need to take into account, both, over-dispersion and the excess of zeroes in the data, when selecting a model to fit the data⁹.

Another way to reiterate the unsuitability of the Poisson model in this case is to show that such a model is unable to predict the excess zeroes found in the data. In the table below, obs refers to actual observations in the data, and Fit_p, Fit_nb and Fit_zip refer to the predictions of the fitted Poisson, negative binomial and zero-inflated Poisson models respectively. It is clear that 78.57% of the locations in the sample have no patent applications, but the Poisson model predicts that only 47.97% would make zero patent applications. Clearly the Poisson model underestimates the probability of zero counts. The negative binomial model, which allows for greater variation in the count variable than that of a true Poisson, predicts that 75.9% of all districts will make no patent applications, much closer to the observed value.

One way to account for the excess zeroes would be to assume that the data comes from two separate populations, one where the number of investments is also zero, and another where the count has a Poisson distribution. The distribution of the outcome is then modelled in terms of two parameters – the probability of always zero and the mean number of patent applications for those locations not in the always zero group. As can be seen from Table 4 above, the zero-inflated Poisson model (fitzip) predicts

⁷ The number of observations are lesser than the number of cases in the dataset owing to missing values for some variables in the model.

⁸ Although there are a total of 604 districts in India, we exclude all districts for which we do not have data for the regressors.

⁹ I am unable to use a log-linear specification since it would result in a loss of all zero observations that constitute an important part of the data.

that 76.36% of all locations will not apply for patents, marginally better than the predictions of the negative binomial model.

An alternative approach to deal with an excess of zeroes would be to use a two-stage process, with a logit model to distinguish between the zero and positive counts, and then a zero-truncated Poisson or negative binomial model for the positive counts. In the case of this paper this would imply using a logit model to differentiate between districts that make no patent applications and those that do, and then a truncated model for the number of districts that apply to at least one patent. These models are referred to as “hurdle models” – a binary probability model governs the binary outcome of whether a count variate has a zero or positive realisation; if the realisation is positive, the ‘hurdle’ is crossed and the conditional distribution of the positives is governed by a truncated-at-zero count model data model (McDowell 2003)¹⁰.

The response variable is ‘count’, i.e. the number of patent applications per district. The Poisson regression models the log of the expected count as a function of the predictor variables. More formally, $\beta = \log(\mu_{x+1}) - \log(\mu_x)$, where β is the regression coefficient, μ is the expected count and the subscripts represent where the regressor, say x , is evaluated at x and $x+1$ (here implying a unit percentage change in the regressor¹¹). Since the difference of two logs is equal to the log of their quotient, i.e.

$\log(\mu_{x+1}) - \log(\mu_x) = \log\left(\frac{\mu_{x+1}}{\mu_x}\right)$, thus one could also interpret the parameter estimate as

the log of the ratio of expected counts. In this case, the count refers to the ‘rate’ of patent applications per district. The coefficients¹² could also be interpreted as incidence rate ratios (IRR), i.e. the log of the rate at which events occur. Table 5 and Table 6 provide the incidence rate ratios using different types of count models¹³.

The IRR score can be interpreted as follows – if R&D expenditures were expected to increase by a percentage unit, the rate ratio for the count of patent applications would be expected to increase by a factor of 1.137, i.e. by 13.7 percentage points (see the coefficient of R&D for the Poisson model in Table 5). More simply, an incidence rate ratio equal to 1 implies no change, less than 1 implies a decrease and more than 1 implies an increase in the rate ratio.

¹⁰ We were unable to achieve convergence when using the zero-inflated and the zero-truncated negative binomial models, and these results are excluded from the paper.

¹¹ This is because the regressors are in logarithms of the original independent variables.

¹² The coefficient results can be made available on request.

¹³ I did not achieve convergence using zero-inflated negative binomial and zero-truncated negative binomial models and these results are thus omitted from the analysis.

Table 5: Incidence Rate Ratios (1999)

Variable	Poisson	Negative Binomial	Zero-inflated Poisson	Zero-truncated Poisson
R&D	1.137***	1.085	1.114***	1.114***
Access	1.142***	1.848*	0.958***	0.958***
HI	0.358***	0.409***	0.724***	0.725***
Localisation	1.081***	1.515*	1.035**	1.035**
Education (technical)	1.079***	1.007	0.875***	0.879***
Education (high-school)	1.747***	3.112***	1.275***	1.274***
Wages	0.973	0.914	0.971	0.972
#	2910	2910	2910	312
AIC	25133.9	3525.1	16516.3	14977.6
BIC	25181.7	3578.9	16612.0	15007.5

Exponentiated coefficients

* p<0.05, ** p<0.01, *** p<0.001

Table 6: Incidence Rate Ratios (2004)

Variable	Poisson	Negative Binomial	Zero-inflated Poisson	Zero-truncated Poisson
R&D	1.271***	1.205***	1.199***	1.199***
Access	1.033	1.133	0.959*	0.958*
HI	0.391***	0.482***	0.796***	0.797***
Localisation	1.106***	1.010*	1.064***	1.064***
Education (technical)	1.429***	1.434***	1.253***	1.252***
Education (high-school)	1.583***	2.847***	1.147***	1.145***
Wages	0.598***	0.590***	0.906**	0.907**
#	3557	3557	3557	401
AIC	15811.9	4066.3	11628.9	9758.9
BIC	15861.3	4121.9	11735.6	9790.8

Exponentiated coefficients

* p<0.05, ** p<0.01, *** p<0.001

As the model selection criteria I also examine and compare the Bayesian information criterion (BIC) and Akaike's information criterion (AIC). Since the models are used to fit the same data, the model with the smallest values of the information criteria is considered better. The first result to note is that the effect of R&D expenditures remains positive and significant, irrespective of the model used. Notably, the effect of the agglomeration measure – localisation – is also positive and significant across the models. Since this measure is disaggregated by industry type, it implies that an increase in clustering within an industry is associated with a positive rise in innovative output in that industry within the district. Recall that the variable 'HI' is a measure of the level of industrial diversity within the district. A higher HI implies higher employment concentration by one industry and lower industrial diversity, and

vice versa. Thus, the negative coefficient (i.e. $IRR < 1$) for HI is evidence of a positive association between more industrial diversity and more innovation.

As a measure of human capital within a location, I also include the proportion of the district population with a high-school degree or higher, and with a technical degree (i.e. a degree in agricultural sciences, engineering, medicine etc). Whilst the coefficient of the proportion with a high school education (or higher) is positive and significant, that of technical education is positive and significant for the second cross-section. Even in the second period, the positive effect of a larger proportion of the population with a high-school degree (or higher) continues to outstrip the effect of that of a technical degree. The variable 'Education (high-school)' includes those with different levels of education. Thus, it is an indication of the innovative capacity of the educated population as a whole, including those with and without advanced degrees.

It is interesting that the effect of 'Access' is often negative and significant, although positive in few cases. A positive effect could be seen as evidence of a spillover effect of expenditures in one region positively affecting innovation in others. A negative effect, however, implies that an increase in R&D expenditures in neighbouring regions is associated with a lower level of innovative output. Since R&D expenditures here refers to the spending by private firms, this effect could be a proxy for how being located close to large firms with large R&D spending could crowd out spending in the smaller location and subsequently negatively affect patent activity in the location.

Since research in this field has previously not been carried out for India, it is impossible to compare with previous results. However, I can make some comparisons with results in other countries. Autant-Bernard (2001), Bottazzi and Peri (2003), Fischer and Varga (2003), Bode (2004) and Rodriguez-Pose and Creszenci (2008) find evidence of inter-regional spillovers, but find that these decay quickly over larger distances. It is possible that in this paper I am unable to measure spillovers at a scale lower than that of the district, and thus note only negative effects of R&D in neighbouring locations. However, Del Barrio-Castro and Garcia-Quevedo (2005) also fail to find any significant effect of R&D expenditures in neighbouring regions for Spain. Ponds et al (2010) find evidence of inter-regional spillovers, and of positive effects of human capital in the Netherlands. However, since their paper concentrates on university linkages, it is difficult to make one-to-one comparisons. Knudsen et al (2008) measure the effect of the presence of scientists and engineers for metropolitan areas in the US and are unable to find a statistically significant effect on patent activity. On the other hand, Carlino et al (2007) find a positive and significant effect of college-educated population on patents per capita in the US. For Italy, Paci and Usai (1999), find that localisation and industrial diversity have positive and significant effect on innovation. On the other hand, for the US, Feldman and Audretsch (1999) find that localisation has a negative effect on innovation, while industrial diversity has a positive effect. However, Aces et al (2002) find that the localisation (labelled location quotient in their paper) has no effect on patent activity. A number of these and other papers also measure the effects of university research on a location's patent activity and find evidence of positive effects – however, owing to lack of data I am unable to analyse the effects of university research in this paper.

At this stage the paper does not claim causality – potentially a number of variables have been omitted and reverse causality biases cannot be ruled out. Nevertheless, the

associations provide some interesting results. The next section addresses some of these concerns and verifies the robustness of the results.

VI. Endogeneity Issues and Robustness Checks

Although all the regressors have been lagged, there could remain endogeneity concerns that would bias the coefficients (or, in this case, the reported incidence rate ratios). One potential source of endogeneity is the reverse causality between the rate of innovation and R&D expenditures. If past innovative activity within a location affects the distribution of R&D, this would lead to biased estimates of β . A second source of bias could be the possibility of variables omitted from the model which could lead to biased estimates of β . To address the possible sources of bias, I perform instrumental variable regressions and examine Durbin and Wu-Hausman test statistics for endogeneity bias.

An ideal instrument to control for reverse causality would influence innovative output only through the distribution of R&D across locations. For instance, the average value of a patent in a location would serve as a good instrument since it would affect innovative output only through its effect on the distribution of R&D expenditures. The average value of a patent would depend on the market for new innovations within a location, and unfortunately I am unable to directly observe this value. I follow Bottazzi and Peri (2003), who in turn follow Ciccone and Hall (1996), and use historical population density as a proxy for the market potential of an invention, which in turn determines the average value of a patent within that location. Early patterns of population density are correlated with market potential, since they continue to effect current population distributions. But they are not expected to have any significant effect on current innovative output, except through their influence on market potential.

There is also a second source of endogeneity bias, in that unobservables in the model, such as some underlying features of the natural geography, could be jointly influencing agglomeration and innovation within a location. A good instrument would allow me to isolate the effect that runs from agglomeration to innovation and account for the omitted variables bias. I follow Lall and Mengistae (2005) and use historic land revenue institutions set up by the British as instruments. Land revenue was the most important source of government revenue and the British instituted three systems defining who was responsible for paying the land taxes. These were (a) landlord based systems (zamindari), (b) individual cultivator-based systems (ryotwari) or (c) village-based systems (mahalwari). These institutions serve as good instruments since the choice of land tenure system was exogenous to regional attributes¹⁴, and because they have been found to have a lasting impact on industrial agglomeration and development (Bannerjee and Iyer 2005). Thus, land revenue institutions affect innovative output only through agglomeration. I link Bannerjee and Iyer's (2005) land revenue classification with the 1991 district boundaries and code the cities according to if the district had a landlord-based system or a village/cultivator-based system.

¹⁴ It was based on the preferences of individual administrators and not on a systematic evaluation of region-specific characteristics.

I run the instrumental variable estimation within a count data model (Mullahy 1997) using a Stata module for IV/GMM Poisson regression (Nichols 2007). I also run a linear regression with an IV specification, and an alternative generalised linear model (GLM) (Hardin et al 2003) to check for the strength of the instrument and to address endogeneity concerns due to measurement errors.

I report the results of these specifications in the tables below, together with the results of diagnostics. Only the results for the instrumented co-efficient are presented here. The tests confirm the validity of the IV specification and the strength of the instrument when the R&D coefficient is instrumented with historical population density. The instrumented coefficient remains positive and significant, and the F-statistic is well above the rule-of-thumb value of 10. I also perform the Durbin-Wu-Hausman test to examine if endogeneity of R&D expenditures could have adverse effects on OLS estimates, and find that the results of the IV estimates are preferable. On the other hand, I am unable to rule out omitted variables bias, since the instrumented coefficient (localisation) is no longer significant and the F-statistic is also below the rule-of-thumb value.

Table 7: IV Estimation 1999 (Instrumented coefficient: R&D)

Variables	IV Poisson	IV Regression	AGLM (Poisson)	AGLM (Negative Binomial)
R&D	1.181*	1.258***	1.300***	1.837**
<i>Other controls</i>	<i>yes</i>	<i>yes</i>	<i>yes</i>	<i>yes</i>
#	2383	299	2257	2257
R^2		0.387		
AIC		964.9		
BIC		994.5		
F-Stat		193.955		

Exponentiated coefficients

* p<0.05, ** p<0.01, *** p<0.001

Table 8: IV Estimation 2004 (Instrumented coefficient: R&D)

Variables	IV Poisson	IV Regression	AGLM (Poisson)	AGLM (Negative Binomial)
R&D	1.21*	1.288***	1.456***	1.829***
<i>Other controls</i>	<i>yes</i>	<i>yes</i>	<i>yes</i>	<i>yes</i>
#	2441	360	2316	2316
R^2		0.467		
AIC		1098.2		
BIC		1129.3		
F-Stat		184.175		

Exponentiated coefficients

* p<0.05, ** p<0.01, *** p<0.001

Table 9: IV Estimation 1999 (Instrumented coefficient: Localisation)

Variables	IV Poisson	IV Regression	AGLM (Poisson)	AGLM (Negative Binomial)
Localisation	0.401	1.321	0.979	1.053
<i>Other controls</i>	<i>yes</i>	<i>yes</i>	<i>yes</i>	<i>yes</i>
#	1763	238	1763	1763
R^2		0.151		
AIC		863.8		
BIC		891.6		
F-Stat		2.931		

Exponentiated coefficients

* p<0.05, ** p<0.01, *** p<0.001

Table 10: IV Estimation 2004 (Instrumented coefficient: Localisation)

Variables	IV Poisson	IV Regression	AGLM (Poisson)	AGLM (Negative Binomial)
Localisation	0.642*	3.988	0.959	0.537*
<i>Other controls</i>	<i>yes</i>	<i>yes</i>	<i>yes</i>	<i>yes</i>
#	1938	279	1938	1938
R^2		0.603		
AIC		1198.1		
BIC		1227.1		
F-Stat		2.645		

Exponentiated coefficients

* p<0.05, ** p<0.01, *** p<0.001

As a robustness check, I also differentiate between different types of applicants - from government-funded research organisations, private-funded research organisations, industry and individuals. For each cross-section, I disaggregate the data according to the type of applicant and re-run the regressions to observe if the effects of the predictor variables vary across each group. Although I carry out the regressions using Poisson and zero-inflated and zero-truncated methods as well, I only report the results of the negative binomial specifications¹⁵. This is to facilitate comparison, but more importantly because the negative binomial models exhibit the best goodness-of-fit statistics. As before, the coefficients are reported as Incidence Rate Ratios for ease of interpretation.

¹⁵ Results from the other models are available on request.

Table 11: Robustness Checks (1999)

Variable	Government-funded	Private-funded	Firms	Individuals
R&D	1.02	1.256***	1.211***	1.700***
Access	0.614**	0.786***	0.846**	1.069***
HI	6.010***	0.651*	0.677***	0.932
Education (technical)	10.60***	0.235***	0.783***	1.697***
Education (high-school)	1.473**	1.555***	3.136***	1.03
Wages	0.485***	0.884**	1.851***	0.863**
N	421	241	569	1496
AIC	3936.3	1290.4	3443.1	6247
BIC	3968.6	1318.2	3477.8	6289.5

Exponentiated coefficients

* p<0.05, ** p<0.01, *** p<0.001

Table 12: Robustness Checks (2004)

Variable	Government-funded	Private-funded	Firms	Individuals
R&D	1.376***	1.169***	1.207***	1.518***
Access	0.708***	0.797*	0.612***	1.133***
HI	1.531***	0.932	0.818*	0.985
Education (technical)	1.091**	1.252***	3.444***	1.050**
Education (high-school)	1.017	1.071	1.725***	1.247***
Wages	0.790*	1.222	0.813**	1.003
N	340	223	547	2164
AIC	2882.8	810.5	3789.3	8992.4
BIC	2913.5	837.7	3823.7	9037.9

Exponentiated coefficients

* p<0.05, ** p<0.01, *** p<0.001

Interestingly, while most coefficients remain significant, there are a few changes with regard to the direction and the size of the effect. Recall that R&D represents only private R&D, and it is worth noting that it continues to have a positive and significant effect on government-funded research¹⁶, although this effect drops off in the earlier cross-section. Much more notably, the effect of the variable ‘Access’, i.e. R&D funding in neighbouring locations (defined as districts within a 500 kilometer radius) now has a negative (i.e. IRR<1) and significant effect on public and private-funded, and industry innovative output, although it continues to have a positive effect on individuals’ innovative output. This could be explained by the tendency of large regions to absorb talent and funding – in other words, being situated close to a regions with higher R&D funding could lead to positive effects on individuals’ propensity to innovate, but may crowd out the same for private industry and organisational

¹⁶ Government-funded research includes both, research carried out within government departments and within government-funded research institutions.

research. The positive effect of the presence of human capital consisting of technically trained individuals continues to be significant and large across the various groups. Lastly, whilst the effect of wages on innovative output is negative, i.e. lower non-agricultural average wages tend to be associated with higher levels of innovative activity (perhaps by contributing to overall productivity through cost reductions), the converse is true for private industry. In other words, wages are positively associated with patent activity. Since I am unable to proxy for the skill-structure of employees, this effect could be capturing the wage premium paid to highly skilled employees in private firms and industry.

VII. Conclusion

This paper contributes to the empirical literature on the geography of innovation in a number of ways. There is little research that has looked at emerging countries, and this is the first study of what drives innovative activity in India. Secondly, responding to the shortcomings of regions or states as the unit of observation, the spatial scale of the study goes down to the level of districts, which in a number of cases correspond to city-level boundaries in India. And lastly, the study finds robust evidence to link the agglomeration of economic activity to innovative activity, and uses instrumental variables to establish causality.

The findings of this study could have important implications for policy. Firstly, the effect of private, firm-level R&D has a positive effect on not just industry innovation, but also spills over into patenting activity carried out by private individuals and even research organisations. While the effect of government R&D has not been studied in this paper, it would be interesting to see to what extent the public sector could further spur innovation. Secondly, education matters. This may be self-evident, but it is worth reiterating. A better-educated labour force increases the ability of a region to innovate, and investing in education would be an efficient means to spur on lagging regions. Lastly, and importantly, public policy may be limited in its ability to encourage innovation. The results indicate that agglomeration economies continue to have a significant effect on innovative activity – this result holds across the two cross-sections and is robust to the type of innovator being examined. While government policy can have a direct effect on improving education and research infrastructure, its capacity to generate agglomeration economies remains unclear.

The study has a few limitations. The analysis of patent applications made to the US and the European Patent Offices are not included. The analysis is also based on patent data, which certify new inventions, and thus innovation citations are excluded. As a number of authors have warned (Griliches 1990, Mansfield 1984, Scherer 1983 – quoted in Feldman and Florida 1994) that the number of patents is not directly equivalent to a measure of innovative output as many patented inventions never become commercially viable products while many successful products are never patented. Nonetheless patents continue to be used as a useful measure of the generation of ideas. In addition the empirical analysis is based on cross-sectional data, albeit at two points in time. Although the time lags between the dependent and independent variables lower the risk of endogeneity, and this paper carries out

instrumental variable estimations to control for simultaneity biases, future studies using panel data are necessary to come to more decisive conclusions.

Bibliography

- Acs, Z. J., Anselin, L., Varga, A. (2002) 'Patents and innovation counts as measures of regional production of new knowledge', *Research Policy* 31: 1069–1085.
- Anselin, L., Varga, A., Acs, Z. (1997) 'Local geographic spillovers between university research and high technology innovations', *Journal of Urban Economics* 42: 422-448.
- Anselin, L., Varga, A., Acs, Z. (2000) 'Geographic and sectoral characteristics of academic knowledge spillovers', *Papers in Regional Science* 79: 435–445.
- Audretsch, D.B. (1998) 'Agglomeration and the location of innovative activity', *Oxford Review of Economic Policy* 14(2): 18-29.
- Audretsch, D.B and Feldman, M.P. (1996) 'R&D spillovers and the geography of innovation and production', *The American Economic Review* 86(3): 630-640
- Autant-Bernard, C. (2001) 'Science and knowledge flows: evidence from the French case', *Research Policy* 30: 1069–1078.
- Barkey, D.L., Henry, M.S. and Lee, D. (2006) 'Innovative activity in rural areas: the importance of local and regional characteristics', *Community Development Investment Review*, Federal Bank of San Francisco.
- Bode, E. (2004) 'The spatial pattern of localized R&D spillovers: an empirical investigation for Germany', *Journal of Economic Geography* 4: 43-64
- Bottazzi, L. and Peri, G. (2003) 'Innovation and spillovers in regions: evidence from European patent data', *European Economic Review* 47: 687-710.
- Carlino, G.A., Chatterjee, S. and Hunt, R.M. (2007) 'Urban density and the rate of invention', *Journal of Urban Economics* 61: 389-419.
- Ciccone, A. and Hall, R.E. (1996) 'Productivity and the density of economic activity', *The American Economic Association* 86(1): 54-70.
- Co, C.Y. (2005) 'Co-evolution of invention activities among cities in New England', *Papers in Regional Science* 85(1): 47-71.
- Del Barrio-Castro, T. and Garcia-Quevedo, J. A. (2005) 'Effects of university research on the geography of innovation', *Regional Studies* 39: 1217–1229.
- Doring, T. and Schnellbach, J. (2006) 'What do we know about geographical knowledge spillovers and regional growth? A survey of the literature', *Regional Studies* 40: 375–395.
- Feldman, M.P., and Audrestch, D.B. (1999) 'Innovation in cities: science-based diversity, specialization and localized competition', *European Economic Review* 43: 409-429.
- Feldman, M.P. and Florida, R. (1994) 'The geographic sources of innovation: technological infrastructure and product innovation in the United States', *Annals of the Association of American Geographers* 84 (2): 210-229.
- Fischer, M. and Varga, A. (2003) 'Spatial knowledge spillovers and university research: evidence from Austria', *Annals of Regional Science* 23: 303–322.
- Fritsch, M. and Slavtchev, V. (2007) 'Universities and innovation in space', *Industry and Innovation* 14: 201–218.
- Girma, S., Gong, Y. and Goerg, H. (2007) 'Foreign direct investment, access to finance, and innovation activity in Chiene enterprises', *Research Paper Series, China and the World Economy*, University of Nottingham.

- Hanson, G. H. (2001) 'Scale economies and the geographic concentration of industry', *Journal of Economic Geography* 1: 255–276.
- Ibrahim, S.E., Fallah, M.H. and Reilly, R.R. (2009) 'Localised sources of knowledge and the effect of knowledge spillovers: an empirical study of inventors in the telecommunications industry', *Journal of Economic Geography* 9: 405-431.
- Jaffe, A.B. (1986) 'Technological opportunity and spillovers of R&D: Evidence from firms' patents, profits and market value', *The American Economic Review* 76 (5): 984-1001.
- Jaffe, A.B. , Trajtenberg, M. and Henderson, R. (1993) 'Geographic localization of knowledge spillovers as evidenced by patent citations', *The Quarterly Journal of Economics* 108 (3): 577-598.
- Kantor, S. and Whalley, A. (2009) 'Do universities generate agglomeration spillovers? Evidence from endowment value shocks', Working Paper 15299, National Bureau of Economic Research, Cambridge, MA.
- Keeble, D. and Wilkinson, F. (1999) 'Collective learning and knowledge development in the evolution of regional clusters of high technology SMEs in Europe', *Regional Studies* 33: 295–303.
- Kelly, M. and Hageman, A. (1999) 'Marshallian externalities in innovation', *Journal of Economic Growth* 4 (1): 39-54.
- Kerr, W.R. (2009) 'Breakthrough inventions and migrating clusters of innovation', Working Paper 15443, National Bureau of Economic Research, Cambridge, MA.
- Knudsen, B., Florida, R., Stolarick, K. and Gates, G. (2008) 'Density and creativity in US regions', *Annals of the Association of American Geographers* 98(2): 461-478.
- Krugman, P. (1991) *Geography and Trade*. Cambridge: MIT Press.
- Menon, C. (2009) 'Stars and comets: an exploration of the patent universe', Discussion Paper 37, Spatial Economics Research Centre, London.
- Mullahy, J. (1997) 'Instrumental-variable estimation of count data models: applications to models of cigarette smoking behaviour', *The Review of Economics and Statistics* 79(4): 586-593.
- Paci, R. and Usai, S. (1999) 'Externalities, knowledge spillovers and the spatial distribution of innovation', *GeoJournal* 49: 381.
- Piergiovanni, R. and Santarelli, E. (2001) 'Patents and the geographic localization of R&D spillovers in French manufacturing', *Regional Studies* 35: 697–702.
- Ponds, R., van Oort, F. and Frenken, K. (2010) 'Innovation, spillovers and university-industry collaboration: an extended knowledge production function approach', *Journal of Economic Geography* 10: 231-255.
- Rajeswari, A.R. (1996) 'Indian patent statistics – an analysis', *Scientometrics* 36(1): 109-130.
- Rodriguez-Pose, A. and Crescenzi, R. (2008) 'R&D, spillovers, innovation systems and the genesis of regional growth in Europe', *Regional Studies* 42 (1): 51–67.

Appendix

Table A. 1: IPC and NIC Matching

IPC	IPC definition	NIC	NIC Definition
		1	Agriculture, hunting and forestry
		2	Forestry, logging and related service activities
			Fishing, operation of fish hatcheries and fish farms;
		5	service activities incidental to fishing
		15	Manufacture of food products and beverages
		16	Manufacture of tobacco products
			Manufacture of wearing apparel; dressing and
		18	dyeing of fur
			Tanning and dressing of leather; manufacture of
		19	luggage, handbags, saddlery, harness and footwear
		36	Manufacture of furniture; manufacturing N.E.C.
A	Human Necessities	85	Health and social work
			Manufacture of wood and of products of wood and
			cork, except furniture; manufacture of articles of
		20	straw and plaiting materials
			Publishing, printing and reproduction of recorded
		22	media
			Manufacture of motor vehicles, trailers and semi-
		34	trailers
		35	Manufacture of other transport equipment
		37	Recycling
			Sale, maintenance and repair of motor vehicles and
		50	motorcycles; retail sale of automotive fuel
		60	Land transport; transport via pipelines
		61	Water transport
		62	Air transport
			Supporting and auxiliary transport activities;
		63	activities of travel agencies
B	Performing operations and transporting	64	Post and communications
			Tanning and dressing of leather; manufacture of
		19	luggage, handbags, saddlery, harness and footwear
			Manufacture of coke, refined petroleum products
		23	and nuclear fuel
		24	Manufacture of chemicals and chemical products
		25	Manufacture of rubber and plastic products
		26	Manufacture of other non-metallic mineral products
C	Chemistry and Metallurgy	27	Manufacture of basic metals
			Manufacture of textiles
D	Textiles and Paper	17	Manufacture of textiles
		21	Manufacture of paper and paper products
E	Fixed Constructions	10	Mining of coal and lignite; extraction of peat
			Extraction of crude petroleum and natural gas;
			service activities incidental to oil and gas extraction
		11	excluding surveying

		12	Mining of uranium and thorium ores
		13	Mining of metal ores
		14	Other mining and quarrying
		41	Collection, purification and distribution of water
		45	Construction
			Sewage and refuse disposal, sanitation and similar activities
		90	
			Manufacture of fabricated metal products, except machinery and equipments
		28	
		29	Manufacture of machinery and equipment N.E.C.
F	Mechanical Engineering		Manufacture of electrical machinery and apparatus N.E.C.
		31	
			Manufacture of office, accounting and computing machinery
		30	
			Manufacture of radio, television and communication equipment and apparatus
		32	
			Manufacture of medical, precision and optical instruments, watches and clocks
		33	
			Financial intermediation, except insurance and pension funding
		65	
G	Physics		Computer and related activities
		72	
H	Electricity	40	Electricity, gas, steam and hot water supply

Notes: IPC: International Patent Classification

NIC: National Industrial Classification