SERC DISCUSSION PAPER 72

Industry and the Urge to Cluster: A Study of the Informal Sector in India

Megha Mukim (Department of International Development, London School of Economics)

March 2011

This work is part of the research programme of the independent UK Spatial Economics Research Centre funded by the Economic and Social Research Council (ESRC), Department for Business, Innovation and Skills (BIS), the Department for Communities and Local Government (CLG), and the Welsh Assembly Government. The support of the funders is acknowledged. The views expressed are those of the authors and do not represent the views of the funders.

Industry and the Urge to Cluster:
A Study of the Informal Sector in India

Megha Mukim*

March 2011

* Department of International Development, London School of Economics

Acknowledgements
I am grateful to Diana Weinhold, Holger Görg, Henry Overman, Waltraud Schelkle, Maarten Bosker, Harry Garretson, Walker Hanlon and Stephen Gibbons for helpful comments and suggestions. I am also thankful to Karan Singh for helping me with the data.

Abstract
This paper studies the determinants of firm location choice at the district-level in India to gauge the relative importance of agglomeration economies vis-à-vis good business environment. A peculiar characteristic of the Indian economy is that the unorganised non-farm sector accounts for 43.2% of NDP and employs 71.6% of the total workforce. I analyse National Sample Survey data that covers over 4.4 million firms, in both unorganised sectors – manufacturing and services. The empirical analysis is carried out using count models, and I instrument with land revenue institutions to deal with possible endogeneity bias. I find that buyer-suppler linkages and industrial diversity make a district more attractive to economic activity, whilst the quality and level of infrastructure are also important. I conclude that public policy may be limited in its ability to encourage relocation of informal firms.

Keywords: agglomeration economies, informal sector, location choice
JEL Classifications: R12, R3, O17

1 Introduction

The informal sector[1] is an important means of livelihood to millions of people in developing countries. Because of its very nature – it is unregulated by government – data collection and subsequent analysis lags far behind that for the formal sector. In India, the informal sector often falls outside the scope for planned development efforts, and thus remains in the shadows with regard to productivity, social security and statistics.

This paper is a first attempt to understand the forces that drive the clustering of informal sector activities in India. It studies how new firms within the Indian unorganised sector choose to locate themselves across districts[2] in the country. Using count models it carries out an empirical test of the decisions of individual firms. In the model, firms compare potential profitability as a function of observable location specific advantages, market access, agglomeration economies and a set of unobserved local attributes of the district. And so, to unpack the location decisions of unorganised sector firms, an econometric analysis of location patterns is carried out to identify the 'revealed preferences' of firms. Firm-level data for the unorganised sector is taken from surveys conducted by the National Sample Survey Organisation (NSSO), which includes information on the number and type of new firms within each district.

It is important to test whether individual firm's decisions are based on agglomeration economies, or on other factors, such as good business environment – the latter being more amenable to change by policy than the former. In theory, if government is interested in encouraging industrial growth in particular regions, it should have a clear understanding of what factors drive firm location decisions. There

[1] A number if countries, including India, often use the terms 'unorganised sector' and 'informal sector' interchangeably.

[2] India is a federal union of 28 states and 7 union territories, which are further sub-divided into 604 districts.

are a few papers that have analysed the case of manufacturing firms in India (see Lall et al 2004, Lall and Chakravorty 2005). However, these studies concern themselves primarily with the formal sector. To the author's knowledge, there has been no previous research that sheds any light on what factors attract smaller, unorganised sector firms to a location. Since the informal sector in India is a significant source of employment (32%) and economic growth (22.6%) in the non-farm sector, there remains a yawning gap in the empirical understanding of the country's industrial location choices.

While the results of the analysis provide an understanding of what drives clustering in informal industries in India, they also add to a rapidly growing body of empirical evidence that tests the theoretical implications of Krugman's economic geography. This paper finds that agglomeration economies have a significant effect on firms' location decisions, and that the ability of incremental policy reforms to counter the effects of geography may be limited. In the case of the unorganised sector, geography could indeed be destiny. However, the paper does not predict how interaction between the forces of agglomeration and good infrastructure might ultimately affect the distribution of economic activity across the country.

The paper is organised as follows. The next section provides a descriptive overview of the clustering of informal sector activity, in both the manufacturing and services sectors. Section 3 starts with a theoretical explanation of the factors influencing the location of economic activity, and presents evidence of how these theories have been tested empirically in the literature. This section also provides an overview of how agglomeration economies may be different for services as compared to manufacturing. Section 4 lays out the estimation framework and discusses the main sources of data. Section 5 presents the results of the model. Section 6 describes the

identification strategy employed. Section 7 concludes and discusses the implications of the findings.

2 Descriptive Analysis

The unorganised sector in India refers to those enterprises whose activities or collection of data is not regulated under legal provision and/or which do not maintain regular accounts. These enterprises are not registered under the Factories Act of 1948. The Act requires all firms engaged in manufacturing to register if they employ 10 workers or more and use power, or if they employ 20 workers or more. Thus, it can be reasonably assumed that all privately-owned manufacturing enterprises meeting these two criteria are said to be in the unorganised sector. All public sector enterprises are automatically assumed to be in the organised sector. Services enterprises are not required to register under the Factories Act (unless they happen to also be engaged in manufacturing activities), and thus, most privately owned services firms are officially classified as being in the unorganised sector. Later in the paper, I analyse enterprises by size to try and control for this problem of definition of what constitutes as unorganised for services firms.

The terms ‘unorganised’ and ‘informal’ sector enterprises are used interchangeably in this paper; however, the latter are a subset of the former. The informal sector comprises mainly of unincorporated proprietary or partnership enterprises, while the unorganised sector includes the same along with cooperative societies, trusts and private limited companies.

The unorganised sector in India continues to occupy a substantial place in the country's economy. Its share in the country's Net Domestic Product (NDP) was 56.7% in 2002-03. The importance of the unorganised sector differs substantially across farm and non-farm activities. For instance, in the same year, its share of agricultural NDP was a whopping 96%, and its share of manufacturing and services NDP was 39.5% and 46.9% respectively.

The unorganised sector's total NDP contribution can be broken down into its services (43.2%) and manufacturing (16.8%) components. Manufacturing enterprises are often registered because they require more licenses and need access to more infrastructure and capital. On the other hand, service activities can be undertaken without many of these pre-requisites.

The importance of the unorganised sector is even starker with regards to employment. In 2004-05, the unorganised sector was a source of livelihood to approximately 86.3% of the country's workforce. Although a large section of the unorganised sector works within agricultural activities, it is pertinent to note that 71.6% of the total employment in the non-farm sector was also unorganised. In other words, although the unorganised sector contributes just over half of the country's NDP, it employs almost 90% of its workforce.

The contribution of the unorganised sector to employment has also remained broadly stable over the last few decades, with that of the formal sector rising very slowly over time. Informal agricultural employment has barely budged around the 99.4 percent mark. In fact the proportion of unorganised sector employment has risen for all these sectors, especially for services and manufacturing by a few percentage points over the period of study (1983-84 to 1999-2000). Sectors like electricity, gas and water supply, and transport and communication have also experienced rapid

informalisation of the their workforce. In other words, the dominance of unorganised employment in the country shows no signs of abating (see Table 3).

Over the last decade, there has been much interest in studying the location and the geographic concentration of economic activity. The clustering of economic activity has important implications for development, through its effect on employment and growth. The Government of India has focussed much attention on trying to encourage industrial activity in secondary cities or to areas where such activity has not previously clustered or even favoured. This effort has been focussed on organised sector activity. And even though the unorganised sector is of critical importance to the economy, there is almost no understanding of what attracts these activities to locations.

Before studying the impact of various factors affecting the location of unorganised firms, I will establish that both sectors, manufacturing and services, show evidence of spatial clustering[3] across different districts in India. A study of what drives spatial concentration of economic activity can only be interesting if such patterns exist in the first place.

There are many methods to ascertain whether firms are uniformly distributed across various locations or if they show patterns of spatial concentration. Clustering in its simplest forms can be shown graphically, or through a bird's eye view of where industry in located by means of geographical maps. Figure 2 provides an actual representation of firm density for the country – the size of the circle is proportional to the number of new informal firm births within the district. The total number of new informal manufacturing and services units exceeded 2 million respectively. The first map illustrates that whilst some districts in the country host a lot of new unorganised

[3] Clustering is a phenomenon in which events or artefacts are not randomly distributed over space, but tend to be organised into proximate groups.

economic activity, others are virtually empty. Also firm births tend to cluster in the same geographical districts, albeit with some differences depending on the type of sub-sector. There are 604 districts in the country, of which informal manufacturing firms are present in 578 districts, and of these around 39 districts account for 50% of all economic activity. On the other hand, informal services firms are present in 556 districts. Of these, around 60 districts account for 50% of all economic activity. In other words, new informal activity is highly concentrated within a few districts in the country.

Of course one could argue that clustering in these districts is simply a factor of the size of the district. And so, the next set of maps carries out the same exercise, but after controlling for the area of the district (in km^2), district population and distance from the coast – and the results show that, keeping in mind the simplest no-clustering (uniform distribution) benchmark, there is evidence of concentration of economic activity in the country. After adding controls, clustering moves from particular districts to clusters of districts. In other words, the per capita rate remains high for the densely populated districts and for their neighbouring districts (see Figure 3).

I also calculate the Theil index for the distribution of new firms for the manufacturing and services' sectors. The Theil index belongs to the family of generalised entropy inequality measures. The values vary between 0 and ∞, with zero representing an equal distribution and higher values representing higher values of inequality[4]. Figure 4 shows the contribution to the Theil index by district. These results correspond closely to the visual clustering presented in the maps. In other words, districts such as Mumbai, Delhi, Kolkata, Bangalore, Hyderabad, Ahmadabad,

[4] The value of the index increases in the inequality of the distribution of firm births by district with respect to total firm births: $T = \frac{1}{N}\sum_{j=1}^{N}\left(\frac{x_j}{\bar{x}}.\ln\frac{x_j}{\bar{x}}\right)$, where x_j is the number of firm births in district j.

Thane, Pune etc are agglomerated even after using different descriptive techniques to control for district-specific characteristics and for the size and the distribution of firms across districts.

Although maps provide a convenient visual representation of the location of new economic activity, more detailed statistics are required to ascertain if there is any evidence of clustering. If economic activity of a particular industry is biased towards a subset of regions, then the industry is said to be 'concentrated'; and if economic activity of a particular region is biased towards a subset of industries, the region is said to be 'specialised'. I use the Theil index to study what regions are specialised, and the Ellison-Glaeser Index[5] to study concentration across industries (see Appendix for construction of these indices). The Theil Index here provides an indication of the over or under-representation of district across a set of given industries, i.e. the distribution of new firms by NIC sector across districts. The results are provided in Table 4, separately for manufacturing and services sectors for district with the most clustering. Again, districts such as Delhi, Mumbai, Kolkata, Bangalore, Hyderabad etc continue to dominate.

Table 5 and Table 6 in provide the Ellison-Glaeser Indices for the two sectors across districts. The EG Index has the property of controlling simultaneously for the employment distribution among firms and regions. In their paper, Ellison and Glaeser (1997) demonstrate that the index takes the value of zero under the null hypothesis of random location conditional on the aggregate manufacturing employment in that region. In other words, the no-agglomeration benchmark is when the value of the index is zero (i.e. $E(\gamma) = 0$). In general, if the EG index is greater than 0.05, the industry is considered to be highly concentrated. I find that manufactures of office,

[5] Duranton and Overman (2005) use a more distance-sensitive measure of concentration. I am unable to estimate their index owing to lack of micro-data on firm location.

accounting and computing equipment, transport and communications equipment, and that of leather products, among others is highly concentrated in a few districts. Services related to research and development, computers and supporting transport and other activities also shows evidence of much concentration.

Having established that there is overwhelming evidence of clustering in unorganised industry across different districts in India, this paper will examine the factors that drive such clustering. In particular it will focus on identifying the role of agglomeration economies in influencing the decision of firms to cluster, i.e. to locate close to one another. It will examine the nature and scale of agglomeration economies using district and NIC 2-digit-level data for unorganised firms in India.

3 Theoretical background and Literature

This section will provide a brief overview of the theoretical understanding of agglomeration economies and outline a few empirical studies of relevance. For an excellent overview of the location theory, see Brulhart (1998) (Table 1, Page 778) that describes the different theoretical schools and lists their principal distinguishing features. Marshall (1919) was the first to identify the benefits from industrial clustering. Clusters of firms, predominantly in the same sector, could take advantage of localisation economies, such as the sharing of sector-specific inputs, skilled labour and knowledge. Thus, cost-saving externalities are maximised when a local industry is specialised. The Marshall-Arrow-Romer (Marshall 1890, Arrow, 1962, Romer 1986) models predict that such externalities predominantly occur within the same

industry. Therefore, if an industry is subject to localisation externalities, firms are likely to locate in a few regions where other firms that industry are already clustered.

The next level is that of inter-industry clustering[6], i.e. when firms in a given industry and those in related industries agglomerate in a particular location. The benefits of clustering would include inter-industry linkages, buyer-supplier networks, and opportunities for efficient sub-contracting. Venables (1996) demonstrates that agglomeration could occur through the combination of firm location decisions and buyer-supplier linkages, since the presence of local suppliers could reduce transaction costs and increase profitability. Inter-industry linkages can also serve as a channel for vital information transfers.

An overall large size of the urban agglomeration and its more diverse industry mix is also thought to provide external benefits beyond those realised within a single sector or due to a tight buyer-supplier network (Henderson 2003). Chinitiz (1961) and Jacobs (1969) proposed that important knowledge transfers primarily occur across industries and the diversity of local industry mix is important for these externality benefits. These benefits are typically called urbanisation economies and include access to specialised financial and professional services, availability of a large labour pool with multiple specialisations, inter-industry information transfers and the availability of less costly general infrastructure. Larger cities also provide a larger home market for end products, make it easier to attract skilled employees. Other factors that make big cities more attractive are urban amenities not available in smaller towns and a large number of complementary service providers such as financial and legal advisers, advertising and real estate services etc.

[6] As Deichmann et al (2005) points out, empirically the distinction between own-industry versus cross-industry is dependent on the level of sectoral aggregation.

Thus, industrial clustering could take place at different levels, which would have different implications for the associated agglomeration economies. A firm could gain from economies of agglomeration that arise from localisation economies, that occur as a result of concentration of firms within the same industry; inter-industry economies, that occur as a result of concentration of firms in related industries in a particular area; and urbanisation economies, that occur across all industries as a result of the scale of a city or region by means of its large markets and urban diversity. It is also pertinent to note that localisation, inter-industry and urbanisation economies are not mutually exclusive – they may occur individually or in combination.

In the empirical literature, there are two broad approaches to identify the determinants of firms' location decisions. One is survey-based or the 'stated preference' approach', for instance to ask firms directly, through an investment climate survey, for instance, about what location factors are important to them. The second approach is a modelling approach or an econometric analysis of empirical patterns used to identify 'revealed preferences' based on the characteristics of the region.

To my knowledge, there are no empirical tests in the literature on factors that could drive the location decisions of informal activity. The established research looks mainly at the formal sector – whether for manufacturing, or services or both. For instance, with regards to formal manufacturing in India, Lall and Meningstae (2005) analyse the productivity of plants sampled from 40 of the country's largest industrial cities and found that differences in clustering across locations were explained by market access, labour regulation and the quality of power supply. With regards to foreign entrants into domestic manufacturing sectors, Head and Reis (1996) show that foreign firms in China preferred to locate in cities where other foreign firms are located. In their paper Head and Mayer (2004) show that downstream linkages made

regions in Europe more attractive to Japanese investors, but the paper does not account for access to suppliers. Cheng and Kwan (2000), and Amiti and Javorcki (2005) also confirm that regional markets and buyer-supplier linkages were important factors affecting the location decisions of foreign firms.

Services firms are theorised to be different from manufacturing. For instance, in some services, product specialisation, rather than standardisation, may be more important in capturing markets (Enderwick 1989), and proximity to competitors, suppliers and markets may be significant determinants relative to agglomeration economies (Bagchi-Sen 1995). And with the introduction of new communication technologies and the ability to slice the service production chain more thinly, it could be argued that proximity would cease to be an important factor in explaining agglomeration economies. Earlier research conducted in North America (Kirn 1987 for the US, and Coffey and McRae 1989 for Canada) found that producer services did not necessarily follow population and manufacturing location patterns – they could locate in peripheral regions and develop an export base. However, more recent research (Dekle and Eaton 1999, Coffey and Shearmur 2002) found evidence that the agglomeration economies exerted a stronger influence in services than in manufacturing, in spite of advances in information and communications technology.

There are a number of reasons why informal activities are different from the formal economy – they are usually an extension of the household economy and start-ups that require little or no capital investment. Informal sector enterprises in India comprise of unregulated micro-enterprises, the bulk of which employ less than five workers, and all of which employ less than 50 workers. Examples of such enterprises are those that produce bidis (Indian cigarettes), small piece-rate suppliers to the textile, weaving or footwear sectors, small shopkeepers etc. The informal sector is also the largest employer of rural migrants in big cities like Mumbai, Kolkata and

Delhi, and like in other countries, the sector serves as the only source of employment to those who are unable to find work in the formal economy. Thus, small enterprises have been viewed as an important means of promoting industrialisation and employment in poor countries.

McGee (1977) noted that the informal sector in South-East Asian cities tended to concentrate in areas of dense population such as nodes of transportation, or where adjacent activities are entertainment complexes, public markets and also in those localities where they could benefit from product complementarities and mutual customer attraction. A priori, there is no reason to assume that informal sector activity remains unaffected by agglomeration economies. Indeed, it could be hypothesised that in the absence of access to formal credit facilities, or alternatively since they are untouched by changes in regulations, the importance of buyer-supplier linkages and informal networks of social interaction could be more important to them than to firms operating in the organised sector. The informal sector in India largely ignores labour regulations, officially recognised collective bargaining processes, taxes or institutional obligations. There is some research (Marjit and Kar 2009) to show that informal manufacturing and self-employed units accumulate fixed assets and invest and that often they are able to do so in times when their formal counterparts are mired in complex regulations.

Production in the formal sector is also dependent on subcontracting among informal firms specialised in some aspect of the vertical production chain. Although parts of the unorganised sector pertain mostly to the production of non-tradables in the economy (think of street vendors and domestic help) they are also an important input to the production of intermediate goods, processed exports and import substitutes, supported by supply side contracts with the formal sector. For instance, informal carpet weavers in Agra operate alongside larger, more formal carpet

designers and exporting firms in the city. And to the extent that the informal sector is linked to its formal counterpart, wages in the sector could be affected structural changes in the formal industrial sector.

With the theoretical and empirical literatures in mind, this paper will concentrate on the extent to which agglomeration economies matter to informal firms' location decisions, and compare them to those in the formal sector. The next section will describe the estimation framework employed and them move on to discussing the results and possible endogeneity bias.

4 Estimation Framework

4.1 Econometric model

A popular model of location choice are conditional logits which assume that a firm evaluates alternative locations at each time period, and would consider relocation if its profitability in another place exceeded that at its current location[7]. The use of a discrete choice framework to model location behaviour stretches back to the 1970s, when Carlton (1979) adapted and applied McFadden's (1974) Random Utility Maximiation Framework to firm location decisions.

Within such a discrete choice framework, a general profit function is used to explain how new firms choose a location. Following McFadden the model assumes a set $J = (1,2,.....j,.....n)$ of possible locations (districts) assuming that location j offers

[7] In reality, relocation can be costly and firms need to take account of sunk investments in production capacity, and other costs of moving. However, these relocation costs are not considered in the model.

profitability level π_{ijk} to a firm i in industry k. The resulting profitability equation yielded by location j to a firm i in industry k is:

$$\pi_{ijk} = \beta Z_{ijk} + \xi_j + \varepsilon_{ijk} \tag{1}$$

where β is the vector of unknown coefficients to be estimated, ξ_j measures unobserved characteristics of the district which can affect the firm's profitability and ε_{ijk} is a random term. Thus, the profit equation is composed of a deterministic and a stochastic component. Under the assumption of independent and identically distributed error terms ε_{ijk}, with type I extreme-value distribution, then it can be assumed that the *i*th firm will choose district *j* if $\pi^i_j \geq \pi^i_l$ for all *l*, where *l* indexes all the possible location choices to the *i*th firm. Thus, the probability that any firm will choose to locate in a district *j* is given by:

$$p_{ijk}(\pi_{ij} \geq \pi_{il} \forall l \neq j) = \frac{e^{\beta Z_{ijk}}}{\sum_{m=1}^{J} e^{\beta Z_{ijk}}} \tag{2}$$

where p_{ijk} is the probability that firm i in industry k locates in district j. If we let $d_{ijk} = 1$ if firm i of industry k picks location j, and $d_{ijk} = 0$ otherwise, then we can write the log likelihood of the conditional logit model as follows:

$$\log L_{cl} = \sum_{i=1}^{N} \sum_{k=1}^{K} \sum_{j=1}^{J} d_{ijk} \log p_{ijk} \tag{3}$$

In practice, however, the implementation of the conditional logit model in the face of a large set of spatial alternatives is very cumbersome[8]. The conditional logit model is also characterised by the assumption of Independence of Irrelevant Alternatives (IIA).

[8] Guimaraes et al. (2003) provide an overview of the problems and how different researchers have attempted to deal with them in the past.

Consequently, the ratio of the logit probabilities for any two alternatives does not depend on any alternatives other than the two considered. More formally, this implies that the ε_{ijk}s are independent across individual firms and choices; all locations would be symmetric substitutes after controlling for observables. This assumption would be violated if districts within particular states were closer substitutes than others outside of the state boundary. The addition of dummy variables for each individual choice would effectively control for choice specific unobservables, amounting to the following specification:

$$\pi_{ijk} = \delta_j + \beta Z_{ijk} + \xi_j + \varepsilon_{ijk} \qquad (4)$$

where δ_js are the alternative specific constants introduced to absorb factors that are specific to each particular choice. In this case all explanatory variables (observable or unobservable) that only change across choices are absorbed by the alternative specific constants. In the presence of large datasets, such as the one I plan on using, this implementation would be impractical because of the large number of parameters to be estimated. And this would still leave the problem of the IIA unsolved.

As an econometric alternative, it can be shown (Guimaraes et al 2003) that the implementation of conditional logit models yields identical results to Poisson regression models when the regressors are not individual specific. They demonstrate how to control for the potential IIA violation by making use of an equivalence relation between the conditional logit and Poisson regression likelihood functions. In a separate paper, Guimaraes et al (2004) provide an empirical demonstration. In this model the alternative constant is a fixed-effect in a Poisson regression model, and coefficients of the model can be given an economic interpretation compatible with the Random Utility Maximisation framework. Since using both models yield identical parameter estimates, I will use Poisson regressions to generate coefficients. See

Mukim and Nunnenkamp (2010) for a comprehensive list of empirical papers that use Poisson models and those that use conditional logits.

Guimaraes et al (2003) show that Equation (3) is equivalent to that of a Poisson model that takes the number of new firms in a district, n_{ijk}, as the dependent variable and includes a set of location-specific explanatory variables. The same results will be obtained if we assume that n_{ijk} follows a Poisson distribution with expected value equal to:

$$E(n_{ijk}) = \lambda_{ijk} = \exp(\alpha d_{ijk} + \beta Z_{ijk}) \quad (5)$$

where $[\alpha, \beta]$ is the vector of parameters to be estimated and d_{ijk} is a vector of K dummy variables, each one assuming the value 1 if the observation belongs to industry k. Thus, the above problem can be modelled as a Poisson regression where the $[\alpha, \beta]$ vector can be estimated regardless of the number of δ_k parameters.

To sum up, I test the importance of economic geography and locational factors by implementing a count model, wherein the count of new firms within a location is modelled as a function of factors common to the location and those common to particular sectors within a location. The original estimation framework is based on a location decision model in which individual firms compare profitability across different locations.

4.2 Specification of variables

The deterministic component of the function consists of the various attributes of the location that can influence the profitability of a firm in that particular location, and the random component consists of the unobserved characteristics of the location, and measurement errors. The dependent variable in the model is the count of new

informal firms at time t, whilst all the explanatory variables in the model are defined at time $t-1$. Section 4.3 below describes the sources of data and the cross-sectional time period for manufacturing and services firms in detail.

The observables in this model are given by:

$$Z_{ijk} : \sigma_{jk}, \Lambda_{kj}, U_j, MA_j, Ed_j, X_j, W_j, WE_j$$

Where:

σ_{jk} represents localisation economies, represented by the share of firms in industry *k* found in location *j*

Λ_{kj} represents inter-industry trading relations measured by the strength of buyer-supplier linkages

U_j represents urbanisation economies in location *j*

MA_j summarises access to markets in neighbouring districts

Other district-level characteristics include:

Ed_j measures the level of human capital in location *j*

X_j captures the quality and availability of infrastructure (electricity and communications)

W_j a vector of factor input price variables in location *j*

WE_j captures the level of wealth) in location *j*

ξ_j measures unobserved characteristics of the district which can affect the firm's profitability. Each firm considers these factors at the time it is making its location decision, but these are not captured in the data. The specifics of the endogeneity problem are dealt with in more detail in Section VI.

The economic geography variables in this model are represented by market access (MA_j), localisation economies (σ_{jk}), inter-industry economies (Λ_{kj}) and

urbanisation economies (U_j). The variables representing business environment are Ed_j (educational attainment) X_j (quality and availability of power and communications' infrastructure) and WE_j (wealth). The remainder of this section provides a detailed description of each of the variables used in the model.

Localisation economies (σ_{jk}) 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 district *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 district.

$$\sigma_{jk} = \frac{E_{k,j}}{E_k}$$

There are several approaches for defining inter-industry linkages: input-output based, labour skill based and technology flow based. Although these approaches represent different aspects of industry linkages and the structure of a regional economy, the most common approach is to use the national level input-output accounts as templates for identifying strengths and weaknesses in regional buyer-supplier linkages (Feser and Bergman 2000). The strong presence or lack of nationally identified buyer-supplier linkages at the local level can be a good indicator of the probability that a firm is located in that region. To evaluate the strength of buyer (supplier) linkages for each industry, a summation of regional (here district) industry employment weighted by the industry's input (output) coefficient column (row) vector from the national input-output account is used:

$$\Lambda_{kj} = \sum_{k=1}^{n} w_k e_{kj}$$

where, Λ_{kj} is the strength of the buyer (supplier) linkage, w_k is industry *k*'s national input (output) co-efficient column (row) vector and e_{kj} is total employment for industry *k* in district *j*. The measure examines local level inter-industry linkages based on national input-output accounts. The national I-O coefficient column vectors describe intermediate goods requirements for each industry, whilst the I-O coefficient row vectors describe final good sales for each industry. Assuming that local industries follow the national average in terms of their purchasing (selling) patterns of intermediate (final) goods, national level linkages can be imposed to the local level industry structure for examining whether district *j* has a right mix of buyer-supplier industries for industry *k*. By multiplying the national I-O coefficient vector for industry *k* and the employment size of each sector in district *j*, simple local employment numbers can be weighted based on what industry *k* purchases or sells nationally.

I use the Herfindal measure to examine the degree of economic diversity, as a measure of urbanisation (U_j) in each district. The Herfindal index of a district *j* (U_j) is the sum of squares of employment shares of all industries in district *j*:

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

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 U_j is one when the entire regional economy is dominated by a single industry. Thus a higher value signifies lower level of economic diversity.

In principle, improved access to consumer markets (including inter-industry buyers and suppliers) will increase the demand for a firm's products, thereby providing the incentive to increase scale and invest in cost-reducing technologies. The proposed model will use the formulation proposed initially by Hanson (1959), which states that the accessibility at point 1 to a particular type of activity at area 2 (say, employment) is directly proportional to the size of the activity at area 2 (say, number of jobs) and inversely proportional to some function of the distance separating point 1 from area 2. Accessibility is thus defined as the potential for opportunities for interactions with neighbouring districts and is defined as:

$$MA_j = \sum_j \frac{S_m}{d_{j-m}^b}$$

Where, MA_j is the accessibility indicator estimated for location j, S_m is a size indicator at destination m (in this case, district population), d_{jm} is a measure of distance between origin j and destination m, and b describes how increasing distance reduces the expected level of interaction[9]. The size of the district j is not included in the computation of market access – only that of neighbouring districts is taken into account[10]. The accessibility indicator is constructed using population (as the size indicator), distance (as a measure of separation) and is estimated without exponent values. The market access measure has been constructed by allowing transport to occur along the orthodromic distance[11] connecting any two districts within a 500-kilometre radius.

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

[10] The final specification includes population to control for the size of district j.

[11] Also known as great-circle distance, it is the shortest distance between any two points on the surface of a sphere.

A distinguishing feature of my approach to evaluating the factors that drive firms to locate in particular districts is that I make use of data on education. I assess quantitatively the role played by the human capital across different districts on the decisions of firms across different industries to situate themselves in a particular district. I include a measure of the effect of education, captured by the education variable - Ed_j. This is defined as the proportion of the population within the district with a high-school education.

I define X_j as a measure of 'natural advantage' through the embedded quality and availability of infrastructure in the district. I use the availability of power (proxied by the proportion of households with access to electricity) within a location as an indicator of the provision of infrastructure. In addition I also use the proportion of households within a district with a telephone connection as an indicator of communications' infrastructure.

W_j is an indicator of input costs in location *j*, and is given by nominal district-level wage rates (i.e. non-agricultural hourly wages). The expected effect of this variable is hard to pin down theoretically. On the one hand, if wages were a measure of input costs then one would expect informal activity to be inversely related to wages, since high costs within a location would make it less attractive. However, it is also important to control for the skill set of the workers since a positive coefficient on wages could be proxying for more skilled-labour. In theory, workers with higher ability could demand a higher wage rate and in turn enjoy a higher level of consumption. Although I am unable to directly control for the ability of the worker, I include 'education' as a proxy for the level of human capital within the district. And thus, the proportion of high-income households (WE_j) within a district is an indicator of the general level of wealth, or more specifically, consumer expenditure within a

district. The variable is constructed using household consumption data and refers to those households that belong to the highest monthly per-capita consumption expenditure group[12].

4.3. Data Sources

The dependent variable, used in the reduced form estimation, is the count of new firms within the informal sector in India – in the manufacturing and in the services sector. The data is drawn from the Fifty-Seventh Round (July 2001-June 2002: Unorganised Service Sector) and the Sixty-Second Round (July 2005-June 2006: Unorganised Manufacturing Enterprises) of the National Sample Survey Organisation. The former household survey contains data on services enterprises in the informal sector (NIC division 38-97), and the latter on manufacturing enterprises in the informal sector (NIC division 15-37). Enterprises are divided into (1) own account enterprises, which are normally run by household labour and which do not hire outside labour on a regular basis, (2) non-directory establishments, which employ one to five workers (including household and hired taken together) and (3) directory establishments, which employ six or more workers (including household and hired taken together).

I extract data on new firms from the question that asks the enterprise its status over the last 3 years (expanding/stagnant/contracting/operated for less than 3 years). I select enterprises that respond in the positive to the latter option, in each of the two surveys. The surveys also contain data on the district within which the enterprise is located. The total number of new services firms counted within the 1999 survey

[12] The actual MPCE category differs depending on the year of the survey, the type of district (rural or urban) and the population of the district.

equals 2,409,204 and the count of new manufacturing firms for the 2004 survey is 2,041,137. In short, I carry out two separate cross-sections, one for unorganised manufacturing firms and the other for unorganised services firms. Since the surveys sample different firm populations, I am do not exploit changes between the two rounds – however, I am interested in looking at what factors drive unorganised manufacturing and/or services firms to a district.

The choice of years is dictated by the data. Whilst data on the dependent variable is drawn from the NSSO Rounds described above, I extract data from the Employment and Unemployment Surveys - Round 55.10 (July 1999 – June 2000) and Round 61.10 (July 2004 – June 2005). The former is the source of explanatory variables for the cross-sectional analysis for services, and the latter for manufacturing. This data, which is disaggregated by industry and district, allows me to construct my agglomeration variables. It is important to keep in mind that since employment data is taken from household surveys, it includes employment within the economy as a whole, and does not differentiate between the formal and the informal sector. In other words, the construction of localisation, input-output and urbanisation economies already assumes linkages between the organised and unorganised sectors. Data on education, electricity and communications infrastructure, and on wages and wealth within the district are also drawn from the household surveys. I use population data from the 2001 Census to construct the market access variable.

5 Results and Discussion

I start with an illustration of the characteristics of the data to explain my modelling choices. The first observation is that the data is over-dispersed. In Table 9, the mean number of new firms per district is around 4,111 for the services sector, and 3,531 for the manufacturing sector. At the same time the respective standard deviations are around 1.6 to 2.3 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. A frequent occurrence with count data is an excess of zeroes – in this case, however, this is not a significant problem. Only 29 districts (of a total of 586) have zero new services units, and 52 districts (of a total of 578) have zero new manufacturing units.

I also check the suitability of the different types of models with regards to their predictive power. '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. Of all the locations in the sample, 4.9% have no new services units, and 9% have no new manufacturing units. In both cases, the Poisson model (Fit_p) predicts that 0% of all districts would have no new units – clearly the model underestimates the probability of zero counts. The negative binomial (Fit_nb), which allows for greater variation in the variable than that of a true Poisson, predicts that 0.66% and 3.25% of all districts will have no new services or manufacturing units respectively. One could also assume that the data comes from two separate populations, one where the number of new firms is always 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 new firms for those locations not in the always zero group. The Zero-inflated Poisson (Fit_zip) predicts that 2.5% and 8.42% of all districts will have no new services or manufacturing units, much closer to the observed value.

An alternative approach to the zero-inflated Poisson is 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 a positive counts. For this data, this would imply using a logit model to differentiate between districts that have no new firms and those that do, and then a truncated model for the number of districts that have at least one new firm. 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).

The response variable is 'count', i.e. the number of new firms 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 at $x+1$ (here implying a unit percentage change in the regressor[13]). Since the difference of two logs is equal to the log of their quotient, i.e. $\log(\mu_{x+1}) - \log(\mu_x) = \log(\frac{\mu_{x+1}}{\mu_x})$, 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 new firms per district. The coefficients[14] could also be interpreted as incidence rate ratios (IRR), i.e. the log of the rate at which events occur.

The IRR score can be interpreted as follows: if localisation were to increase by a percentage unit, the rate ratio for the count of new manufacturing firms would be

[13] This is because the regressors are in logarithms of the original independent variables.
[14] The non-exponentiated coefficient results can be made available on request.

expected to decrease by a factor of 0.382, i.e. by 61.8[15] percentage points (see the coefficient of localisation in model 3 in **Table 10**). In other words, if input linkages were to increase by a percentage point, the rate ratio for the count of new services firms would be expected to increase by a factor of 1.247 i.e. by 24.7 percentage points (see the coefficient on input in model 3 in **Table 11**)

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. 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. I also control for the size of the district (population), and the total employment within the district (wherever possible) and include state dummies. The economic geography variables are represented by localisation, input, output, urbanisation and market access, whilst business environment variables are represented by education, telephone, electricity, wages and wealth. The results of zero-truncated negative binomial models (for both manufacturing and services), which have the best goodness-of-fit statistics, are provided in Table 10 and Table 11 – results from the other models are presented in Table 12 and Table 13.

Localisation has a negative and significant effect – since localisation refers to the clustering of firms within the same industry within a location, this could be evidence that clustering leads to competition of firms within the same industry. Linkages to final goods' suppliers have a positive and significant effect on the attractiveness of a district to new informal manufacturing activity. On the other hand, it is not clear why such firms seem to have a negative association with regards to co-

[15] 0.618 = 1 – 0.382

location with intermediate goods' buyers. Market access, i.e. being located close to larger, more populated districts again seems to have no effect on how attractive a district is to informal manufacturing activity.

With regard to business environment variables, the effect of education and telecommunications infrastructure seems to be insignificant. The negative coefficient (0.622<1) on electricity could be explained by the phenomenon of manufacturing subsidising residential power in many parts of the country, and the presence of more households with access to power could potentially increase the costs of subsidisation for the manufacturing sector. Wages also seem to be unrelated to a location's attractiveness – as mentioned before I am unable to directly account for the skill set of the worker. However, I do include the proportion of wealthy households within the district as a control – this would allow me to control for the ability of some workers to demand higher wages, and also provide an indication of the demand within a district. Since I control for the general quality of human capital within the district, I interpret the positive and significant coefficient on wealth as informal manufacturing activity being attracted to districts with higher consumption expenditures.

The results for informal service firm births show that the higher the intra-industry concentration, the lower the attractiveness of the location. Since informal services refer mainly to small shopkeepers and households providing services, this means that localisation may be capturing the effect of competition within a location. The effect of input linkages is positive and significant across different models implying that informal services tend to be attracted to those industries that they supply to. In the case of services, unlike manufacturing, new units are also attracted to their intermediate goods' industries (1.169>1). Industrial diversity within a district seems to have a negative and significant effect - recall that since a higher Herfindahl index implies lower industrial diversity, the direction of the sign of the coefficient could be

evidence of a positive association between more industrial diversity and more profits, or greater attractiveness of the district.

With regard to the business environment, access to electricity has a positive and significant effect, whilst education and communications infrastructure do not seem to matter much. The size of the district, i.e. population, strongly attracts informal services activity. This is intuitive since one would expect clustering from personal consumer services (such as hairdressers, or rickshaw drivers) that supply the final demands for consumers and thus need to be located close to urban populations. The total level of employment, both in formal and informal activity, within a district, on the other hand, has a negative effect and is somewhat significant.

In summary, the effects of localisation and input-linkages, and the absence of the effects of education or communications, are broadly stable across different models employed for both the manufacturing and the services sector (see Appendix C). Access to power seems to matter negatively for manufacturing, and positively for services. The size (i.e. population) of the district also makes a location more attractive to informal activity.

6 Endogeneity Issues, Robustness and Other Exercises

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). The underlying assumption within the model is that if a particular location offers some inherent features that improve the profitability of certain economic

activities, firms will be attracted to that location. Such inherent features may be related to natural endowments or regulatory specificities, but they could also have to do with essentially un-measurable factors such as local business cultures. How to isolate the effect that runs from agglomeration to performance thus represents a considerable challenge. With regard to the proposed analysis, the presence of these unobservable sources of a location's natural advantage complicates the estimation procedure, particularly in identifying the contribution of production externalities to the location decision of firms.

Ellison and Glaeser (1997) point out that the effects of unobservable sources of 'natural advantage' (i.e. positive values of ξ_{jk}) will not be separately identified from those of production externalities between firms that arise simply from firms locating near one another. Simply including the number of firms or employment in a particular industry, which is a commonly used indicator in empirical studies evaluating localisation economies, will not be able to distinguish whether firms are attracted by a common unobservable, whether they derive benefits from being located in close proximity to one another, or whether it is some combination of the two. As it is impossible to get data on all the factors relevant to a firm's location decision, it would be helpful to find an instrument for own industry concentration that is not correlated with the unobservable sources of natural advantage ξ_{jk}.

I follow the identification strategy used by Lall and Mengistae (2005) who address this problem by using historic land revenue institutions, set up by the British and detailed by Bannerjee and Iyer (2005), 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 are of interest to the analysis for a three reasons. First, the British decision on which land tenure system to adopt depended more on the preferences of individual administrators rather than a systematic evaluation of region-specific characteristics. Thus, the choice of institutional arrangements is largely exogenous to regional attributes. Second, landlords were allowed to extract as much as they wanted from their tenants, thus making their behaviour predatory, leading to high inequality and low general investment in their districts. Further, as most wealthy landlords were not cultivators themselves, this reduced pressure on the state to deliver services important to farmers as well as general public goods. Third, rural institutions have considerable bearing on urban and industrial development. Rural class structures and social networks do not disappear once people move to cities. Thus, these land-tenure systems serve as good instruments since they have been found to influence agricultural investment, profitability and general industrialisation in the post-independence period, and since the choice of institution was largely exogenous, they are not correlated with any observable features of the underlying natural geography of the region.

However, it should be noted that land revenue institutions are not perfect instruments. These institutions had long-lasting effects on many aspects of the district, not only on its general level of industrialisation. Thus, all measures of agglomeration - localisation, input-output and urbanisation - could be treated as endogenous. In theory, these institutions could also serve as an instrument for the level of educational attainment or of power infrastructure within the district.

Following Lall and Mengistae (2005), I link Banerjee 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. I then use instrumental variable techniques in my estimation, and in separate

specifications, I instrument localisation and urbanisation with the choice of land revenue system. 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 run a simple OLS and a linear regression with an IV specification, using standardised counts as the dependent variable. I also run an alternative generalised linear model (GLM) (Hardin and Carrol 2003) to check for the strength of the instrument and to address endogeneity concerns due to measurement errors.

The results of the specifications are presented together with the results of diagnostics in Tables 14-17. The tests confirm the validity of the IV specification and the strength of the instrument when the urbanisation coefficient is instrumented with land revenue institutions, for both manufacturing and services. This is not the case when localisation is instrumented with land revenue institutions – where the F-statistic is well below the rule-of-thumb value of 10. I also perform the Durbin-Wu-Hausman test to examine if endogeneity of urbanisation and localisation could have adverse effects on OLS estimates, and find that the results of the IV estimates are preferable.

A comparison of the exponentiated coefficients from the Poisson models (3-6), simple, instrumented and AGLM, shows that in the case of manufacturing, the coefficient on urbanisation and localisation remain relatively stable and remains significant in a few cases after instrumenting. In the case of services, after instrumenting with land revenue institutions, urbanisation ceases to be significant, whilst localisation has a much stronger negative effect. First stage results are reported in Table 18.

6.1 Robustness check: Controlling for size

As a robustness check, I carry out the same exercise by differentiating between firms of different sizes. I divide the sample of enterprises into those that are small (i.e. employ less than 5 workers) and large (i.e. they employ more than 5 workers). In the case of unorganised manufacturing, almost 90 per cent of the firms in the sample, thus defined, are small-scale enterprises. For informal services, small-scale enterprises account for 93 per cent of the sample. The sample could also be divided into own-account enterprises (OAE) and establishments. Own-account enterprises do not employ any hired workers on a regular basis, whilst establishment enterprises employ one or more workers on a regular basis. Around 68 per cent of all informal manufacturing, and approximately 70 per cent of all informal services enterprises are own-account enterprises.

When I compare manufacturing firms by their sizes, I make a few interesting observations. Localisation economies continue to have a strong negative effect on small-scale or own-account enterprises. In addition, these enterprises are attracted to those they sell to but not those they buy from, unlike their larger counterparts (see 'Establishments') that also seem to be attracted to their intermediate suppliers. Most importantly, the size of the district, i.e. the population explains an important part of what makes a location attractive to small-scale and OAE enterprises.

Some of these results also hold for small-scale or OAE informal services enterprises. Localisation continues to have a negative and significant effect – implying that new births do not take place in locations with more existing firms in the same sector. Services firm, irrespective of their size seem to be co-located with those they supply to. But smaller enterprises are now also attracted to intermediate suppliers, as are establishments. The level of industrial diversity has a positive impact on all establishments, except those with more than 6 workers, where the result is insignificant. Access to electricity has a positive impact and again, the size of the

district makes a location more attractive to small-scale and OAE enterprises and establishments.

In summary, the results are broadly similar to those obtained before, except that the impact of certain factors seems to be stronger for small-scale firms than for larger establishments in the data. In their analysis of Italian firms Lafourcade and Mion (2003) also find that small firms are more spatially concentrated than large ones and are more sensitive to input-output linkages. Additionally, as the data is unable to differentiate between formal and informal services, controlling for the size of the firm provides a reasonable approximation of informality, and excludes large services enterprises that are not formally registered under the Factories Act, but which in all other ways are run like formal-sector enterprises.

6.2 Unorganised versus organised

I also carry out the same exercise for the organised manufacturing and services sector in India, to check how the results differ. I use data for both manufacturing and services firms from the Prowess database, and data from the Annual Survey of Industries (ASI) for manufacturing firms. Prowess is a corporate database that contains normalised data built on a sound understanding of disclosures of over 18,000 companies in India. The ASI contains data on over 140,000 manufacturing firms in India. I then re-run the regressions for new firms for the two cross-sections – 1999-2000 and 2004-2005. Although I carry out the regressions using Poisson, zero-inflated and zero-truncated methods, I only report the results of the negative binomial specifications[16]. This facilitates comparison, but more importantly the negative

[16] Results from the models are available on request.

binomial models exhibit the best goodness-of-fit statistics. As before, the coefficients are reported as Incidence Rate Ratios for ease of interpretation.

Since I have data on much fewer firms when using the Prowess dataset for the organised sector, most of the predictor variables are no longer significant. The effect of localisation is no longer negative, nor significant, with regards to organised services industry[17] – contrast this with the negative and significant effect for unorganised services for the same variable. This could be since formal services consist mostly of finance, insurance, IT firms etc, which may benefit more from knowledge spillovers when in proximity to one another, as compared to informal services firms, such as small shop-keepers, rickshaw drivers etc, which would suffer from higher competition with more proximity. Supplier linkages, i.e. proximity to those industries to which a formal service firm sells its products to, also make a location more attractive – this is similar to the positive and significant results for informal services. Data on formal manufacturing from the ASI provide some evidence of positive spillovers from locating close to firms within the same industry. There is also evidence that formal manufacturing firms like locating close to those they source from, but not necessarily close to those they supply to. The effect of telecommunications, education or power infrastructure is not consistent across different years. The size (population) of a district has a strong positive effect on formal services industries, which is interesting, but seems to have a negative effect on formal manufacturing. The latter could be explained by urban regulations that prevent heavy industries from clustering near large population settlements in cities and towns.

One might also expect that the rate of informal activity would be higher in places where there are barriers to entry to formal activity. In other words, it may be

[17] This result is also similar to the coefficient on large-scale services enterprises (see Table 20), providing evidence that large service enterprises are run just like other formal sector enterprises registered under the Factories Act.

possible that informal activity serves as a substitute to the formal sector. If this were the case, one might expect to see a negative correlation between the informal and formal firm births. In the data, I find that both the count and the rate of new firm activity are positively correlated at the geographical level of the state and that of the district. And so I investigate the inter-linkages between the types of sectors that could be driving these correlations.

6.3 Measures of co-agglomeration

While the data treats formal and informal manufacturing and services as separate units, in reality these firms are inter-linked in a number of ways. The agglomeration variables (localisation, input, output and urbanisation) have been constructed taking total employment, i.e. across the formal and informal sector, into account. However, this does not tell us anything about the linkages between and across formal and informal, manufacturing and services firms. Following Ellison and Glaeser (1997, 2010) I compute pair-wise coagglomeration measures for all 2-digit industries for manufacturing and services, across the organised and the unorganised sector (see Appendix A for construction of the Index). I have at my disposal data from four different sources: organised manufacturing data comes from the Annual Survey of Industries, unorganised manufacturing and services data comes from two different surveys of the National Sample Survey Organisation, and organised services data comes from the Prowess database.

Clearly, the Prowess database contains very few firm observations as compared to data from the NSSO and the ASI. I use the Annual Survey of Industries instead of Prowess for manufacturing firms, as the former is a richer source of data, even though the latter contains data on manufacturing units. Since Prowess accounts

for such a small proportion of firms in the sample, using this database gives an inflated value of coagglomeration. In other words, owing to the small size of these sectors when data for total employment in pooled, and the small number of firms in the dataset causes the coagglomeration index to be very volatile. Thus, I drop data from Prowess, and construct coagglomeration measures using the remaining databases. Subsequently, I am unable to construct coagglomeration measures for formal services. Table 23 lists the 20 most coagglomerated sectors. Similar to the EG agglomeration index, the no-coagglomeration benchmark is when the value of the index is zero (i.e. $E(\gamma) = 0$). In general, if the EG coagglomeration index is greater than 0.05, the industries are considered to be highly concentrated.

Certain coaggomerations, such as office and computing maintenance and market research activities with education, i.e. primary, secondary, distance learning education activities, seems intuitive – one might expect these industries to use similar labour pools. However, others, such as the coagglomeration of manufactures of apparel with education, or that of recreational and entertainment activities with recycling, is not clear.

Earlier results found that buyer-supplier linkages explained a large proportion of new informal activity within a district. I will now verify to what extent these linkages are correlated with the final coagglomeration indices observed in my data. Whilst the earlier analysis made no distinction between organised and unorganised industries, this analysis teases out the importance of each type of activity (i.e. formal or informal) for each type of industry (i.e. manufacturing and services). To relate the measure of coagglomeration to a single measure of linkages between a pair of industries, I follow Ellison et al (2010) and construct an input-output index (see Appendix A for construction of the Index). I then relate this single measure for each pair of industry to the coagglomeration measure also constructed for each pair on

industry – except that the latter are also constructed separately for formal and informal manufacturing. The table below provides the correlation values for each pair of coagglomerated industries with the standard input-output index.

Since I do not have data on labour market pooling and knowledge spillovers, in this section I try to discern the effect of input-output spillovers only. A major limitation of the EG index is that it does not distinguish between spillovers and natural advantages to explain the coagglomeration of firms – and I will thus be unable to single out the effect of buyer-supplier linkages from that of natural advantage. A high correlation may be an indication that the pair of industries are coagglomerated owing to input-output linkages, while a low correlation may be an indication that other factors, such as say, labour market pooling or technological spillovers may underlie the observed coagglomeration.

I find that although coagglomeration and input-output linkages are positively associated, the level of correlation is quite low. Coagglomeration between formal manufacturing and formal and informal manufacturing and services does seem to have some correlation with the standard input-output measures, perhaps indicating that these buyer-supplier linkages may explain the coagglomeration to some extent. Interestingly, the standard input-output measure is negatively associated with the coagglomeration of formal services with itself – implying that other linkages may be more important. The same outcome is true for coagglomeration of informal services.

It could also be argued that input-output linkages and coagglomeration are endogenous – in other words, firms may use the outputs of (or sell to) particular sectors simply because these sectors are coagglomerated. If it is assumed that input-output linkages are determined by given production technologies and that the national input-output vectors are representative at the local scale, then I can rule out scenarios in which firms would adjust their inputs or outputs according to what was locally

available. If this were true, I would also expect to find a higher correlation between my measures of input-output linkages and coagglomeration.

The results for input and output linkages to explain the attractiveness of a location to informal manufacturing activity was significant and positive – in other words, being located closer to buyers or suppliers made a location more attractive to new units. The coagglomeration exercise conducted above shows that input-output linkages are in fact positively correlated with the EG measure of coagglomeration, which is what I would expect in light of my earlier results. Similarly, with regards to informal services, although input linkages made a location more attractive to new informal services units, output linkages had a negative effect. The standard input-output measure in the above analysis is an un-directional measure of the input and output variables and thus it could be capturing the negative effect of output linkages found in the earlier regression analysis.

7 Conclusion

This paper seeks answers to the following question: What factors influence the spatial distribution of informal economic activity within India? The main aim of the paper is to understand what drives the process of spatial variations in industrial activity, i.e. in identifying the factors that determine location decisions. It is important to understand why economic activity tends to concentrate geographically because if one can explain geographic concentration, then one can go some way towards explaining important aspects of international trade and economic growth. The importance of this research is underscored by two inter-related factors – that the clustering of economic activity has

important implications for economic development and that the contribution of the informal sector to economic growth and employment makes it a potent tool in influencing regional economic policy.

The empirical analysis finds that economic geography factors have an important effect on informal firms' performance, and thus their decision to locate in a particular area. In the case of formal manufacturing in India, Lall and Mengistae (2005) find that there is a pattern in the data whereby geographically disadvantaged cities seem to compensate partially for their natural disadvantage by having a better business environment than more geographically advantaged locations. The findings in this paper are that economic geography factors, such as input-output economies, do in fact positively impact the attractiveness of a district to new informal activity, whilst localisation seems to be capturing competition, and so it has a negative and significant effect. The analysis finds that the presence of education and telecommunications infrastructure seems to matter little. This is an indication that governments may be limited in their ability to narrow regional disparities in hosting of informal economic activity, which is a source of growth and employment.

This research also makes an important contribution to the empirical literature on industrial development and economic geography. To my knowledge, there are no papers that have examined the location of informal industry, although a handful study the effects of agglomeration economies and business environment on the spatial concentration of manufacturing in emerging countries. In large developing countries the informal sector accounts for an important proportion of domestic product and employment, and any study that does not account for the sector is scarcely representative. In addition, whilst the theoretical development of new economic geography has received much attention in the literature, there is still much scarcity of empirical tests for developing countries.

In addition the use of land revenue institutions as an instrument helps to rule out omitted variables bias by controlling for the difference between first and second nature economic geography, although these instruments are far from perfect. In summary, this paper provides evidence of the validity of the forces emphasised by new economic geography and location theory approaches. The study does not attempt to perfect the theory of economic geography, but it does attempt to confront the existing tenets with data on unorganised industry in India.

The policy implications of the research and its findings are of significant importance – policy-makers need to have an understanding of the relative importance of existing agglomeration economies and business environment if they are interested in influencing the decisions of informal activity. With the importance of this sector and its potential effect on employment and economic growth, such an understanding could provide a powerful tool for spreading growth and employment to geographically less-advantaged regions. This analysis finds that governments may find it an uphill task to encourage informal economic activity to locate to regions that it has not previously favoured.

Tables

Table 1: Share of unorganised activity (2002-03)

Industry	Organised (% of NDP)	Unorganised (% of NDP)	Total
Agriculture, forestry, fishing	4.1	95.9	100
Mining, manufacturing, electricity and construction	60.5	39.5	100
Services	53.1	46.9	100
Total	**43.3**	**56.7**	**100**

Source: National Account Statistics 2005

Table 2: Distribution of Employment (2004-2005)

		Number of workers (millions)	Distribution of workers (%)
Agriculture	*Organised*	6.1	2.4
	Unorganised	252.8	97.6
		258.9	***100***
Non-Agriculture	*Organised*	56.5	28.4
	Unorganised	142.1	71.6
		198.5	***100***
Total	*Organised*	62.6	13.7
	Unorganised	394.9	86.3
		457.5	***100***

Source: NSSO Sample Survey 2004-2005

Table 3: Employment by sector (%)

Industry	1983-84		1987-88		1993-94		1999-2000	
	Org	*Unorg*	*Org*	*Unorg*	*Org*	*Unorg*	*Org*	*Unorg*
Agriculture, forestry and fishing	0.6	99.4	0.7	99.3	0.6	99.4	0.6	99.4
Mining and quarrying	55.5	44.5	44.2	55.8	40.7	59.3	43.2	56.8
Manufacturing	19.7	80.3	17.3	82.7	16.1	83.9	14.9	85.1
Electricity, gas and water	90.7	9.3	71.3	28.7	69.7	30.3	79.0	21.0
Construction	17.7	82.3	10.1	89.9	10	90	6.5	93.5
Trade, hotels and restaurants	2.1	97.9	1.8	98.2	1.6	98.4	1.2	98.8
Transport, storage and communication	38.8	61.2	34.8	65.2	29.7	70.3	21.5	78.5
Services	40.3	59.7	36.8	63.2	31.7	68.3	34.8	65.2

Source: Sakhtivel and Joddar 2006[18]

Table 4: Theil Index for the unorganised sector

District	Manu	District	Serv
Mumbai	255.43	Kolkata	984.42
Ludhiana	146.34	Mumbai	958.80
South Tripura	100.84	Delhi	361.93
Kolkata	80.03	Purba Champaran	248.53
Delhi	52.53	Medinipur	226.19
Ahmadabad	47.11	Ernakulam	175.90
Jaipur	44.08	Pune	169.70
South 24 Parganas	43.08	Thane	161.71
Coimbatore	42.63	Bangalore	139.19
West Tripura	42.19	Hyderabad	137.65
Surat	39.93	Lucknow	131.88
Thane	39.70	Kanpur Nagar	128.59

[18] Organised employment figures are obtained from annual reports (1983 and 1988) and Quarterly Employment Review (1994 and 2000).

North 24 Parganas	39.52	West Tripura	104.66
Haora	37.08	South 24 Parganas	99.99
Murshidabad	36.44	Jammu	96.08
Srinagar	34.17	Thiruvananthapuram	95.27
Hyderabad	34.00	Madurai	92.67
Varanasi	32.53	West Godavari	90.62
Virudhunagar	31.18	North 24 Parganas	90.12
Vellore	29.69	Barddhaman	86.76

Table 5: Ellison-Glaeser Index (Unorganised Manufacturing)

NIC	Description	EG Index
30	Office, accounting and computing machinery	0.204
35	Other transport equipment	0.105
32	Radio, television and communications equipment	0.069
33	Medical, precision and optical instruments, watches and clocks	0.045
19	Tanning and dressing of leather; manufacture of luggage, handbags saddlery, harness and footwear	0.023
31	Electrical machinery and apparatus	0.021
34	Motor vehicles, trailers and semi-trailers	0.017
23	Coke, refined petroleum and nuclear fuel	0.016
27	Basic metals	0.013
16	Tobacco Products	0.012
29	Machinery and equipment	0.010
24	Chemical and chemical products	0.010
25	Rubber and plastic products	0.009
21	Paper and Paper products	0.008
22	Publishing, printing and reproduction of recorded media	0.008
17	Textiles	0.007
26	Other non-metallic mineral products	0.006
36	Furniture	0.004
20	Wood and cork products (except furniture)	0.003
28	Fabricated metal products (except machinery and equipments)	0.003
18	Wearing apparel; Dressing and dyeing of fur	0.002
15	Food products and Beverages	-0.007

Table 6: Ellison-Glaeser Index (Unorganised Services)

NIC	Description	EG Index
73	Research and development	0.287
61	Water transport	0.206
72	Computer and related activities	0.099
63	Supporting and auxilliary transport activities; activities of travel agencies	0.015
90	Sewage and refuse disposal, sanitation and similar	0.013

	activities	
70	Real estate activities	0.005
91	Activities of membership organisations	0.004
71	Renting of machinery and equipment without operator and of personal and household goods	0.003
74	Other business activities	0.003
60	Land transport; transport via pipelines	0.002
80	Education	0.002
93	Other service activities	0.002
85	Health and social work	0.001
92	Recreational, cultural and sporting activities	0.001
55	Hotels and restaurants	0.000
64	Post and communications	0.000

Table 7: Descriptive Statistics

Variable	Expected sign	#		Mean	
		manufacturing	*services*	*manufacturing*	*services*
New firms		567	572	3,531	4,111
Localisation	+	557	469	0.003	0.002
Input	+	557	462	4213.2	3821.3
Output	+	557	462	2189.6	8237.7
Urbanisation	-	578	586	0.41	0.33
Market Access	+	574	582	869363	871313
Education	+	578	480	0.074	0.056
Electricity	+	578	486	0.633	0.559
Telephone	+	578	486	0.368	0.083
Wealth	+	578	486	0.051	0.054
Wages	-/+	574	483	100.94	93.47

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

Table 8: Predictor Variables

	Variable	Indicator	Source(s)	Availability 1999-2000	Availability 2004-2005
Economic Geography	Localisation	Intra-industry concentration	NSSO	√	√
	Input/Output economies	Buyer/Supplier linkages	NSSO	√	√
	Urbanisation	Economic Diversity	NSSO	√	√
	Market Access	Neighbouring markets	Orthodromic distance calculations	√	√

Business Environment	Education	Persons with a High-School education	NSSO	√	√
	Electricity	Persons with access to electricity	NSSO	√	√
	Telephone	Households with a telephone connection	NSSO	√	√
	Wages	Non-agricultural hourly wages	NSSO	√	√
	Wealth	High-income households	NSSO	√	√

Notes: NSSO - National Sample Survey Organisation

Table 9: Characteristics of the Data

	Services			**Manufacturing**		
Variable	#	Mean	Std. Dev.	#	Mean	Std. Dev.
count	586	4111.27	6749.53	578	3531.38	8207.68
count>0	557	4325.32	6856.00	526	3880.49	8525.32
Obs	586	0.0495	0.2171	578	0.0900	0.2864
Fit_p	480	0.0000	0.0000	570	0.0000	0.0000
Fit_nb	480	0.0066	0.0025	570	0.0325	0.0227
Fit_zip	480	0.0250	0.0632	570	0.0842	0.1438

Table 10: Manufacturing IRRs

Variable	Zero-truncated Negative Binomial		
	[1]	[2]	[3]
Localisation	0.385***		0.382***
Input	4.207***		4.173***
Output	0.754***		0.772***
Urbanisation	0.859		0.84
Market Access	1.101		1.064
Education		0.882	1.217
Telephone		0.700***	1.029
Electricity		1.181	0.622**
Wages		1.162	0.826
Wealth		1.048	1.021
Population	1.856	2.019	3.149***
Employment	1.194	1.844	
#	3762	5975	3673
Pseudo R^2	0.022	0.011	0.023
AIC	35199	47070.9	34608.6
BIC	35460.8	47358.8	34894.2

Exponentiated coefficients

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

Table 11: Services IRRs

Variable	Zero-truncated Negative Binomial		
	[1]	[2]	[3]
Localisation	0.709***		0.711***
Input	1.240***		1.247***
Output	1.170***		1.169***
Urbanisation	0.678***		0.688***
Market Access	1.008		1.008
Education		0.998	0.867
Telephone		1.144***	1.057
Electricity		1.068	1.246**
Wages		0.965	0.894
Wealth		1.04	1.067
Population	3.134***	1.857***	2.600***
Employment	0.700*	0.948	0.663*
#	2655	5069	2594
Pseudo R^2	0.03	0.018	0.03
AIC	35056.8	59430.2	34503.8
BIC	35298	59704.5	34773.4

Exponentiated coefficients
* p<0.05, ** p<0.01, *** p<0.001

Table 12: Incidence Rate Ratios (Manufacturing[19])

Variable	Poisson	Negative Binomial	Zero-inflated Poisson	Zero-inflated Negative Binomial	Zero-truncated Poisson	Zero-truncated Negative Binomial
Localisation	0.361***	0.382***	0.411***	0.484***	0.411***	0.482***
Input	4.317***	4.173***	3.163***	2.876***	3.160***	2.876***
Output	0.900***	0.772***	1.026***	0.862**	1.026***	0.864**
Urbanisation	0.894***	0.84	0.912***	0.97	0.905***	0.963
Market Access	1.014***	1.064	1.014***	1.036	1.013***	1.031
Education	1.132***	1.217	1.099***	1.149	1.084***	1.14
Telephone	1.115***	1.029	1.051***	1.064	1.056***	1.078
Electricity	0.563***	0.622**	0.544***	0.630***	0.540***	0.632***
Wages	0.863***	0.826	0.813***	0.763	0.833***	0.778
Wealth	1.073***	1.021	1.075***	1.019	1.071***	1.019
Population	2.345***	3.149***	2.982***	2.661***	2.642***	2.099**
Employment	1.339***				1.159***	1.259
#	3673	3673	3673	3673	2046	2046

[19] I exclude incidence rate ratios from the Zero-Inflated Poisson and Negative-Binomial models since convergence was not reached.

Pseudo R^2	0.48	0.023			0.453	0.034
AIC	3589192.7	34608.6	2635260.3	33597.5	2630313.1	29092.9
BIC	3589366.6	34894.2	2635614.2	33957.6	2630566.1	29351.6

Exponentiated coefficients
* p<0.05, ** p<0.01, *** p<0.001

Table 13: Incidence Rate Ratios (Services)

Variable	Poisson	Negative Binomial	Zero-inflated Poisson	Zero-inflated Negative Binomial	Zero-truncated Poisson	Zero-truncated Negative Binomial
Localisation	0.783***	0.711***	0.811***	0.771***	0.811***	0.772***
Input	1.234***	1.247***	1.219***	1.202***	1.219***	1.200***
Output	1.161***	1.169***	1.151***	1.151***	1.151***	1.151***
Urbanisation	0.748***	0.688***	0.769***	0.729***	0.769***	0.729***
Market Access	1.123***	1.008	1.149***	1.07	1.149***	1.08
Education	0.917***	0.867	0.910***	0.863*	0.910***	0.855*
Telephone	1.046***	1.057	1.077***	1.081*	1.077***	1.085*
Electricity	1.308***	1.246**	1.248***	1.187*	1.248***	1.186*
Wages	0.844***	0.894	0.831***	0.912	0.831***	0.915
Wealth	1.094***	1.067	1.082***	1.059	1.082***	1.06
Population	2.019***	2.600***	1.965***	2.369***	1.965***	2.331***
Employment	0.686***	0.663*	0.698***	0.643**	0.698***	0.647**
#	2594	2594	2594	2594	2259	2259
Pseudo R^2	0.431	0.03			0.405	0.032
AIC	2115024.8	34503.8	1942625.7	34135.4	1940842.9	32366.8
BIC	2115288.5	34773.4	1942965.6	34481.2	1941100.4	32630

Exponentiated coefficients
* p<0.05, ** p<0.01, *** p<0.001

Table 14: Manufacturing (Instrumented variable: urbanisation)

Variables	OLS (1)	2SLS (2)	Poisson (3)	IV Poisson (4)	AGLM (Poisson) (5)	AGLM (Negative Binomial) (6)
Urbanisation	0.919	1.184	0.84	0.769**	0.891	0.763***
Other controls	*yes*	*yes*	*yes*	*yes*	*yes*	*yes*
#	3673	3673	3673	3673	3673	3673
F-Stat		29.61				

Notes: For specifications (3), (4), (5) and (6), the dependent variable is raw counts; Exponentiated coefficients. For specifications (1) and (2) the dependent variable is standardised counts
* p<0.05, ** p<0.01, *** p<0.001

Table 15: Manufacturing (Instrumented variable: localisation)

Variables	OLS (1)	2SLS (2)	Poisson (3)	IV Poisson (4)	AGLM (Poisson) (5)	AGLM (Negative Binomial) (6)
Localisation	0.756***	0.439	0.382***	0.361***	0.0177	0.0368
Other controls			*yes*	*yes*	*yes*	*yes*
#	3673	3673	3673	3673	3673	3673
F-Stat		2.03				

Notes: For specifications (3), (4), (5) and (6), the dependent variable is raw counts; Exponentiated coefficients. For specifications (1) and (2) the dependent variable is standardised counts
* p<0.05, ** p<0.01, *** p<0.001

Table 16: Services (Instrumented variable: urbanisation)

Variables	OLS (1)	2SLS (2)	Poisson (3)	IV Poisson (4)	AGLM (Poisson) (5)	AGLM (Negative Binomial) (6)
Urbanisation	0.671***	0.385	0.688***	2.344	0.973	3.585
Other controls			*yes*	*yes*	*yes*	*yes*
#	3673	3673	3673	3673	3673	3673
F-Stat		15.39				

Notes: For specifications (3), (4), (5) and (6), the dependent variable is raw counts; Exponentiated coefficients. For specifications (1) and (2) the dependent variable is standardised counts
* p<0.05, ** p<0.01, *** p<0.001

Table 17: Services (Instrumented variable: localisation)

Variables	OLS (1)	2SLS (2)	Poisson (3)	IV Poisson (4)	AGLM (Poisson) (5)	AGLM (Negative Binomial) (6)
Localisation	0.856***	0.109	0.711***	0.265***	1.588	Failed to converge
Other controls			*yes*	*yes*	*yes*	*yes*
#	2594	2594	2594	2594	2594	2594
F-Stat		0.15				

Notes: For specifications (3), (4), (5) and (6), the dependent variable is raw counts; Exponentiated coefficients. For specifications (1) and (2) the dependent variable is standardised counts
* p<0.05, ** p<0.01, *** p<0.001

Table 18: First-Stage Results

	Manufacturing		Services	
	localisation	*urbanisation*	*localisation*	*urbanisation*
Land Revenue Institutions	-0.062	0.132***	0.028	0.103***
(Std. Error)	(0.044)	(0.024)	(0.073)	(0.026)
Constant	-16.755***	2.541***	-17.393***	-0.556*
	(0.587)	(0.361)	(0.853)	(0.333)
#	3673	3673	2594	2594
R^2	0.859	0.723	0.652	0.732

Table 19: Manufacturing enterprises by size

	Zero-truncated Negative BInomial			
	Small-scale	Large-scale	OAE	Establishments
Localisation	0.508***	0.789*	0.443***	0.705***
Input	2.635***	1.491**	3.062***	1.349***
Output	0.877**	1.083	0.862*	1.208**
Urbanisation	1.129	0.808	1.094	0.942
Market Access	1.064	1.705	1.061	0.966
Education	1.15	0.749	1.111	1.106
Telephone	1.024	1.057	1.017	1.402**
Electricity	0.717**	0.501*	0.684**	0.762
Wages	0.761*	1.391	0.775	0.918
Wealth	1.026	1.052	1.02	0.978
Population	1.815*	1.51	2.290**	0.997
Employment	1.257	1.734	1.197	1.317
#	2569	570	1539	1074
Pseudo R^2	0.028	0.029	0.035	0.022
AIC	35322.5	6501.6	22141.7	13294.1
BIC	35591.7	6679.7	22387.3	13508.2

Exponentiated coefficients
* p<0.05, ** p<0.01, *** p<0.001

Table 20: Services enterprises by size

	Zero-truncated Negative BInomial			
	Small-scale	Large-scale	OAE	Establishments
Localisation	0.789***	0.843	0.791***	0.744***
Input	1.189***	1.367***	1.189***	1.112**
Output	1.148***	0.882*	1.147***	1.205***
Urbanisation	0.755***	1.345	0.765**	0.706***
Market Access	1.093	1.178	1.106	1.087
Education	0.867*	1.241	0.875	0.841*
Telephone	1.077*	1.104	1.047	1.179***

Electricity	1.158*	1.389	1.189*	0.934
Wages	0.892	1.412	0.922	0.945
Wealth	1.037	0.895	1.058	1.014
Population	2.175***	2.542	2.299***	2.387***
Employment	0.687*	0.380*	0.631*	0.789
#	3834	495	2163	2166
Pseudo R^2	0.023	0.033	0.029	0.023
AIC	50384.6	4820.9	30223	24487
BIC	50672.2	4997.5	30484.3	24748.3

Exponentiated coefficients
* p<0.05, ** p<0.01, *** p<0.001

Table 21: Organised Manufacturing and Services

	Services (Prowess)		Manufacturing (ASI)	
Variable	*1999*	*2004*	*1999*	*2004*
Localisation	0.971	0.951	1.305***	1.117
Input	1.225***	1.200***	0.768***	0.857*
Output	0.906***	1.019	1.195***	1.072
Urbanisation	1.003	0.865*	0.439***	1.174
Market Access	0.988	0.966	2.361***	0.744
Education	0.888**	1.093	0.783	1.563
Telephone	0.958	1.342***	1.002	0.316*
Electricity	1.177**	0.641***	0.944	37.11*
Wages	1.083	0.590***	0.966	9.348***
Wealth	1.050*	0.95	0.783*	1.1
Population	1.416*	1.870**	0.887	0.000153***
Employment	0.764*	0.552**	1.546	368.6**
#	2477	2120	864	457
Pseudo R^2	0.049	0.055	0.163	0.102
AIC	13243.9	11321.4	2588.5	1226.2
BIC	13464.8	11536.4	2745.7	1341.7

Exponentiated coefficients
* p<0.05, ** p<0.01, *** p<0.001

Table 22: Industry Data Sources

Type	Source	Frequency	Percent	Cumulative
Organised	ASI	40694	8.42	8.42
Organised	Prowess (manufacturing)	684	0.14	8.56
Organised	Prowess (services)	367	0.08	8.64
Unorganised	NSSO (manufacturing)	80591	16.67	25.31
Unorganised	NSSO (Services)	361040	74.69	100
Total		483376	100	

Table 23: Most Coagglomerated Industries

Industry1	Type*	Industry2	Type	Coagg index
Apparel and fur	Or	Education	Unor	**0.2321**
Repair/Maintenance of office and computing equipment	Unor	Education	Unor	**0.1715**
Education	Unor	Market research, consulting, bookkeeping etc	Or	**0.1429**
Recreation, motion picture, TV, radio activities	Or	Recycling	Or	**0.1274**
Medical, precision and optical instruments	Unor	Repair/Maintenance of office and computing equipment	Or	**0.1009**
Apparel and fur	Or	Repair/Maintenance of office and computing equipment	Unor	**0.0791**
Apparel and fur	Or	Market research, consulting, bookkeeping etc	Or	**0.0669**
R&D	Unor	Market research, consulting, bookkeeping etc	Or	**0.0611**
Sewage and refuse disposal, sanitation	Or	Leather	Or	**0.0574**
Office, accounting and computing equipment	Unor	Market research, consulting, bookkeeping etc	Or	**0.0523**
Coke and refined petroleum	Or	Collection, purification distribution of water	Or	**0.0511**
Repair/Maintenance of office and computing equipment	Unor	Market research, consulting, bookkeeping etc	Or	**0.0509**
Radio, TV, Communication Equipment	Or	Market research, consulting, bookkeeping etc	Or	**0.0464**
Auxiliary transport, storage and warehousing	Unor	Auxilliary transport, storage and warehousing	Or	**0.0445**
Furniture, jewellary, musical instruments etc	Or	Market research, consulting, bookkeeping etc	Or	**0.0427**
Repair/Maintenance of office and computing equipment	Unor	Market research, consulting, bookkeeping etc	Or	**0.0417**
Sea, coastal, inland water transport	Unor	Leather	Unor	**0.0390**
Furniture, jewellary, musical instruments etc	Or	Radio, TV, Communication Equipment	Unor	**0.0387**
Market research, consulting, bookkeeping etc	Or	Office, accounting and computing equipment	Or	**0.0380**
Market research, consulting, bookkeeping etc	Or	Medical, precision and optical instruments	Or	**0.0377**

*Type refers to the organised (Or) or unorganised (Unor) sector

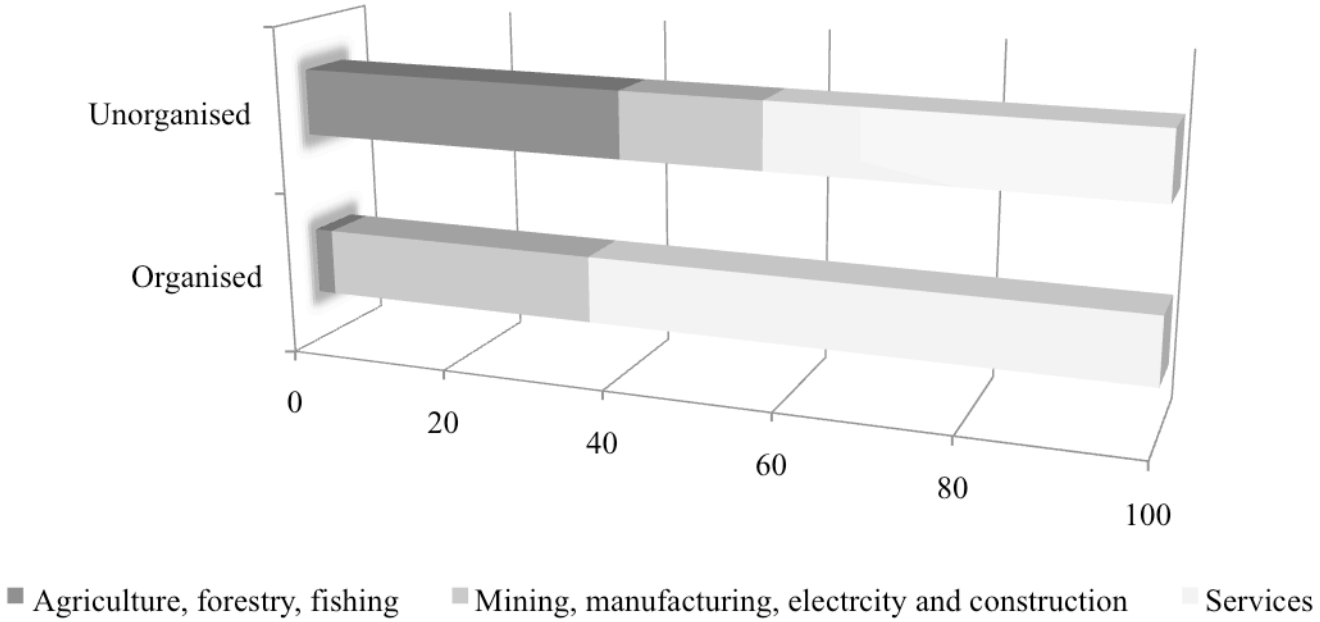
Table 24: Coagglomeration and input-output correlations

Industry 1	Industry 2	Correlation Index
Formal Manufacturing	Formal Manufacturing	0.0531
Formal Manufacturing	Formal Services	0.0688
Formal Manufacturing	Informal Manufacturing	0.0536
Formal Manufacturing	Informal Services	0.0529
Formal Services	Formal Services	-0.0382
Formal Services	Informal Manufacturing	0.0771

Formal Services	Informal Services	0.0175
Informal Manufacturing	Informal Manufacturing	0.0314
Informal Services	Informal Services	-0.0271
Informal Manufacturing	Informal Services	0.0502

Figures

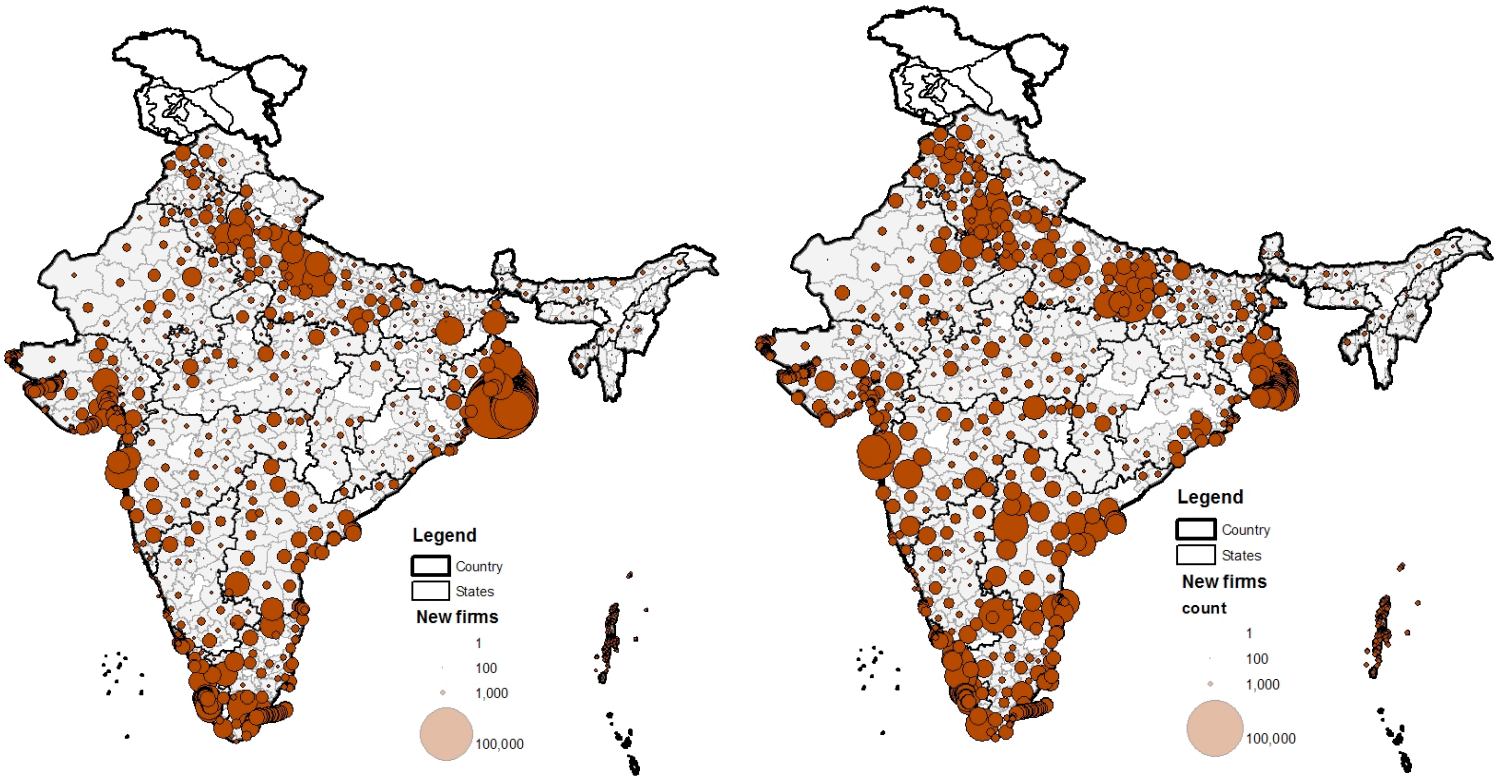
Figure 1: Share of activity as a % of sectoral NDP (2002-03)



Source: National Account Statistics 2005

Figure 2: Distribution of Informal Activity

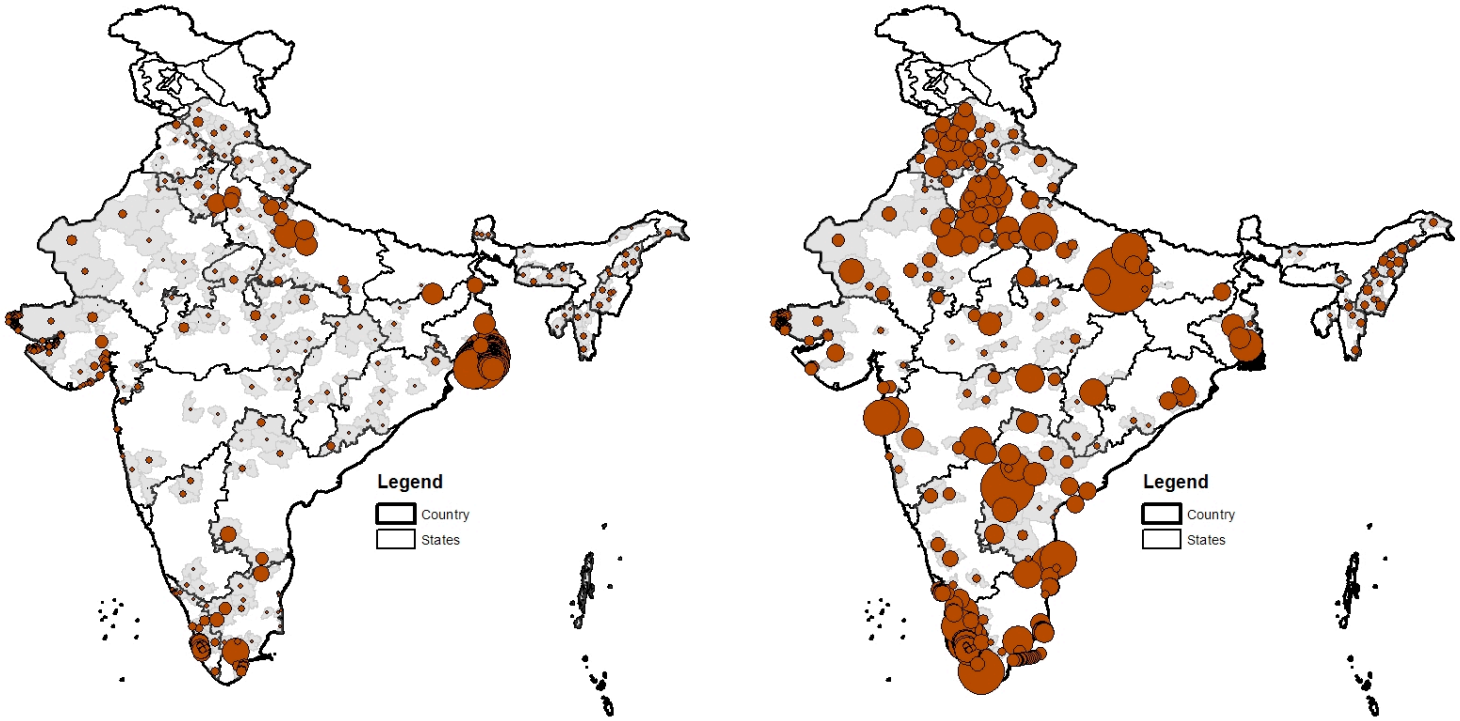
Manufacturing **Services**



Source: Food and Agricultural Organisation (GAUL) and Prowess

Figure 3: Distribution of Informal Activity (with controls)

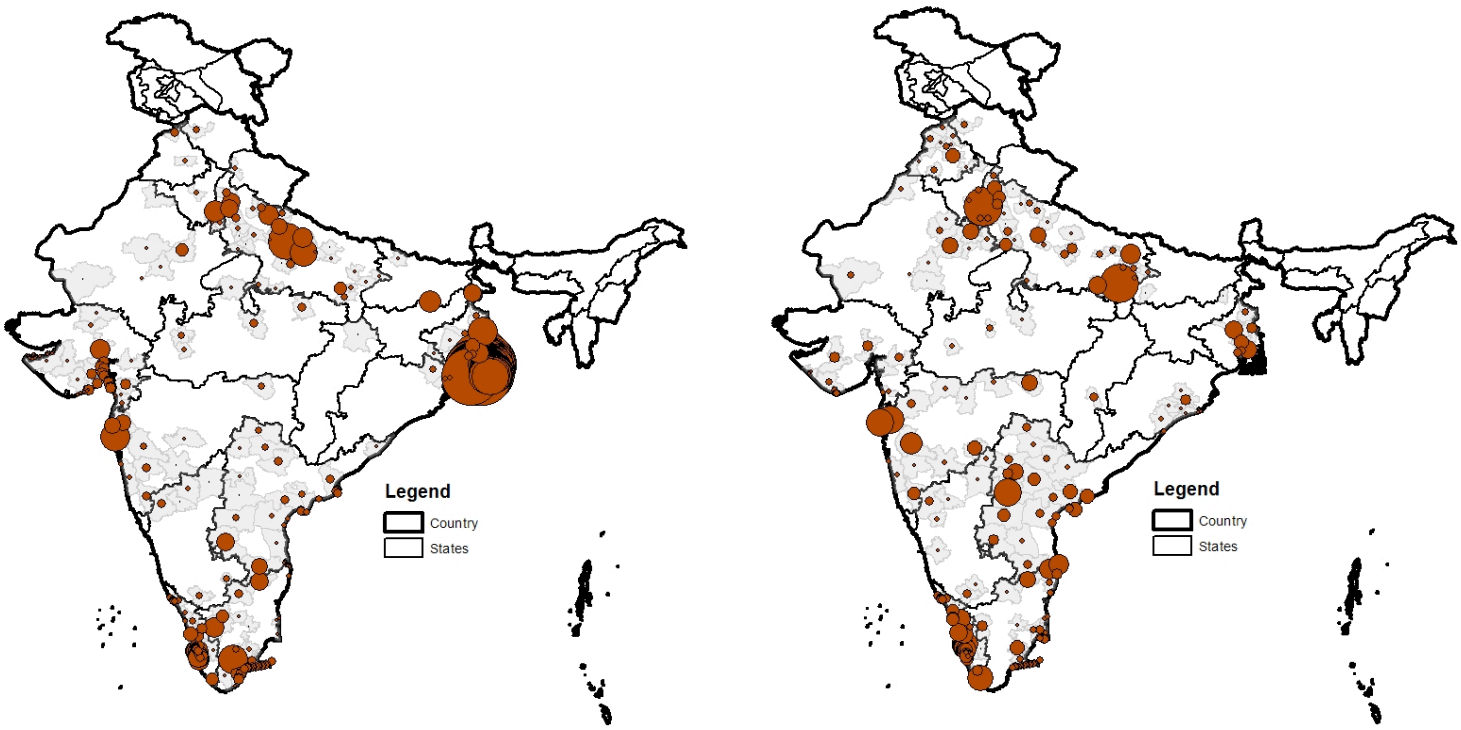
Manufacturing **Services**



Source: Food and Agricultural Organisation (GAUL), Prowess, Census 2001

Figure 4: Distribution of Informal Activity (contribution to the Theil Index)

Manufacturing **Services**



Source: Food and Agricultural Organisation (GAUL), Prowess, Census 2001

Appendix

Ellison and Glaeser (1997) Index:

The EG Index for industry k is equal to:

$$\gamma_k = \frac{G - \left(1 - \sum_j x_j^2\right) H_k}{\left(1 - \sum_j x_j^2\right)\left(1 - H_k\right)}$$

where G for industry k is defined as:

$$G = \left(s_j^k - x_j\right)^2$$

and s refers to the share of total employment of district j for industry k, x refers to the share of district j in total employment, and H is the plant employment Herfindahl index, corresponding to the sum of the squares of the share of employment of each plant over the total employment of the industry.

Theil Index:

The Theil index for specialisation here measures the extent of over or under representation of a district with regards to employment across a set of industries. The value of the index is[20]:

$$T_r = \sum_k \frac{x_{jk}}{x_j}\left(\log\frac{x_{jk}}{x_j} - \log\frac{x_k}{x}\right)$$

where:

x_{jk} refers to employment in industry k in district j
x_j refers to total employment in district j
x_k refers to total employment in industry k
x refers to total employment

Ellison and Glaeser (2010) Coagglomeration Index:

The EG coagglomeration index applies to industry pairs, and for industries i and j it is defined as:

$$\gamma_{ij}^c = \frac{\sum_{m=1}^{M}(s_{mi} - x_m)(s_{mj} - x_m)}{1 - \sum_{m=1}^{M} x_m^2}$$

where m indexes geographic areas (here, districts), s_{mi} is the share of industry i's employment contained in area m, x_m measures the aggregate size of area m (which is modelled as the mean employment share in the district across manufacturing/services industries).

Ellison and Glaeser (2010) Input-Output Index:

$Input_{i \leftarrow j}$ is defined at the share of industry i's inputs that come from industry j.
$Output_{i \rightarrow j}$ is defined as the share of industry i's outputs that are sold to industry j.
To construct a proxy for the linkages between a pair of industries, I follow Ellison et al (2010) and define unidirectional versions of the input and output variables by:

[20] See Brakman et al (2005) for more on the calculation of the index for concentration.

$$Input_{ij} = \max\left\{Input_{i\leftarrow j}, Input_{j\leftarrow i}\right\} \text{ and}$$
$$Output_{ij} = \max\left\{Output_{i\rightarrow j}, Output_{j\rightarrow i}\right\}$$

The combined variable is then defined as:

$$InputOutput_{ij} = \max\left\{Input_{ij}, Output_{ij}\right\}$$

References

Amiti, M. and Javorcki, B.S. (2005), 'Trade costs and location of foreign firms in China', International Monetary Fund Working paper No. 55, Washington DC.

Arrow, K.J. (1962), 'The economic implications of learning by doing', Review of Economic Studies, 29, pp. 155-173.

Bagchi-Sen, S. (1995), 'FDI in US Producer Services: A Temporal Analysis of Foreign Direct Investment in the Finance, Insurance and Real Estate Sectors', Regional Studies, 29(2), pp. 159-170.

Bannerjee, A. and Iyer, L. (2005), 'History, institutions and economic performance: the legacy of colonial land tenure systems in India', American Economic Review, 95 (4), pp. 1190-1213.

Brakman, S., Garretson, H., Gorter, J., van der Horst, A. and Schramm, M. (2005), 'New Economic Geography, empirics and regional policy', Working Paper no. 56, CPB Netherlands Bureau for Economic Policy Analysis.

Brulhart, M. (1998), 'Economic Geography, industry location and trade: The evidence', World Economy, 21(6), pp. 775-801.

Carlton, D. (1979), 'Why new firms locate where they do: an econometric model', in W. Wheaton (ed.), *Interregional movements and regional growth*, Washington DC.

Carlton, D. (1983), 'The location and employment choices of new firms: an econometric model with discrete and continuous endogenous variables', Review of Economics and Statistics, 65 (3), pp. 440-449.

Cheng, L.K. and Kwan, Y.K. (2000), 'What are the determinants of the location of foreign direct investment? The Chinese experience', Journal of International Economics, 52 (2), pp. 379-400.

Chinitz, B. (1961), 'Contrasts in Agglomeration: New York and Pittsburgh', American Economic Review, 51, pp. 279-289.

Coffey. W.J. and Shearmur, R.G. (2002), 'Agglomeration and dispersion of high-order service employment in the Montreal metropolitan region 1981-96', Urban Studies, 39 (3), pp. 359-378.

Deichmann, U., Kaiser, K., Lall, S.V. and Shalizi, Z. (2005), 'Agglomeration, transport and regional development in Indonesia', Policy Research Working Paper No. 3477, World Bank, Washington DC.

Dekle, R. and Eaton, J. (1999), 'Agglomeration and land rents: evidence from the prefectures', Journal of Urban Economics, 46 (2), pp. 2001-2014.

Duranton, G. and Overman, H.G. (2005), 'Testing for localization using micro-geographic data', Review of Economic Studies, 72 (4), pp. 1077-1006.

Ellison, G. and Glaeser, G.L. (1997), 'Geographic Concentration of U.S. Manufacturing Industries: A Dartboard Approach', Journal of Political Economy, 105 (5), pp. 889–927.

Enderwick, P. (1989), *Multinational Service Firms*, Routledge, London

Feser, E.J. and Bergman, E.M. (2000), 'National industry cluster templates: a framework for applied regional cluster analysis', Regional Studies, 34, pp. 1-20

Guimaraes, P., Figueiredo, O. and Woodward, D. (2000), 'Agglomeration and the Location of Foreign Direct Investment in Portugal', Journal of Urban Economics, 47 (1), pp. 115-135.

Guimaraes, P., Figueiredo, O. and Woodward, D. (2003), 'A Tractable Approach to the Firm Location Decision Problem', Review of Economics and Statistics, 85 (1), pp. 201-204.

Guimaraes, P., Figueiredo, O. and Woodward, D. (2004), 'Industrial Location Modeling: Extending the Random Utility Framework', Journal of Regional Science, 44 (1), pp. 1-20.

Hanson, G.H. (1959), 'How accessibility shapes land use', Journal of the American Institute of Planners, 25, pp. 73-76.

Hardin, S. and Carroll, R. (2003), 'Instrumental variables, bootstrapping, and Generalized Linear Models', Stata Journal, 3(4), pp. 351-360.

Head, K. and Mayer, T. (2004), 'Market potential and the location of Japanese investment in the European Union', Review of Economics and Statistics 86(4), pp. 959-972.

Head, H. and Reis, J. (1996), 'Inter-city competition for foreign investment: Static and dynamic effects of China's Incentive', Journal of Urban Economics, 40 (1), pp. 38-60.

Henderson, J.V. (2003), 'The urbanization process and economic growth: the so-what question', Journal of Economic Growth, 8, pp. 47-71.

Jacobs, J. (1969), *The Economy of Cities*, MIT Press: Cambridge.

Kirn, T.J. (1987), 'Growth and change in the service sector of the United States: a spatial perspective', Annals of the Association of American Geographers, 77, pp. 353-372.

Krugman, P. (1991), 'Increasing returns and economic geography', Journal of Political Economy, 99 (3), pp. 483-499.

Lafourcade, M. and Mion, G. (2003), 'Concentration, spatial clustering and the size of plants: disentangling the sources of co-location externalities', CORE Discussion Paper No. 2003/91, Belgium.

Lall, S.V., Koo, J. and Chakravorty, S. (2003), 'Diversity Matters: The Economic Geography of Industrial Location in India', Policy Research Working Paper 3072, World Bank, Washington DC.

Lall, S.V. and Chakravorty, S. (2005), 'Industrial Location and spatial inequality: Theory and Evidence from India', Review of Development Economics, 9 (1), pp. 47-68.

Lall, S.V., Shalizi, Z. and Deichmann, U. (2004), 'Agglomeration economies and productivity in Indian industry', Journal of Development Economics, 73 (2), pp. 643-673.

Lall, S.V. and Mengistae, T. (2005) 'The impact of business environment and economic geography on plant-level productivity: an analysis of Indian industry', Policy Research Working Paper No. 3664, World Bank, Washington DC.

Marjit, S. and Kar, S. (2009) 'A contemporary perspective on the informal labour market: theory, policy and the Indian experience', Economic and Political Weekly, XLIV (14), pp. 60-71.

Marshall, A. (1890), *Principles of Economics*, Macmillan: London.
Marshall, A. (1919), *Industry and Trade*, Macmillan: London.
McDowell, A. (2003), 'From the Help Desk: Hurdle Models', Stata Journal, 3 (2), pp. 178-184.
McFadden, D. (1974), 'Conditional logit analysis of qualitative behaviour', in P. Zarembka (ed.), *Frontiers in Econometrics*, Academic Press: New York.
McGee, T.G. (1977), 'Hawkers in South East Asia Cities: Planning for a Bazaar Economy', International Development Research Centre, Ottawa.
Mullahy, J. (1997), 'Instrumental-variable estimation of count data models: applications to models of cigarette smoking behaviour', Review of Economics and Statistics, 79(4), pp. 586-593.
Mukim, M. and Nunnenkamp, P. (2010), 'The location choices of foreign investors: a district-level analysis in India', Working Paper No. 1628, Kiel Institute for the World Economy, Germany.
Nichols, A. (2007), 'IVPOIS: Stata module to estimate an instrumental variables Poisson regression via GMM, Statistical Software Components S456890, Boston College Department of Economics.
Romer, P.M. (1986), 'Increasing returns and long-run growth', Journal of Political Economy, 94 (5), pp. 1002-1037.
Sakhthivel, S. and Joddar, P. (2006), 'Unorganised sector workforce in India: trends, patterns and social security coverage', Economic and Political Weekly, May 27, pp. 2107-2114.
Venables, A.J. (1996), 'Equilibrium locations of vertically linked industries', International Economic Review, 37 (2), pp. 341-359.

Spatial Economics Research Centre (SERC)
London School of Economics
Houghton Street
London WC2A 2AE

Tel: 020 7852 3565
Fax: 020 7955 6848
Web: www.spatialeconomics.ac.uk

SERC is an independent research centre funded by the Economic and Social Research Council (ESRC), Department for Business Innovation and Skills (BIS), the Department for Communities and Local Government (CLG) and the Welsh Assembly Government.