

Essays in Applied Spatial Economics

A thesis presented by

Carlo Menon

to

The Department of Geography and Environment

London School of Economics and Political Science

in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

London, United Kingdom

December 2009

UMI Number: U615712

All rights reserved

INFORMATION TO ALL USERS

The quality of this reproduction is dependent upon the quality of the copy submitted.

In the unlikely event that the author did not send a complete manuscript and there are missing pages, these will be noted. Also, if material had to be removed, a note will indicate the deletion.



UMI U615712

Published by ProQuest LLC 2014. Copyright in the Dissertation held by the Author.
Microform Edition © ProQuest LLC.

All rights reserved. This work is protected against
unauthorized copying under Title 17, United States Code.



ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

TRESEJ
F
9274



1242521

Declaration

I certify that the thesis I have presented for examination for the MPhil/PhD degree of the London School of Economics and Political Science is solely my own work other than where I have clearly indicated that it is the work of others (in which case the extent of any work carried out jointly by me and any other person is clearly identified in it).

The copyright of this thesis rests with the author. Quotation from it is permitted, provided that full acknowledgement is made. This thesis may not be reproduced without the prior written consent of the author.

I warrant that this authorization does not, to the best of my belief, infringe the rights of any third party.

Statement of conjoint work

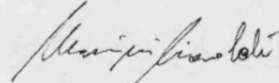
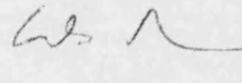
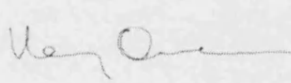
London, 26 November 2009

This is to certify that Chapter 2 of the thesis by Carlo Menon titled "Does urbanisation affect rural poverty? Evidence from Indian districts" is the product of conjoint work with Massimiliano Cali (Dept. of Geography and Environment, London School of Economics). While both students have contributed to all the sections of the work, Massimiliano has mainly focussed on the theory and methods of analysis (sections 2.1-2.3), and Carlo has mainly focused on the variable construction and the empirical analysis (sections 2.4-2.5).

Henry G. Overman

Carlo Menon

Massimiliano Cali



Abstract

The thesis is composed of four chapters, which investigate different topics in the field of applied urban and spatial economics.

The first paper develops an original empirical approach to investigate the role played by labour markets in explaining the pattern of industrial agglomeration in the United States. The methodology allows us to i) obtain an estimate of industrial agglomeration which significantly improves on existing indices, and ii) provide a ranking of industries according to their responsiveness to labour market determinants of agglomeration. Results show that labour market determinants explain around one quarter of the variation in spatial agglomeration across industries.

The second paper assesses whether urbanization alleviates rural poverty in surrounding areas in India, using a panel dataset at district level for the period 1981-1999. We find that the effect is substantial and systematic; this is largely attributable to positive spillovers from urbanisation, rather than to the movement of the rural poor to urban areas per se.

The third paper investigates an extremely peculiar characteristic of the US patent dataset: there is a large group of inventors who develop one or a few patents during a long period of analysis ("comets"), while a very small group of "stars" inventors develop a huge number of patents. In light of that, the paper first explores the location pattern of comets and stars, and then assesses whether the activity of star inventors is beneficial to the production of comet patents in the same city and technological category.

The fourth paper describes the effects of bank liberalization on the geographical penetration of branches in the city of Antwerp (BE). Our results show that, coinciding with the strongest wave of the deregulation and concentration process, banks systematically exit from low income neighbourhoods.

Acknowledgments

There is a long list of people which I sincerely want to thank for having helped me in reaching the final stage of my PhD experience.

In chronological order, the first is Stefano Magrini. He first spotted some research potential in me, even if I almost failed the exam of his undergraduate course at the University of Venice. Since then, he has always been supportive; he has been the first person to ask for frank advice on research issues, as well as on career decisions (or a long-distance holiday by motorbike).

My PhD supervisor Henry Overman is also responsible for a crucial contribution to my thesis, and, more generally, for my personal growth as a researcher. Since the very first meeting, he taught me how to shape sensible research questions (and how to kill rapidly the less sensible ones). I learned a lot from both his hints and his criticisms, and I really appreciate the genuine interest and the time he devoted to my work. As an economist, I feel now radically transformed compared to that day when I entered his office for the first time, and this is mostly because of his support.

Many other people among the departmental staff have been supportive for my research: Steve Gibbons, Ian Gordon, Paul Cheshire, Olmo Silva, Riccardo Crescenzi and Giordano Mion have always been helpful and generous with hints and suggestions. My PhD colleagues in the department Rodrigo Alegria, Alejandra Rodriguez, Roberto Picchizolu, and Filipe Lage de Sousa provided another invaluable source of support, especially in the first year of my PhD; in the following two years, frequent discussions with Rosa San-

chis, Neil Lee, Max Nathan, Massimiliano Cali, and Felix Weinhardt played a major role in shaping the papers which are included in my thesis. Massimiliano deserves a further acknowledgment for having involved me in our common project on urbanization in India: the (many) days we spent together cleaning data, running regressions, and discussing results are among the most precious I remember of my PhD. Similarly, my coauthors Marieke Huysentruyt and Eva Lefevere offered me another invaluable opportunity for learning by working together.

Writing a (hopefully good) thesis requires a significant degree of personal serenity and, possibly, happiness. Thus, the many pages which follow are infused with the support of my family: little Elena, who bore the sacrifice of passing her childhood with a older brother which was always abroad, and who never agreed to hiding her inside his cabin luggage; my brother Daniele, who has always been keen and quick in sharing with me his Matlab wisdom by email; my mother Elisabetta, for standing my frequent mental absence, and the even more frequent physical one; and my father Claudio, who always encouraged me to become an academic (probably partly because he does not exactly know what it means).

The final though is for Anna: there are no words to thank her enough for the unconditional support and faith with my research work, and - more simply - for having accompanied me into the happiest period of my life.

Table of Contents

Abstract	4
Acknowledgments	6
Table of Contents	8
Introduction	12
Spatial economics: anatomy of a discipline	12
Methodological issues in spatial economics	18
Overview of the thesis	27
Final remarks	35
1 The Bright Side of MAUP: An Enquiry into the Determinants of Industrial Agglomeration in the United States	36
1.1 Introduction	36
1.2 The Determinants of Concentration	39
1.3 The MAUP and the Gerrymandering Approach	44
1.3.1 The Dark Side	44
1.3.2 The Bright Side	46
1.3.3 A Real World Example	53
1.4 Building the Counterfactuals	54
1.4.1 The Noise	54
1.4.2 The Pseudo Statistical Areas (PSAs)	58

1.5	Results	61
1.5.1	Interpretation	62
1.5.2	Testing the significance	65
1.5.3	Discussion of findings	66
1.5.4	Comparison of the “CBSA minus Noise” values with the Ellison-Glaeser Index	69
1.5.5	Analysis at 4-digit level	74
1.5.6	Does the labour market matter for concentration?	75
1.5.7	Further robustness: regression of the CBSA-PSA difference on industry variables	77
1.6	Conclusions	78
1.A	The PSA algorithm	80
1.B	OLS regression	85
1.C	Data	87
2	Does urbanisation affect rural poverty? Evidence from Indian Districts.....	89
2.1	Introduction	89
2.2	Urbanization and rural poverty: Channels.....	94
2.2.1	Second round effects.....	95
2.2.2	Disentangling first and second round effects	101
2.3	Empirical methods	104
2.3.1	Data and variables	107
2.3.2	Results.....	116
2.4	Conclusions	135

2.A	Methodological note to the construction of poverty measures	138
3	Stars and Comets: an Exploration of the Patent Universe	140
3.1	Introduction	140
3.2	Patents, localized knowledge spillovers, and the size of innovation	142
3.3	Stars and Comets	147
3.3.1	Preliminary evidence on location of stars and comets	154
3.3.2	Why should stars positively affect comets?	159
3.4	Analysis	161
3.4.1	Instrumental Variable Estimation	164
3.5	Results	168
3.5.1	Robustness tests	171
3.6	Conclusions	177
3.A	Data	179
3.B	Alternative definitions of comets and stars	181
3.C	IV estimation diagnostics	187
3.D	Ancillary tables and results	191
4	Bank Location in the city of Antwerp: Evidence from Microgeographic Data	195
4.1	Introduction	195
4.2	Background and Data	198
4.2.1	Deregulation in the Belgian retail banking industry, 1991 - 2006	198
4.2.2	The city of Antwerp	200
4.3	Data	202

4.3.1 Measure of bank presence, entry, and exit	205
4.3.2 Measures of bank "desert" and bank choice	209
4.3.3 Neighbourhood characteristics	214
4.4 Empirical Analysis	215
4.4.1 Patterns of Bank Location.....	215
4.4.2 Bank desert and bank choice	229
4.5 Conclusion.....	234
4.A Regressions using bank counts	236
4.B Spatial econometric models of desert and choice variables	237
5 Conclusions	242
References	256
List of tables	269
List of figures	273

Introduction

The aim of this introduction is to present a critical overview of a relatively young and undefined discipline: spatial economics. In particular, we focus on the relations with other disciplines (including economic geography and mainstream economics), on data problems, and on identification issues. In doing so, we discuss the most important methodological issues the discipline faces, building on the experience acquired from the applied works presented in the other chapters of the thesis. Finally, we conclude the introduction with a quick overview of the contents of the thesis.

Spatial economics: anatomy of a discipline

Spatial economics is a border discipline between mainstream economics and economic and human geography, although the communication between the two academic groups is rather difficult and infrequent. Recently, a number of contributions have explored the scope for cooperation between spatial economics and economic geography (e.g. Overman, 2004; Duranton and Rodriguez-Pose, 2005; Duranton and Storper, 2006). In what follows, we try to offer an original view on the issue, arguing that although the difference in the research methods of the two disciplines is probably too big to be ever dissolved, nevertheless spatial economics can profitably exploit the similarity of interests with economic geography. Related to that, we will also briefly discuss some methodological issues in the discipline.

Economic geography and spatial economics

In this section we briefly outline the core methodological differences between the two disciplines, which may explain why the dialogue between mainstream economic geography and mainstream spatial economics is so difficult and infrequent. In the following discussion, two *caveat* are needed: first, economic geography and spatial economics are both dense and composite disciplines, and there surely are branches and authors in both the disciplines which do not fit the generalization we propose; second, the epistemological debate we briefly summarize is of course much more complex than it may look from the following lines. We do not aim at a complete treatment of the issue, we just want to sketch the conceptual frame of the discussion necessary to further the discussion, and to suggest why we think that spatial economics may speak to a wider audience than economic geography.

In order to synthetically illustrate the milestones of the research methods in mainstream economic geography, we report a couple of short quotations from David Harvey, probably the most influential economic geographer of the last three decades:

I suspect [...] theory is all too often understood as a bundle of stationary, already fully specified arguments and propositions, ready-made to be applied to and tested against the real world. [...] Theory should be understood instead as an evolving structure of argument sensitive to encounters with the complex ways in which social processes are materially embedded in the web of life. (Harvey, 2006, p.78-79)

We reach out dialectically (rather than inward deductively) to probe uncharted seas from a few seemingly secure islands of concepts. (Harvey, 1985, p. XVI)

From these quotations, we learn that economic geography analyses relies on a *a priori* background theory which is applied to, rather than tested by, the observation of reality. It is also the only way by which we can obtain a general rule from one or few cases, because the background theory is necessary in order to disentangle the contingent from the

independent, the particular from the general. Indeed, how could a researcher identify the universal “guiding forces” from just a case study? How can she/he prove their universality? It is only possible if our observations are theoretically interpreted. This is what makes the Harvey’s approach different from the “scientific method”.

An economist would object that the weakness of this approach stems from the failure to solve the doxa/epistème opposition, i.e., the net separation between the opinion and the proved knowledge. Therefore, any advancement of the theory risks to be, or at least to appear, tautological and self-referential.

Economists may seem equally or even more fundamentalist about their research methods, but their "positivist" approach (by which we mean: testing a theoretical predictions using real world observations), instead, do begin from a *tabula rasa*, i.e., the specific theoretical hypothesis under test are formally not requested by the testing procedure. In reality, this is variably true, but economists’ papers are generally clear in stating which assumptions are needed to reach a given results. In passing, however, we note that economists are often reluctant to discuss the epistemology of their research methods, and the most popular PhD programs in economics do not include any formal training in the subject.

According to Kuhn (1962), the scientific method in general is incapable of final empirical validation because facts are not independent from theories used to test them. Harvey goes far beyond, claiming that every scientific method is ideological:

Scientific method, it is often argued, guarantees the objectivity and ethical neutrality of factual statements [. . .]. The peculiarity of this view is that the claim to be ethically neutral and ideologically free is itself an ideological claim. [. . .] I am arguing that the use of a particular scientific method is of necessity founded in ideology. (Harvey,

1974, p. 256)

In other words, every scientific method deepens its roots on the ground of seemingly secure islands. Formally, this is very hard to object: even our actual knowledge of the reality grounds on the hypothesis of the trueness of our sensorial perceptions. This, in turn, will make vain the economists' effort to support the universality and neutrality of their approach.

Nevertheless, we can still speculate on the nature of the islands. In an economics-style scientific method, the position and the genealogy of islands are declared, while in a dialectical context their existence is tentative – in the first case, they are indicated in the naval map and their coordinates are given, while in the second case their existence is based on the evolving tales of old sailors we meet at night in the port.

Therefore, even though we admitted that every scientific method is ideological, we should reflect on the order of magnitude of the ideological content: this being not a subjective judgment of value, but rather an objective assessment of their potentiality of seeing our results accepted by a wider audience.

Spatial economics, the real world, and the geography redemption

So far, we argued the advantage of the “economic method” based on being less ideological and thus accepted (or acceptable) by a wider audience. But it comes at a cost: the emphasis on building rigorous empirical methodologies very often requires an abstraction from the real world. To many, economics is getting further away from the real world, and spatial economics may not be exception. The sentence “for economists, real life is a special case”

is ceasing to be a joke to become a serious truth, and even (or especially) the most influential theorists seem to have a rather short view on the rules which govern the real, everyday world.¹ However, direct experience of real world is a crucial aspect for the construction of our theories.

Spatial economics may have an advantage in this field, as compared to other branches of economics: the production of case studies by economic geographers on similar fields of research can be a fruitful source of information on the real world. The mainstream economist may be inspired by the eclectic and "holistic" view of the geographer and include in their rigorous analysis some elements which are not generally contemplated in economics, like e.g. political powers or relational capital. On the theoretical side, it may foster the study of economic models which can better mimic the real world; on the empirical side, it can help economists to understand which are the limits and the resources of the real world data we use for their empirical works.

A well-known example is the book by AnnaLee Saxenian (1994) on innovation culture in the Silicon Valley: a classic geographic-style case study based on more than one hundred interviews, it is now cited in almost any economic paper on industrial agglomeration or localized knowledge spillovers. In her book, Saxenian compared the California's Silicon Valley with the Route 128 in Massachusetts: while the two areas were similarly specialized in electronics during the 70s, they faced rather different fortunes in the 90s, with the former becoming the world center for semiconductor design, and the latter seeing a season of relative decline. According to the author, the main reason of the success

¹ The recent financial crisis has further challenged the reputation of economics as a discipline, especially in the field of macroeconomics and finance (cf. the Economist, 2009).

of businesses located in California's Silicon Valley in the '90s is due to the fact that the area "developed a decentralized but cooperative industrial system, while Route 128 came to be dominated by independent, self-sufficient corporations". Saxenian's argument has been integrated into several econometric works, which tested whether small firms are more innovative, and more incline to cluster and network. On the other side, these sorts of organizational differences are difficult to quantify, which creates the need for the case-study approach (Rosenthal and Strange, 2004).

Another interesting example is the adoption of Social Network Analysis tools - widely used in human geography and sociology to map and quantify relational capital - in spatial economics papers: for instance Breschi and Lissoni (2009) assess to which extent the evidence of localized knowledge spillovers is due to market mediated professional networks, rather than non-priced informal externalities. By applying social network analysis to patent data in the US, they find that, after controlling for inventors' mobility across workplaces, the residual effect of spatial proximity on knowledge diffusion is greatly reduced, as compared to results of previous works on the subject. In our opinion, this is a good example of how tools which are not part of the traditional economist's toolbox can be borrowed from sister disciplines in social sciences to provide solid evidence on topics at the top of the (spatial) economics research agenda.

The fourth section of the handbook chapter of Rosenthal and Strange (2004) also provides some good examples on how case studies may integrate a "regression approach" to agglomeration economics. Beyond Saxenian, they also describe the study of the New York Metropolitan Region by Hoover and Vernon (1959), which contained some original (at the

time) ideas on external economies and frequent face-to-face communication between small businesses; according to the authors, they constitute the main advantages of urbanization. The conclusion which followed from the study was that small firms and large cities should be strongly associated. Rosenthal and Strange however highlight that Holmes and Stevens (2000) investigated the relationship across industries and cities with an econometric approach and found evidence going in the opposite direction, i.e., large firms are more likely to be found in clusters. According to Rosenthal and Strange, therefore, this is an example of the danger of generalizing case study findings, which often focus on cases and localities which are the *exception* rather than the *rule* (as stressed also by Overman, 2004).

Summing up, spatial economists may profit by the body of research produced by geographers on a number of topics and issues which are of interest for both the disciplines. Geographers' case studies may be extremely useful to acquire a deeper understanding of the real world, to take into consideration forces and factors which generally escape economic modelling, and to integrate into economics frameworks and tools from other disciplines. At the same time, this will not necessarily imply an acceptance of the validity of general theory based on a small number of case studies, nor of the "exceptionalistic claim", i.e., general theories are meaningless, as everything is an exception.

Methodological issues in spatial economics

In the following, we briefly discuss some of the most important and urgent issues in spatial economics: the Modifiable Areal Unit Problem (MAUP), the sourcing of data, and the identification strategy.

The Modifiable areal unit problem (MAUP)

The "modifiable areal unit problem" arises from the large and unpredictable variation that economic estimates may have depending on the size and shape of the adopted spatial units. Every geographical area may be divided in many different ways, and commonly used statistics may present huge variation among them. Furthermore, differently from international comparison (where country borders have an economic and political meaning that is not comparable with any other geographic classification) in a sub-national setting researchers face a variety of administrative and functional divisions – each of them with its pros and cons – with the result that the choice of the spatial unit is often arbitrary, or constrained by data availability.

The MAUP has been widely investigated and debated: the first contributions go back to the '30s and - more recently - the work by Openshaw (1984) Arbia (1989) and many others have provided a clear and precise insight on the issue. However, the interest in the issue has been mostly confined to the field of quantitative geography, while in spatial economics the MAUP has seldom been taken into explicit consideration. A few meritorious exceptions are the series of papers by Cheshire and co-authors based on the Functional Urban Regions (e.g. Cheshire and Magrini, 2009; Cheshire and Hay, 1989), in which the spatial units are meant to be "geographically meaningful" in relation to the enquired phenomenon, or studies based on continuous definition of space (e.g., Duranton and Overman, 2005; Marcon and Puech, 2003).

However, in general the attempts to offer adequate solutions on the data side have been limited and isolated, especially on the European context. The impression is that the MAUP is, by now, a dark shadow nestled in the subconscious of the discipline.

We do not necessarily think that the MAUP is always a serious bias for spatial economics estimates. It could be possible, as Briant et al. (2007) maintain, that in some circumstances the problem is negligible. The issue, in our opinion, is more general: in every spatial economics paper, we should be able to explain why the MAUP may, or may not, be a concern, rather than hoping that the bias is not that big. Furthermore, we should reflect carefully on the "spatial extent" of the phenomenon under scrutiny and on the properties of the adopted spatial units. This is very seldom done, and it is unfortunate, as it could disclose new, interesting directions of investigation. Furthermore, this becomes even more paradoxical in the light of the huge efforts spent in improving the efficiency of econometric estimates, e.g. by means of complex econometric models able to control for a spatial autoregressive structure in the error terms. Probably, the same amount of efforts spent on improving the quality and appropriateness of data would be much more rewarding in terms of precision and reliability of results.

A careful consideration on the MAUP is also crucial for policy evaluation. Regional policies always target spatial units, thus an assessment of which administrative level is the best to "contain" a given economic socio-phenomenon is extremely valuable to inform policy makers. To the best of our knowledge, there is little discussion in the literature about that. A step in this direction can be found in Cheshire and Magrini (2009), which build a policy incentive variable based on "the ratio of the total population of the largest (relevant)

jurisdiction representing the FUR to the population of the FUR as a whole" (p. 101), where the FUR (functional urban region) is the adopted spatial unit of analysis. The authors find the variable to have a positive effect on the economic growth of European cities from 1978 to 1994, which suggests that a close geographical matching with the administrative entity bearing local political power is beneficial for the FUR. This may be due to the fact that the effects of policies do not spill-over to areas where voters and stakeholders do not reside, and therefore all the benefits are internalized (which in turn would imply better policies).

Finally, spatial economists should also keep in mind that an awareness of the MAUP and, more generally, of the peculiarities of working with geographical and sub-national data should be considered to be a strategic advantage as opposed to other economists dealing with the same kind of data. Rather than ignore them, they should stress these problems in order to highlight their distinctive contributions.

The data

From the previous discussion we learnt that the replicability of results is one of the milestones of the scientific method in economics. Therefore, our empirical methods should be clearly explained, in order to allow every interested reader to replicate and verify our analysis. However, for an empirical analysis to be replicated knowing the methodology is not enough - we also need the data. Therefore, using publicly available data – and sharing the database once the research is published – is reinforcing the scientific solidity of a research.

Unfortunately, however, at subnational scale publicly available data for some macroregions - namely Europe - are scarce and of bad quality (Combes and Overman, 2004, offer

a good discussion of the issue), which is particularly unfortunate, also in the light of the previous discussion on the MAUP. In recent years, the situation has slightly improved in terms of data availability, as Eurostat published many new series starting from the early 2000s. Nevertheless, it has remained steadily critical in terms of spatial classification and coverage. The spatial classification is entirely based on administrative units which are extremely heterogenous among EU member countries in several dimensions (size, political and economic meaning, spatial extension, etc.). Just to give a simple example, although the Eurostat webpage describing the NUTS classification claims that at the NUTS3 level the total population spans between a minimum of 150,000 and a maximum of 800,000,² a rapid check with Eurostat demographic database³ shows that 19 NUTS3 of many different countries are actually over 2 million of population (and for only a fraction of them the NUTS3 coincides with the corresponding NUTS2), while 45 units of the same classification do not reach 50,000 inhabitants. Considering that the variability in population is just a symptom of more general sources of heterogeneity (e.g. rural-urban), there is probably no need to say that results based on such a sample may be highly misleading in many circumstances.

There are, however, also some good news about data. The collection, visualization, storing, elaboration of micro(geographic) data has been massively affected by the progress of computer science, and new datasets offer many valuable opportunities to researchers in spatial economics. Contrary to other fields of economics, it seems that recent research in spatial economics is not taking fully advantage of that. For instance, empirical papers

² http://epp.eurostat.ec.europa.eu/portal/page/portal/region_cities/regional_statistics/nuts_classification visited on the 13th of August 2009

³ The dataset we used is the "Annual average population by sex (reg_d3avg)", freely available from the Eurostat website.

using remote sensing or land use data are very rare, although these sources offer unprecedented information, in terms of geographic detail and coverage. The discipline is putting unbalanced efforts in refining theoretical and empirical methodologies, as compared to the efforts spent on improving the quality of the used data, although there are many examples showing that the latter way is also extremely rewarding in terms of advancements in the discipline. One of these is the work by Burchfield et al. (2006), which is based on a grid of 8.7 billion 30x30 meter cells obtained by merging high altitude photographs for previous periods (around 1976) with satellite images for the '90s. Building on these data, the authors were able to assess with extreme precision the determinants of urban sprawl in US cities, eventually finding that ground water availability, temperate climate, rugged terrain, decentralized employment, early public transport infrastructure, uncertainty about metropolitan growth, and unincorporated land in the urban fringe are all factors fostering sprawl. Another interesting example is the very recent working paper by Henderson et al. (2009), which combines night light satellite digital maps with rainfall data, in order to assess the effects of climatic shocks on urban growth in Africa at a very detailed spatial scale. It is worth noticing that both the papers are rather simple in terms of identification strategy, but they are able to disclose new and relevant information thanks to the accuracy and novelty of data.

Geographical data elaborated through a Geographic Information System (GIS)⁴ offer also huge opportunities for instrumental variables (as many prominent development economists have recently realized). For instance, Duflo and Pande (2007) assess the effects

⁴ For a detailed discussion of application of GIS in economics, see Overman (2008).

of dams on agricultural productivity across Indian districts. To overcome a potential reverse causality issue, they argue that river gradient affects a district suitability for dams and therefore use the latter variable - appropriately measured with a GIS - as an instrument in a two-stages least squared regression. Rosenthal and Strange (2005) instead address a classical urban economic question, i.e., whether density and proximity to human capital enhance productivity, using geologic features (landslide hazard, seismic hazard, bedrock, and surface water) as instruments to address measurement error in agglomeration variables and endogeneity in the wage-agglomeration relationship. In this case, the validity of the instrument is given by the fact that tall buildings need specific geologic conditions to be built; therefore, geologic variables are correlated with density, but arguably have no independent effect on productivity.

Furthermore, geolocated micro-geographic data also allow to assess the *distance decay* of many economic phenomena. For instance, Duranton and Overman (2005), using detailed information on the location of British manufacturing plants, find that most of the industry agglomeration take place within 50 km. Henderson and Arzaghi (2008), instead, study agglomeration effects on productivity on a sample of advertising agencies located in Manhattan, finding that these effects are strong but dissipate at a distance of around 750 meters.

Finally, there is a last consideration to mention about this kind of geophysical variables which are becoming increasingly available to the researchers thanks to remote sensing digitalized maps or other sources. Very often, these variables are better measured, more easily understandable, and more meaningful than most of the commonly used variables in

spatial economics (e.g. regional GDP, regional Gross Value Added, industry employment or shipment). It follows that analyses based on geophysical data may be more reliable, appealing, and comprehensible than traditional papers based on mainstream economic measures collected, with many efforts and huge costs, by national statistical offices.

The quest for identification

As in other fields of economics, a solid identification strategy, able to provide solid evidence on the direction of causality of the inferred association, is extremely rewarding in terms of article publishability. Beyond traditional identification issues (e.g. reverse causality, omitted variables, and the like), in spatial economics a common challenge to a clean identification strategy is the presence of unobserved spatial factors creating spurious correlation between the dependent and independent variables. E.g., we may observe that housing prices are higher in neighbourhoods with better schools, and thus we may suppose that good schools *cause* higher housing prices; but it could also be that more expensive neighbourhoods are inhabited by more pushy parents, which is unobserved and may have a positive effect on pupils' performance at school. Therefore, the effect of school quality on housing price may be overestimated.

Popular identification strategies in spatial economics aimed at controlling for spatial unobservables include fixed effect regressions and the spatial discontinuity approach. To the extent that often datasets in spatial economics are large, fixed effects for spatial units at wider level of classifications in a cross-section, or for every spatial unit in a longitudinal panel, are an easy way (but sometimes too approximate) to control for some of the spa-

tial unobservables. Spatial discontinuity approaches, instead, are based on the assumption that while spatial unobserved factors vary continuously across space, the variable of interest may present some discontinuities, e.g. due to administrative borders. Probably the most popular application of this strategy is based on school district boundaries: in a few countries, including the US and the UK, pupils are forced to attend the school to which their house is allocated to. To the extent that neighbourhood social composition and unobserved amenities are not affected by the district border, by restricting the sample to observations close to the border and including a border dummy (or differencing the variables across the borders) it is possible to eliminate most of the spatial unobservables, but still keeping the variation in school quality (Black, 1999, constitutes the seminal paper in the field). The methodology has also been used to estimate the effects of state policies on the location of manufacturing in the US (Holmes, 1998) and the impact of local taxation at municipality level on the location and growth of firms in the UK (Duranton et al., 2007).

The “quest for identification” is overall beneficial to the discipline but does not need to become an obsession. The questions which can be answered with a clean identification strategy are necessarily limited, and sometimes the answers are obvious or irrelevant. Furthermore, we should admit that sometimes correlation can be extremely informative as well. In a few circumstances, knowing that two things are happening “together” is already an improvement of our knowledge of the real world – the issue is to qualify the “together” adding the right temporal and especially spatial coordinates. In addition to identification and causality, this should be the ultimate aim of spatial economics – which is the spatial decay of economic phenomena? Which are the reasons of such a spatial decay?

In the discipline, we probably need more attention about the spatial units we use, and their economic meaning. Spatial units are not neutral to spatial economic analysis – understanding at which spatial scale two socioeconomic variables are associated may be as much important as understanding the direction of the causality link.

We have already mentioned that empirical methods may often be misleading because of spurious correlation, which may depend on simultaneity, omitted variables, or poor proxies. In a spatial setting, poor proxies are important sources of bias. We are not working with rate of saving and interest rate. We are working with rural poverty, urbanization, innovation, knowledge spillovers, agglomeration externalities, etc.: all these variables are hard to measure and therefore the proxies we use need to be clearly explained and discussed.

In this thesis the author put a lot of efforts in corroborating the robustness of the findings. The direction of causality is always deeply investigated and, when necessary, the instrumental variable estimation is adopted in order to minimize endogeneity biases. At the same time, this is always seen as a means, rather than an end.

Overview of the thesis

The thesis is composed by four papers. Although they investigate four different topics and use four different data sources, they share some common factors.

First, they are all informed by the epistemological debate which we exposed in the previous section. Simple theoretical hypotheses are discussed, and empirical methodologies are developed accordingly, clearly stating all the assumptions we need in order to transform our results into insights on the theories we are testing.

Second, they pose conceptually simple and clear research questions in the field of urban economics – broadly defined – which have not been satisfactorily addressed by previous literature. In particular, the first three papers investigate different kind of urban externalities - labour market pooling, urbanization spillovers in rural areas, and knowledge spillovers, respectively; while the fourth paper relates to the effects of neighbourhood segregation in cities.

Third, an appropriate empirical method is delineated in order to find evidence able to provide a clear answer to the research questions under scrutiny. These methods may either be standard econometrics analysis, as in the second, third, and fourth paper, or original techniques developed specifically for the given research question, as it is the case in the first paper.

Fourth, particular attention is devoted in all the papers to the "spatial nature" of the data and of the phenomena under scrutiny. For instance, in the first paper we analyze the effect of labour market pooling as a determinant of industrial agglomeration and we recognize that the spatial extension of commuting flows is the best approximation of different labour market areas. In the second paper, we argue that the Indian district is a suitable spatial units for assessing the kind of urban externalities we outline in the theoretical part. In the third paper, we assume, and partially test, that the effect of knowledge spillovers is mostly relevant at metropolitan area level, and we therefore adopt the Metropolitan Statistical Areas as the unit of analysis.

In the following, we present a brief overview of the individual papers included in the thesis.

First paper

The first paper investigates the role played by the labour market in explaining the pattern of industrial agglomeration in the United States. The paper stems from a careful examination of two major issues in the analysis of the determinants of industrial concentration: the weakness of a parametric approach, and a Modifiable Areal Unit Problem (MAUP).

From a theoretical point of view, industrial agglomeration is a rather simple issue. Alfred Marshall (1890) theory based on knowledge spillovers, input-output linkages, and labour market pooling is still the most cited reference on the topic. However, empirical evidence is far from being conclusive, especially regarding the role played by the different determinants listed by Marshall. The need for more (and different) empirical research on the topic is therefore urgent. Ellison et al. (2009) paper, forthcoming in the *American Economic Review* (developed contemporaneously to our paper) is probably the first systematic attempt on this direction providing robust evidence on the issue; however, it investigates co-agglomeration patterns (how different industries concentrate in space), rather than agglomeration patterns (how plants of the same industry concentrate). Previous empirical investigations (e.g.: Rosenthal and Strange, 2001 and 2004; Ellison and Glaeser, 2001) approached the issue by regressing the degree of industries' concentration on proxies of input intensities. As we discuss in the paper, these models explain a very limited amount of the (large) variation of the dependent variable, thus results are not conclusive - which in turn cast doubts on the scope for a parametric approach.

At the same time, another empirical weakness stems from the Ellison and Glaeser (EG) index, which is used in the mentioned papers to quantify concentration patterns. Al-

though the EG index represents a significant improvement on the so-called "first generation concentration index", and it has been widely accepted as the standard method to measure industrial agglomeration, it is not free from criticisms. In the paper, we provide an extensive overview of its limits.

The second main issue which the paper deals with is the MAUP. In the context of industrial agglomeration, the MAUP is particularly relevant, as plant location is a process which is weakly related to administrative unit boundaries. Plant location is a "point event", i.e., something we would draw as a point in a map, rather than as a polygon. Therefore, "taking points on a map and allocating them to units in a box" (Duranton and Overman, 2005), treating distance as a binary variable in/out, can be extremely misleading.

Building on these two conceptual challenges, we develop an original procedure which allows us to depart from traditional parametric approaches to industrial agglomeration, and exploits the information contained in the variation of the results given by different spatial classifications of the geographical area under scrutiny. Using data for the United States for the year 2000, we are able to rank the manufacturing industries - at a very detailed classification level - according to the absolute and relative importance of the labour market determinants in their concentration patterns. We then show that labour market accounts for one quarter of the variation in agglomeration across industries. Furthermore, we provide evidence that the EG index does not generally take the value of zero under the null hypothesis of random spatial allocation of industries, and we calculate a more precise measure of industrial agglomeration. These results are obtained using publicly available micro-geographic data, elaborated with GIS software and with two original zoning algorithms.

Second paper

The second paper assess how urbanization affects rural poverty in surrounding areas in India, using a panel dataset at district level for the period 1983-1999.

Urbanization and rural poverty are among the most important socioeconomic phenomena in developing countries. Most of the less developed countries are urbanizing rapidly, and in these years, for the first time in the history, the world urban population is overtaking the rural one (UN, 2008). At the same time, most of the world poor still live in rural areas. In 1993 rural areas accounted for 62% of the world population and for 81% of the world's poor at the \$1/day poverty line; in 2002 - after a period of intensive urbanization - the same figures stood at 58% and 76% respectively (Ravallion et al., 2007). Although urban poverty is increasing as well, rural poverty is still predominant.

The second paper of the thesis constitutes one of the first attempts - to our knowledge - to investigate the effects of urbanization on surrounding rural areas. We identify two general effects which could be in play: a first round effect due to the migration of rural poor to urban areas, and a second-round effect involving urban externalities alleviating poverty in rural areas. While the first round effect could in fact leave the overall poverty level unchanged, the second round effect may substantially reduce the number of people living in poverty. We therefore identify four main mechanisms through which a "second round effect" may take place, i.e., backward linkages, rural non-farm employment, remittances, agricultural productivity, rural land prices and consumer prices.

Subsequently, we estimate the effect of urbanization on rural poverty on a panel of Indian districts in the period 1983-1999. Conditional on the empirical challenges which the

paucity and unreliability of data sources generally pose for developing countries, we run a series of different specifications, accounting for a number of unobserved fixed effects and for a potential reverse causality bias. We eventually conclude that urbanization has indeed a strong poverty-reducing effect in nearby areas, and that the second-round effects seem to dominate.

Third paper

The third paper exploits a rich database on applications to the US Patent Trademark Office to investigate an extremely peculiar characteristic of this dataset: there is a large group of inventors who develop one or a few patents during a long period of analysis (the "comets"), while a very small group of "stars" inventors develop a huge number of patents. We especially explore the spatial component of this feature of the data, by assessing how the two categories of inventors are located across US cities, and whether the concentration of stars may be beneficial for the productivity of comets.

Although the analysis of patents' data has become extremely popular in the economic literature, this specific aspect has been overlooked, to our knowledge. Patenting has generally been treated as the output of an identical "innovation black-box", and the hypothesis that inventors situated at the tails of the distribution might be linked to different kinds of innovation processes has not been discussed. However, this latter aspect could be relevant, as the innovation literature has suggested that the production of new ideas and products is an extremely heterogenous process, with small and big firms acting in a markedly different way.

In the light of that, the paper first explores the location pattern of comets and stars, finding that comets are more likely to be found in more diverse cities, and with more small firms, than stars. Subsequently, we assess whether the activity of star inventors is beneficial to the production of comet patents in the same city and technological category. We indeed find that the effect is rather strong. A number of policy implications follow from these results.

Fourth paper

In the last two decades, Belgium - as the rest of Europe - has experienced a massive restructuring of the retail banking industry. The fourth paper of the thesis is a detailed description of the empirical work related to a wider project aimed at analyzing the effects of these changes on the geographical penetration of branches. While most of the focus in the economics literature has been on analyzing the pace and extent of liberalization at the level of a nation state or a larger geographical region, this study looks at these issues at a more disaggregated level, shifting focus to the very local scale, i.e., the most relevant to the everyday consumer of bank services.

The intense and fast process of deregulation and concentration which took place in the late '90s makes Belgium an interesting quasi-natural experiment for assessing the impact of these processes on branch geographic penetration. At the same time, the city of Antwerp also fits particularly well the analysis, as it is a city with very high income disparities, and strong pattern of neighbourhood segregation.

Local access to finance services providers is especially important for low income people. Generally, this component of the population has low within-city mobility, and, at the same time, scarce financial education and need of guidance for saving and budgeting. Very often, having a bank account is a prerequisite to access social benefits. In the light of that, we are also keen to check whether the decrease of geographic penetration of banks has been systematically more intense in low income neighbourhoods.

Therefore, using detailed neighbourhood-level data for the metropolitan area of Antwerp, changes in the patterns of bank location are examined in three subsequent periods, corresponding to distinct stages in the deregulation process. We investigate how geographical branch penetration and potential choice among different bank groups have developed over time, and how these developments relate to neighbourhood characteristics (income *in primis*).

The analysis poses two major methodological challenges. The first one is accessing the right kind of data, and the second one is properly measuring them. We deal with the first problem by constructing our own original dataset, recording the address of all retail banking branches from telephone directories. About the second one, we dedicate particular attention to developing the right empirical tools in order to provide precise evidence on the phenomena investigated. In particular, we exploit GIS software to build a neighbourhood-specific metric of bank presence which minimizes the border bias and other kinds of MAUP-related distortions. We also develop two original measures of bank "desert" and bank choice.

Our results show that in the second period, in coincidence with the strongest wave of the deregulation and concentration process, banks are systematically exiting from low

income neighbourhoods. In the two other periods, there is no evidence of such a pattern. The result is robust across a number of different measures and specifications.

Final remarks

In this introduction we discussed some of the most important and interesting issues, in our opinion, related to spatial economics. We first defined the discipline, by the means of a critical comparison with economic geography. We especially stressed the epistemological differences between the two subjects, and we reached the conclusion that spatial economics might be of interest for a wider public, because the ideological content of the adopted scientific method is somehow less strong and controversial. On the other side, we also draw attention to the fact that epistemological debates are generally absent in economics.

Subsequently, we discussed some of the methodological problems which scholars in the discipline are facing, especially related to data treatment and sourcing. In detail, we consider the issues of the Modifiable Areal Unit Problem (MAUP), of the pros and cons of data sources for spatial analysis, and identification problems. We reached the conclusion that spatial economists should exploit the peculiarities of spatial data in order to emphasize their distinctive contributions to the scientific debates.

Finally, we presented an overview of the individual papers included into the thesis.

Chapter 1

The Bright Side of MAUP: An Enquiry into the Determinants of Industrial Agglomeration in the United States

1.1 Introduction

The high degree of spatial concentration of firms belonging to the same industry is a striking real world fact. A widely accepted theoretical explanation of its determinants has been proposed more than one century ago by Alfred Marshall (1890), who identifies the labour market pooling, input sharing and knowledge spillovers as the main drivers of the process.

In this paper, we develop a new methodology aimed at measuring industrial agglomeration and at disentangling the effects of the “labour market” determinants. Starting from an exploration of the Modifiable Areal Unit Problem (henceforth MAUP), i.e., the apparently unpredictable dependence of results on the size and the shape of spatial units, we argue that this variation can rather be interpreted as useful information, once in control of the process generating the spatial classification.

More specifically, we compare the level of concentration of each industry in three differently shaped datasets: the first one uses the commuting-defined US metropolitan areas (Core Based Statistical Areas – CBSAs), and is meant to “maximize” the effect of all the determinants. The second one is composed by randomly aggregated (contiguous) counties, and is expected to be comparable to the CBSA one except for not maximizing the labour

market effects. The third one is a random aggregation of non contiguous Zip Code Areas (ZCA) and is aimed at capturing only the noise component.

Functional areas based on the self-containment of commuting flows – like the CBSAs in the US – are generally used as an approximation of the spatial extent of individual labour markets. At the same time, the effects of the determinants of agglomeration which are not dependent on the labour market (i.e., input-output linkages, market access and transport costs, natural advantages) are not affected by the spatial extent of different labour markets; the only spatial characteristics which matter, in this case, are location and distance.

As a comparison, we create an alternative spatial classification by randomly aggregating counties into internally connected spatial units (Pseudo Statistical Areas, or PSAs), ideally equivalent to the CBSA classification in everything but the self-containment of the commuting flows. We then argue that the amount of concentration of each industry measured using the CBSA definition should be bigger than that measured in PSA dataset, and that the difference is proportional to the importance of the labour market determinants for that industry.

We also contribute to the existing literature by developing a new technique for correctly estimating the amount of spatial concentration of each industry. It is widely known that the traditional concentration measures (e.g., Gini or Krugman indices) are affected by the “dartboard effect” bias⁵ (Ellison and Glaeser, 1997), i.e., the amount of spurious concentration given by the “lumpiness” of industrial establishment and the discrete classi-

⁵ The definition comes from the metaphor used by Ellison and Glaeser (1997): if an industry exhibits high concentration of plant employment, then traditional indices will find positive concentration just for a statistical effect, even if the underlying spatial process is completely random (that is, even if one randomly throws plants to a map).

fication of the space. Another, less known, source of bias for concentration indices - which affects also the "new generation" of indices, like the Ellison and Glaeser gamma - is essentially geographic and is given by the arbitrary aggregation (or disaggregation) of events in a continuous space using exogenously defined spatial units.

We propose a different approach, which consists in estimating the "noise" with a Monte Carlo procedure, and then in filtering out the industry-specific estimated noise from the estimates of industrial concentration. The vector of industry-specific values of the noise is given by the average concentration value (as measured by a gross concentration index) in a distributions of 1000 random counterfactuals, each of them obtained by i) randomly "shuffling" plants across the space and ii) randomly aggregating small (not necessarily contiguous) portions of space (US Zip Code Areas – ZCAs) into spatial units of the same size of the real ones (CBSAs). Step i) captures the "lumpiness" effect, and step ii) the geographical bias. By applying this procedure we obtain a distribution of "spurious concentration" for each industry, which we can easily use to estimate the amount of "true" concentration (and to test its significance).

Thus, we will assess the level of concentration of each industry in a way which meets all the criteria listed by Combes and Overman (2004): measures are comparable across activities and spatial scales, they take a unique known value under the null hypothesis of no systematic component in the location process, it is possible to report their significance, the spatial and industrial classification are controlled for, and the estimation technique is related to explicit assumptions about theory.

To sum up, we calculate the industrial concentration for the manufacturing 6-digit sectors in three different settings: in the first one the spatial unit is the CBSA, in the second one (the Noise) it is a random aggregation of non contiguous ZCAs, and in the third one (the counterfactual) it is a random aggregation of contiguous counties. The size distribution and the number of spatial units will be the same in the three datasets. We calculate the difference between the first two values as an index of industrial concentration, and the difference between the first and the third values as an estimate of the relative importance of labour market determinants for each industry.

1.2 The Determinants of Concentration

Industrial clustering is a striking real world fact and economists have been speculating on its determinants since Marshall's (1890) seminal taxonomy, which recognize in labour market pooling, knowledge spillovers, and input-output linkages the main drivers of industrial agglomeration. In this paper, we reclassify the aforementioned Marshallian determinants under the more general categories of "labour market determinants" and "other agglomeration determinants". Examples of labour market determinants may be the need for specific skills, for low wages, or for an environment where workers can enjoy frequent interactions with other high skilled workers, or also some local amenities which make the location particularly appealing for specific groups of workers. More generally, we define as a labour market determinant every economic factor which is related to the local labour market and

constitutes an asset (e.g. lower wage costs, or higher productivity) for a firm located in that area.⁶

Conversely, examples of non labour market determinants are input-output linkages between firms, access to local natural resources, lower prices for inputs other than labour.

Labour market determinants may have a non-linear, and even non-monotonic, relationship with the variables that are commonly used to proxy the input intensity of industries (average wage, total labour compensation over total value of shipment, capital intensity, etc.). For instance, a given industry may target a local endowment of low-wage labour force, another a highly skilled one. Both the industries may consider labour supply as the most important determinant of their location, but they completely differ in the average wage level. In addition, the hi-skill industry may be capital intensive, which implies that it allocates a low share of total costs to labour compensation, while the low-skill one is likely to employ a large labour force, thus bearing a bigger share of labour cost. Alternatively, capital intensive production may require also repetitive and unqualified labour, which translates into a low average wage. In these and similar situations, results from a cross-industry linear regression of a concentration index on measures of input intensity can be misleading.⁷

This can be just one of the several reasons why empirics of concentration have not been conclusive so far. Contributions on the subject can be separated into two general categories: the description (or measurement) of the concentration pattern, and the inference

⁶ The reader should note that the term “determinants” is not interchangeable with “externalities”, the former being more general than the latter. This is important, as some of the determinants we are interested in, e.g. the availability of specialised workers, are not externalities.

⁷ One could theoretically interact industry input intensity with local factor endowments (as done, for instance, by Midelfart-Knarvik et al., 2000), but obtaining the necessary geographic and industrial data is impossible in most of the cases; moreover, one would still need to assume the process is monotonic in the interaction variable.

on its determinants at industry level. It is obvious that if the former is misleading, also the results of the latter are unreliable.

The first issue has been recently critically surveyed by Combes and Overman (2004), who effectively point out all the limits of “attempts to collapse the entire structure of industrial production down to one number that can be compared across time and across countries” (p. 2855). These limits are particularly evident in the so-called first generation concentration (and specialization) indices – namely the Krugman index and Location Gini index – in the light of the failure to control for the aforementioned “dartboard” bias, and more generally to meet the target requirements identified by Combes and Overman.

The second generation indices, i.e., the Ellison and Glaeser’s “gamma” index (EG henceforth) and similar (Ellison and Glaeser, 1997; Maurel and Sedillot, 1999), represent a significant improvement,⁸ but are still fraught with problems. The EG index for industry k is equal to:

$$\gamma = \frac{G_k - (1 - \sum_i x_i^2) H_k}{(1 - \sum_i x_i^2)(1 - H_k)} \quad (1.1)$$

Where G is defined as

$$G_k = \sum (s_k^i - x_i)^2 \quad (1.2)$$

where s_k^i and x_i correspond to the share of total employment of region i for industry k and in the aggregate, respectively, and H_k is the plant employment Herfindahl index,

⁸ Actually this assertion is questionable: although the second generation indices are more theoretical informed, on the other side in a few cases a gross employment index is more policy relevant. e.g., in assessing how an industry shock affects regions. In such a case, the only thing that matters is the concentration of employment, irrespectively of the dartboard bias.

corresponding to the sum of the squares of the share of employment of each plant, over the total employment of the industry. The EG index has the nice property of controlling for the size distribution of plants within industries, and of regions within the study area (where the size is quantified with employment). The authors demonstrate that their index takes the value of zero under the null hypothesis of random location conditional on the aggregate manufacturing employment in that region. Formally, the index derives from a model where firms choose their location according to natural advantages (first order spatial process), and intra-industry spillovers (second order spatial process). The processes are observationally equivalent, as both translate into a local industry employment share higher than the aggregate one.

The EG index generated a *Pax Romana* in the field, as a consequence of its strong intuitive appeal and easy computation. However, there are a few aspects which may still be improved. First, the Herfindahl index takes into account only the numerosity, the average size, and the variance of the size distribution of plants and regions;⁹ under the underlying statistical and theoretical hypotheses, this is sufficient to prove the unbiasedness of the index. However, more flexible approaches have proved to give rather different results (e.g. Duranton and Overman, 2005). our methodology will therefore exploit all available information on the two size distributions without need to rely on any statistical assumption. Second, equating the probability of a plant to locate in a given region to the region's aggregate share of employment may not be the most logical null hypothesis (especially if

⁹ As it is widely known, the Herfindahl index can be expressed as $1/n + n\sigma^2$, where σ^2 is the variance of the employment shares of plants. It therefore depends on the number of plants and regions, their average size, and the variance of their size.

there are many small regions and few plants for industry), as the size of plants may be endogenously determined by the industry pattern of concentration (as shown by Holmes and Stevens, 2002). We will follow a different approach, based on the number of manufacturing plant sites, rather than on the employment share (which is the same “null hypothesis” adopted by Duranton and Overman, 2005). Third, the variance of the employment size of regions is not the only geographical characteristic which contributes to generate the bias. As Arbia (2001), Overman and Combes (2004), and Duranton and Overman (2005) convincingly argued, it is the whole process of “taking points on a map and allocating them to units in a box” (Duranton and Overman, 2005) that is arbitrary and likely to introduce a spurious component in the results. This happens because our “boxes” are generally not regular nor homogeneous in both shape and size. Furthermore, in the process we lose all the spatial information embedded in the data, and distance is collapsed to a binary variable in/out.

Regarding empirical inference on the *mechanisms* that could lead to agglomeration, only a few contributions provide evidence on the Marshallian microfoundations of agglomeration economies at industry level (Ellison and Glaeser, 2001; Rosenthal and Strange, 2001 and 2004). These studies are based on a linear regression of the EG index on industry-specific input intensity proxies. Results are not conclusive, however, for many possible reasons. The first one is data scarcity, both at geography and industry level: detailed data on manufacturing activity disaggregated by sector are hard to find, also at national level. It follows that the concentration pattern and the input intensity of industries are extremely difficult to quantify. Generally scholars use shares of total expenditure as proxies but these

are clearly endogenous, as firms chose locations (and therefore concentrate) in order to minimize costs. This has been acknowledged (e.g. Rosenthal and Strange, 2001) but not satisfactorily solved, to the best of our knowledge. Additionally, as already mentioned before, if the EG index contains a bias, this is transferred to the regression output. Second, the effect of the determinants may be non linear and, more generally, difficult to parametrize. Third, path dependencies, local idiosyncratic dynamics, and unobserved factors may play important roles in explaining industrial concentration. All these elements provide the need for developing an alternative tool to explore the topic.

The need for further research in the field has also been stressed, in a very recent paper, by Ellison et al (forthcoming). This paper proposes a novel approach, based on the analysis of coagglomeration patterns (as measured by the EG co-agglomeration index), rather than agglomeration, and expressing them as a function of the degree by which different industries share the same inputs. They apply their methodology to US data and cope with endogeneity issues by using UK industry variables as instruments. Their results suggest that input-output linkages and, secondly, labour market pooling are the most important factors in explaining coagglomeration.

1.3 The MAUP and the Gerrymandering Approach

1.3.1 The Dark Side

Every geographical area may be divided in a theoretically infinite number of ways, and economic estimates may present huge variation among them. Furthermore, differently from

international comparison (where country borders have an economic and political meaning that is not comparable with any other geographic classification) in a sub-national setting researchers face a variety of administrative and functional divisions – each of them with its pros and cons – with the result that the choice of the spatial unit is often arbitrary, or constrained by data availability. The complex variation which the results of investigations based on “modifiable units” are prone to is called the “Modifiable Areal Unit Problem” (MAUP). The issue had first been raised by Gehlke and Biehl (1934), who essentially focussed on the scale problem. Openshaw and Taylor (1979) provided evidence of how the “shape” component of the MAUP plays an important role too, and numerous other contributions in the field of quantitative geography have explored in detail the nature of the bias generated by the MAUP in a number of different statistical and econometric specifications (e.g. Arbia, 1989; Fotheringham and Wong, 1991; Amrhein, 1994; Steel and Holt, 1995). More recently, Briant et al. (2007) reconsider the role of MAUP with an application to French data. They perform standard economic geography analyses (applied to agglomeration, concentration, and trade), using administrative, functional, and random (geometric) spatial units. Although they find some variations in the results, they eventually reach the conclusion that “the MAUP induces much smaller distortions than economic misspecification” (p. 25).¹⁰

Arbia (1989) shows how the distortions arising from scale and shape effects would be minimized if the units of analysis were: i) identical, in terms of shape, size and neigh-

¹⁰ Two caveats, however, have to be kept in mind while assessing their findings: first, the random counterfactual is based on a single iteration, thus is not possible to test the statistical significance of their results; second, the French political geography may present some peculiarities which limit the extendibility of their conclusions, as the authors themselves acknowledge.

bouring structure; and ii) spatially independent. Given the difficulty to meet both the prerequisites with available geographical data, in the last years the MAUP has generally become part of the subconscious of spatial economists and regional scientists, and has seldom been taken into explicit consideration. In the rare case it happened, efforts to deal with it have been concentrated on obtaining a dataset of spatial units which are “geographically meaningful” in relation to the enquired phenomenon,¹¹ or in getting rid of the spatial unit altogether by using a continuous definition of space (e.g., Duranton and Overman, 2005). Scholars in quantitative geography have been more active and some of them tried to develop “optimal zoning” of areal units in order to minimize the MAUP (e.g. Openshaw and Rao, 1995), or to assess how aggregation effects are affected by different spatial classifications (Amrhein, 1994; Wong, 1997).

1.3.2 The Bright Side

The idea that the MAUP, once under control of the researcher, may become a powerful tool has already been suggested by Openshaw (1977), but to the best of our knowledge no contribution has tried yet to test economic theories by comparing spatial systems with different economic properties (other than the size of spatial units). The attempt to develop such a methodology, with an application to industrial spatial concentration, is the main goal of this paper.

More precisely, we develop a “gerrymandering” approach to i) correctly measure the extent to which different industries agglomerate, and to ii) disentangle the effect of labour

¹¹ See Cheshire and Hay (1989) and Magrini (1999 and 2004) for some examples.

market determinants from the effects of the other determinants of agglomeration. In the following, we discuss the methodology related to ii), while the first point will be treated in section 4.1.

In order to disentangle the effect of labour market determinants, we compare the level of concentration of each industry in the US “travel-to-work” regions (CBSAs – Core Based Statistical Areas) against the level of concentration of the same industries in a distribution of spatial units of the same average size (PSAs – Pseudo Statistical Areas), obtained by randomly aggregating the same sample of counties which form the CBSAs. Before illustrating the details, we introduce a brief definition of the CBSAs and PSAs.

The CBSAs: the US Office of Management and Budget defines the CBSAs by identifying a central county with a significant share of urban population and by subsequently aggregating the neighbouring counties which have high commuting linkages with the central one. CBSAs are merely statistical entities, do not have any political or administrative meaning, and may cross a State border. Essentially, the aim of the CBSA definition is containing in the same spatial unit the place of work and of residence of most of the workers.¹² Whether this can effectively identify the “borders” of local labour markets is debated, also in the light of the slippery definition of the latter: the labour market may be defined as a continuum not only in the spatial, but in almost any of its dimensions (Cheshire, 1979). Furthermore, workers may have different commuting patterns according to their income and skills. Nevertheless, some consensus has emerged on the comparative advantages of a zoning procedure which maximizes the self-containment of commuting flows (see Cheshire

¹² More information on the CBSA definition can be found in the US Census website (<http://www.census.gov/population/www/c>) and in the part IX of the Office and Management and Budget Federal Register of 27/12/2000 (OMB, 2000).

and Hay, 1989, p. 21-25, for a detailed discussion),¹³ and the criterion is now adopted by many statistical institutes around the world (e.g. in the UK, France, Italy).

The PSAs: the Pseudo Statistical Areas, or PSAs, are internally connected spatial units, created by randomly aggregating counties in order to build a counterfactual for the CBSAs. Their number and size distribution is identical to the CBSAs and, ideally, the only difference from the latter is in the containment of the commuting flows, as in the PSAs the self-containment is not maximized (as the aggregation of counties is random). More details on how they are built and their properties will follow - at this stage, we just need to focus on the absence of any spatial linkage to local labour markets.

We can now introduce the following two assumptions:

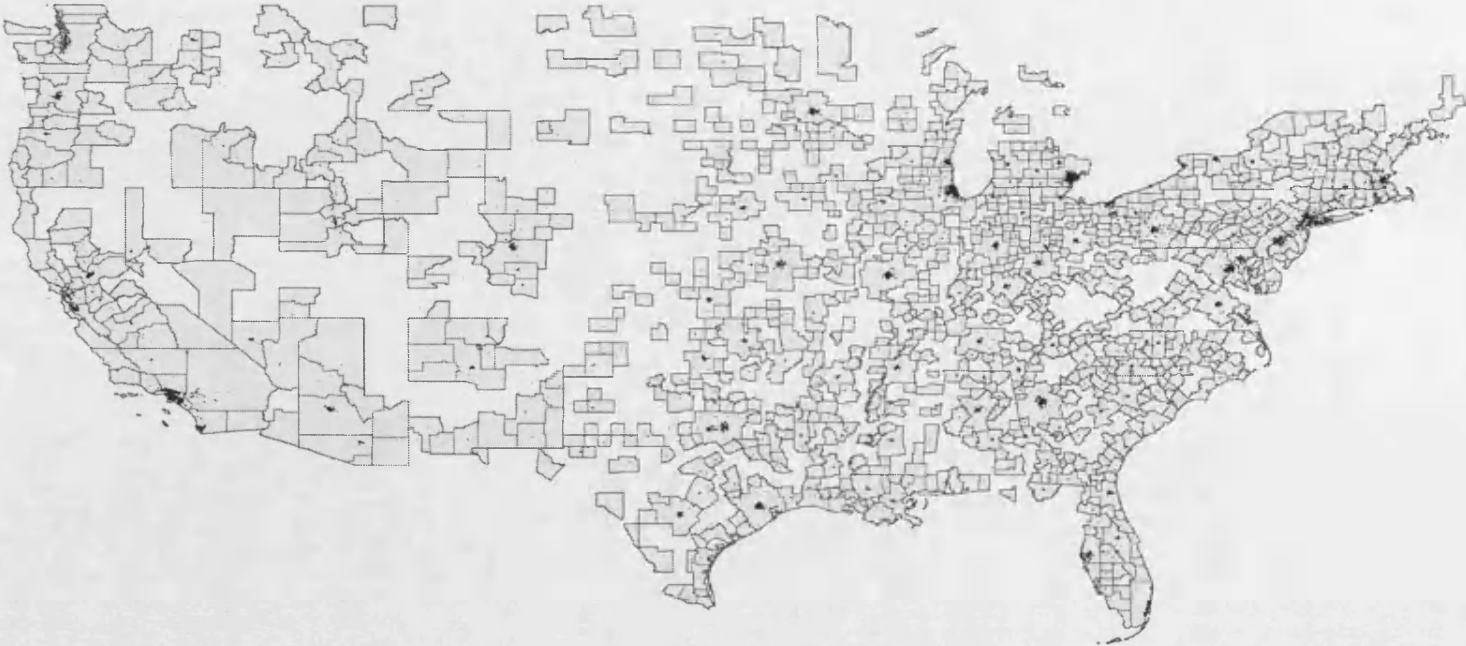
1. The intensity of all the determinants of concentration, except the labour market ones, varies continuously across the physical space.¹⁴ The reader should note that this is implicitly assumed – for all the determinants – in virtually all the previous empirical analyses of concentration/agglomeration, either because they do not discuss the properties of the adopted spatial unit, or because they adopt a distance-based measures of concentration.

2. A representative (information constrained) firm located within a given CBSA expects to face a labour supply which is spatially constrained by the extent of the estimated commuting area, i.e, by the borders of the CBSA. This is a weaker assumption than claim-

¹³ The rationale for that rests on the consideration that the most immediate channel of adjustment and price-clearing within a spatial labour market is occupational mobility constrained to residential immobility (people change the job but not the house). In the following of the paper I will present an empirical exercise which corroborates the “labour market homogeneity” hypothesis.

¹⁴ Considering that the sample is limited to the counties belonging to the CBSAs, i.e., to counties where a significant level of population or employment is present, the notion of distance we use is corrected for the general spatial distribution of economic activity.

Fig. 1.1. CBSAs and populated places



Note: the picture shows the CBSAs (light gray polygons) and the populated areas (dark dots) in the US

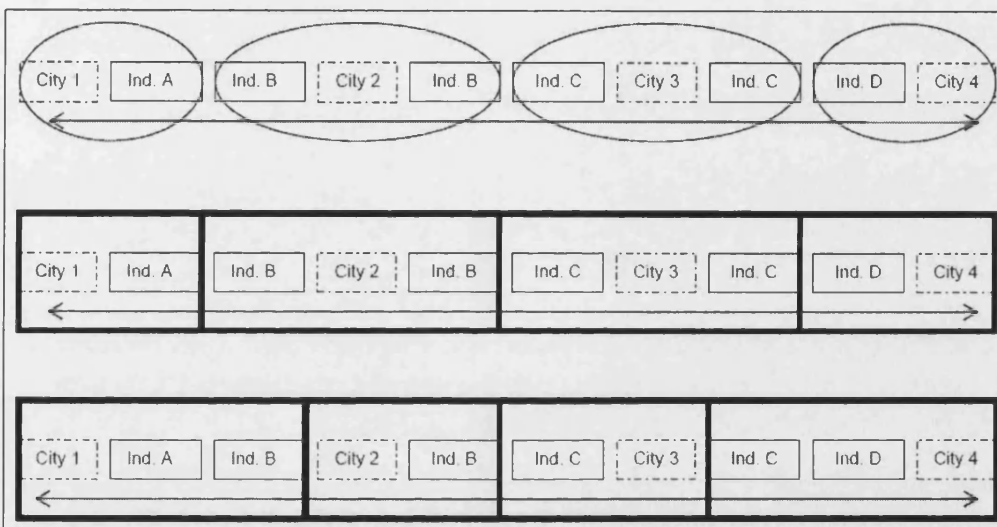
ing that CBSAs truly identify local labour markets, which is the property we discussed earlier.

It follows that if an industry is highly dependent on labour market characteristics, its heterogeneous distribution across space will follow the heterogeneous distribution of the labour endowment. Considering that higher spatial concentration implies higher heterogeneity among spatial units, the amount of concentration determined by labour market externalities is expected to be higher in a dataset in which the spatial units are defined in a way that maximizes the “within homogeneity” and the “between heterogeneity” of labour market characteristics (CBSAs), than in any other comparable dataset (PSAs). Conversely, if the spatial distribution of a given industry is unaffected by labour market externalities, the geography of its concentration patterns is independent on commuting flows – according to assumption one, it depends only on physical distance.

In order to clarify the concept, we introduce a simple example (figure 1.2). Consider a one-dimensional space where there are four cities (1, 2, 3, and 4) and six industrial districts, belonging to four different industries (A, B, C, and D). Workers commute from cities to the nearest industrial district, forming the commuting area delimited by the ellipsoids in the upper diagram of figure 1.2. Labour is the only input and the location of the different industries is only due to labour market determinants. A commuting-based classification (like the CBSAs) will subdivide the space into the four spatial units reported as rectangular polygons in the second line of the diagram, thus minimizing the commuting flows across them. In the third line, a random classification is drawn; as compared to the second line, units have the same size, but the commuting flows are not taken into account. It

immediately appears from the example that the amount of concentration we can measure using the commuting-based spatial classification is bigger than what we would find using any other spatial classification.¹⁵

Fig. 1.2. A stylized example



Note: the figure shows a stylized example of different spatial classifications. The first line reports the commuting flows from city 1,2,3, and 4 to industries A, B, C, and D. The second line reports a functional spatial classification based on commuting flows, while the third line presents an alternative spatial classification which the same size distribution of the previous one.

The reader may argue that the methodology is affected by a reverse causality problem, that is, the commuting flows are determined by industrial clustering and the labour market areas are shaped after the industrial clusters, rather than being their determinant. We think that this may hardly be a concern, for the following reasons:

¹⁵ The methodology may recall the so-called “regression discontinuity approach”, which has recently been applied in a geographical setting by Holmes (1998) and Duranton et al. (2006), among others. However the apparent analogy is misleading, because the discontinuity I exploit in this case is only approximate, given that we expect that some commuters will cross CBSAs borders.

a. We are considering only the manufacturing sector, which employs less than the 20% of the workforce, while the commuting flows are calculated on the whole sample of workers.

b. Generally cities and commuting linkages are shaped by exogenous or path-dependent factors, like physical geography or long term investments in commuting infrastructures.

c. Different industries use different natural resources and are affected by different determinants. Thus the reverse causality would be an issue only in extremely specialized towns, where only one or few industries employ most of the labour force. Only in this circumstance commuting flows might be “endogenous” to non-labour determinants.

d. The industry statistics are calculated over a sample of 876 CBSAs. Even if the aforementioned arguments are not convincing relatively to a few specific case, it is extremely unlikely that they are systematically violated across the whole sample.

However, even if the previous points do not hold, which in turn would question the true direction of causality driving our results, the methodology proposed here may still keep most of his informative content by showing the degree of association between industrial clusters and individual agglomerations of labour. That is to say, industries supposedly clustering because of the effects of other determinants end up generating integrated labour markets which spatially match the extent of the clusters. Even if we do not know what came first (and it is hard to define theoretically as well, given the “snowball” effect of agglomeration economies), it would still be a step forward in our understanding of the economics of agglomeration.

Summing up, in this paper we therefore investigate, industry by industry, the degree of spatial association of labour markets and the agglomeration of plants, across a sample of 876 spatial units and more than 320,000 manufacturing plants. Under some additional assumptions, we interpret this association as an evidence of labour market determinants of agglomeration.

1.3.3 A Real World Example

Figure 3 reports a closer view of the CBSA map, in order to discuss the adopted methodology with a real world example. The thick lines report CBSA borders, the thin lines the County borders, and the black polygons the “populated places”. The digital maps are available from the US National Atlas website and refer to the period of analysis. The general pattern that emerges is a collection of small one-county micropolitan areas which surround a few big metropolitan areas, composed of several counties. In both the micropolitan and metropolitan areas there is a populated agglomeration, normally a medium or big city, at the centre of spatial units, and smaller towns around.

Looking at the picture, it should clearly appear how – in a world where the spatial extent of different labour markets does not matter for agglomeration – there is no reason to expect that the CBSA classification would better match the clustering of economic activities than any other random aggregation of counties. For instance, if in the North and South parts of the Indianapolis CBSA (at the top-left of the picture) we register the presence of plants of the same industry and this industry is not present in Anderson, Columbus and Bloomington (neighbouring CBSAs on the southern side), this is because of a specific need

of that industry for the labour force residing in Indianapolis. Conversely, if plant location is driven by input-output linkages with other firms in Indianapolis city, these plants can be located in Columbus or in Bloomington with roughly the same probability of locating in a site at a similar distance inside the Indianapolis CBSA.

1.4 Building the Counterfactuals

1.4.1 The Noise

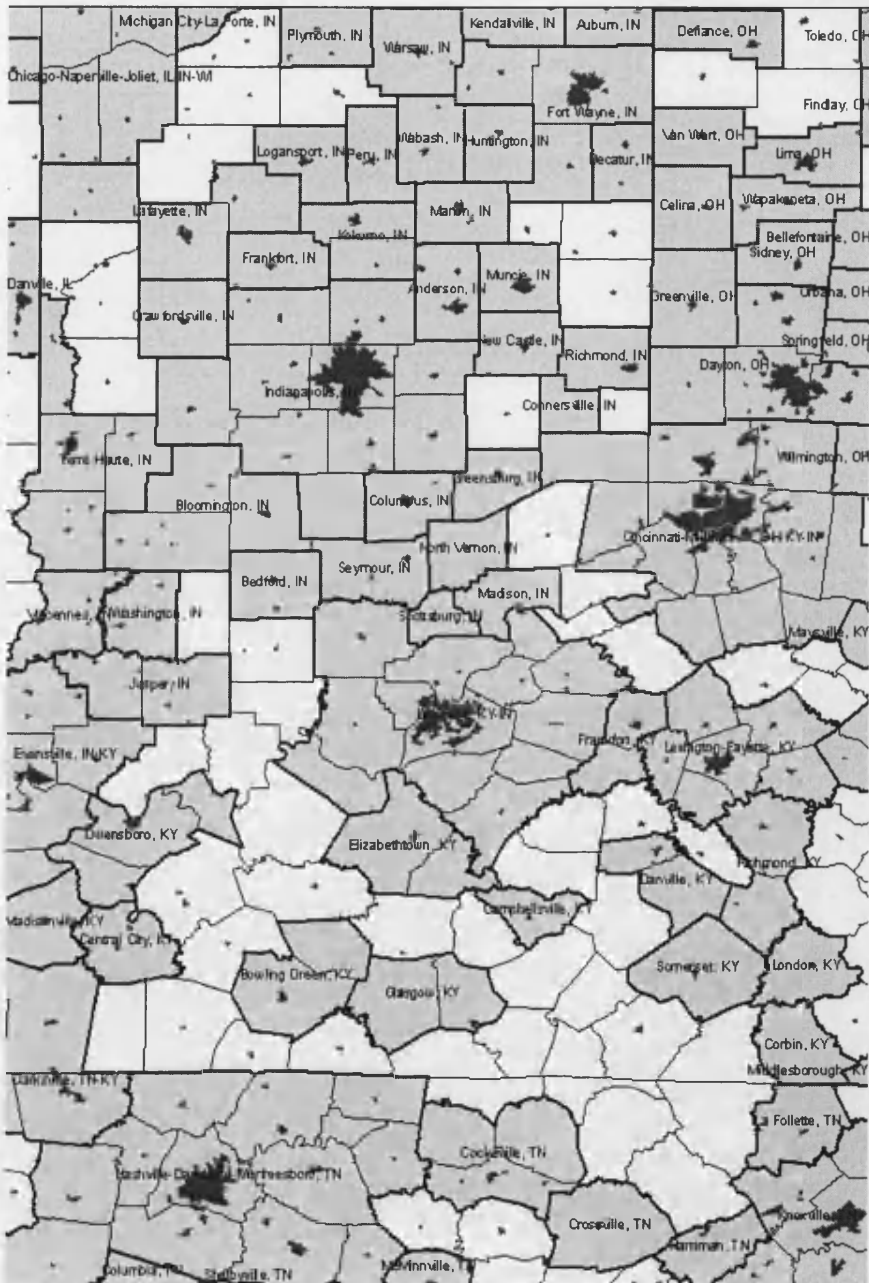
As we mentioned earlier in the paper, the amount of concentration detected using a gross concentration index may be seriously biased because of the “dartboard effect” (due to the “lumpiness” of industrial establishments) and the MAUP.

Trying to eliminate the bias without renouncing to discrete spatial units may be extremely complex and beyond the scope of this paper. It is relatively easy, instead, to create a counterfactual where the amount of concentration measured by a gross concentration index is totally spurious, i.e., is given only by noise, in order to have an estimate of the bias generated by the given industry and spatial classifications (and their interaction).

Therefore, we apply a simple technique which exploits all the information contained in the plant employment distribution and in the spatial classification system. our approach consists in composing a distribution of 1000 datasets, each of them created by applying the following two-step algorithm to the original sample of plants under analysis:

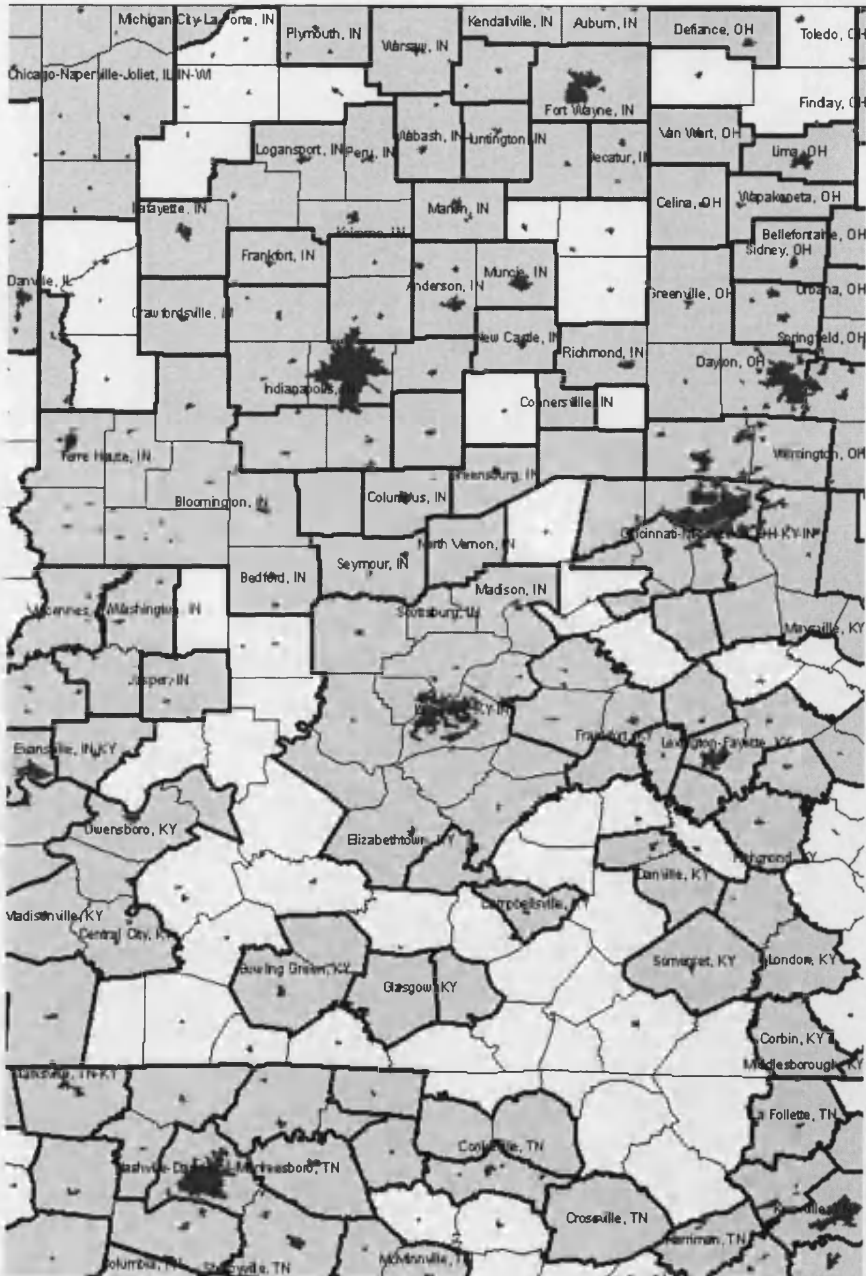
- a) The plants are “shuffled” across ZCAs (Zip Code Areas, the smallest spatial units at which industry data are available), simulating a scenario where plant location is

Fig. 1.3. CBSA classification, Indianapolis area



Note: the picture reports the CBSA borders in the area around the city of Indianapolis (IN). The bold lines are CBSA border, the thin lines are county borders, and the white areas are counties which are not included in the CBSA classification. The darker polygons represents populated areas. Source: author's elaboration on US National Atlas shape files.

Fig. 1.4. PSA classification (single iteration), Indianapolis area



Note: the picture reports a single iteration of the PSA borders in the area around the city of Indianapolis (IN). The bold lines are PSA border, the thin lines are county borders, and the white areas are counties which are not included in the CBSA classification. The darker polygons represent populated areas. Source: author's elaboration on US National Atlas shape files.

random given the spatial distribution pattern of all manufacturing plants. This means that every ZCA ends up with a random – in term of employment and industry – sample of plants, but with exactly the same number of plants it originally had. The reader should note that this is equivalent to reshuffling plants across plant sites, in the spirit of Duranton and Overman (2005).

b) The ZCAs are randomly aggregated – without any contiguity constraint – into bigger spatial units. The number of these spatial units is equivalent to the number of CBSAs, and every time a new spatial unit is created, its maximum employment size is drawn (without replacement) from the actual CBSA size distribution, in order to mimic the total employment distribution of the CBSA dataset.¹⁶ This step is meant to absorb the geographical bias embedded in the CBSA dataset, by reproducing an equivalent spurious aggregation of points into comparable “boxes” deprived of any spatial meaning (internal connectivity). Replicating the CBSA size distribution is important, as gross concentration indices are highly sensitive on the size distribution of the chosen spatial unit (generally, concentration is positively correlated with the variance of the size).

At every iteration,¹⁷ the G concentration index and the EG index are calculated and stored. At the end of the process, we obtain a distribution of values of the indices for each industry. Interestingly, for many industries the average G values are definitely high,

¹⁶ As the size limit cannot be met exactly and cannot be overcome, the units are slightly smaller than their equivalent CBSAs, and a few ZCAs are left over from the aggregation process. All these are then randomly assigned to the existing units, with probability proportional to the size of the units, in order to preserve the original size distribution.

¹⁷ The assignment of plants to PSAs is a slow procedure – it takes around nine minutes with a standard PC. Repeating it 1000 times will take around 150 hours. Therefore, in order to speed up the algorithm, the step b) is repeated only every 50 iteration. However, the random variation of the results at every iteration is assured by the reshuffling of the plants in step a).

while the values of the EG index are generally close to zero, but with significant exceptions, especially for industries with a small number of establishments. A more precise description of the “Noise pattern” is reported in the results section.

1.4.2 The Pseudo Statistical Areas (PSAs)

In order to obtain a relative estimate of the importance of the labour market determinants for each industry, we need a counterfactual in which the spatial units are in all comparable to the CBSAs, except for the containment of the commuting flows. This implies two main requirements: i) the “Pseudo Statistical Areas” (PSAs) must be internally connected, and ii) they must follow the same size distribution of the CBSAs (in terms of total employment and area).

Therefore, we composed an algorithm which randomly assigns the 1707 counties to 876 internally connected spatial units (i.e., every county composing the PSA must be contiguous to at least one of the counties composing the same PSA). In theory, it would be possible to aggregate ZCAs – like in the Noise – rather than counties. The latter, however, are preferred for three reasons: first, counties are the “building blocks” for CBSAs; second, differently from CBSAs, counties are also administrative, not just statistical, entities, therefore their borders may account for unobserved discontinuities (e.g. policies, taxation, etc); third, building a contiguity matrix of ZCAs is not easy, as their borders are not exactly defined, and their huge number would create computational difficulties (the contiguity matrix would have around 1,6 billion elements). This, of course, does not imply that the county

is the best spatial unit to measure industrial concentration in the US; rather, it is the best building block to define a counterfactual to the CBSAs definition.

The functioning of the algorithm – explained in depth in appendix A – can be summarized here as follows: for every county that has not been assigned already, a random neighbour is chosen, and both the counties are added in a PSA-to-be. A vector including all the neighbours of the two counties is then created, from which a random contiguous county is chosen and included to the PSA-to-be. The process continues until the size and employment limit is reached, or all the counties around the PSA-to-be are assigned. In order to maximize the degree of internal connectivity, and therefore to avoid shaping PSAs as long rows of counties, the likelihood for a county to be added to a PSA is exponentially proportional to the number of contiguous counties already included in that PSA. For instance, if an unassigned county i is surrounded by counties which have already been assigned to the forming PSA, its probability to be assigned is much higher than it is for a county that has only one contiguous neighbour already assigned to the forming PSA.¹⁸

The algorithm also closely mimicks the employment size distribution of CBSAs. This is obtained by imposing on every forming PSA a total employment limit drawn (without replacement) from the actual CBSA size distribution.¹⁹

Although there is a clear trade-off in replicating the size distribution without limiting the randomness of the aggregation, on average the moments of the PSA distribution are

¹⁸ The neighbour vector is sorted in decreasing order according to the number of times every county is repeated; then a random number is drawn from an exponential distribution with the lambda parameter equal to three (thus skewed to the left) and bounded by the length of the vector of counties. The random number correspond to the position of the chosen county in the vector; lower is the number, more frequently the county is repeated in the vector. This implies that counties which are repeated most are more likely to be selected.

¹⁹ Preliminary versions of the algorithm showed that results of the G and EG indices are not independent on the size distribution of the spatial units

close to the ones of the CBSA distribution (Table 1.1, first two rows). While the focus is primarily on the employment distribution, the algorithm contains also some instructions aimed at replicating the CBSA area distribution (Table 1.1, third and fourth rows). The joint replication of both the distributions (area and employment) is important for two reasons: first, it avoids that the difference in industrial concentration between the CBSA and PSA dataset contains a spurious component due to a different size distribution; second, it also contributes keeping the distributions of central and outlying counties across spatial units similar in the two datasets. In fact, as central counties are much denser populated than the outlying ones, a different repartition of them would necessarily end up in a different employment or area distribution.

The outcome of a single iteration of the algorithm are shown in Figure 4, which reports the same area of Figure 2, substituting the CBSA borders with the PSA ones. The picture shows how the size and the shape of the spatial units are extremely similar to the original CBSA classification. Furthermore, there are two other considerations which corroborate the robustness of the results to potential isolated “strange geometries”: first, they would generate a bias only for industries with a significant share of plants located in that area; second, and most important, the algorithm executes 1000 iterations, which implies that distortions may arise only if the “weird geometries” are always created in the same location. At every iteration, a gross concentration index (the G index, defined in the following section) is calculated and stored.

As we mentioned earlier in the paper, we assume that the PSA counterfactual will be less heterogeneous in terms of labour market characteristics than the CBSA dataset. We

did a simple exercise to test this assumption: we calculated the coefficient of variation of an immediate proxy of labour market characteristics – the unemployment rate – in the CBSA and PSA dataset. If our assumption is true, we should find that the CBSA dataset presents an higher variation in the unemployment rate. Results are indeed supportive: in the CBSA dataset the coefficient of variation is equal to 0.377. In 1,000 iterations of the PSA dataset we obtain a mean of 0.331, a 90th percentile of 0.348, and a maximum of 0.366. It implies that the variation of the unemployment rate is higher in the CBSA dataset than in any of the 1,000 counterfactuals. This therefore support the hypothesis that CBSAs are significantly more heterogeneous in labour market characteristics.

Table 1.1. CBSA and PSA (average across 1000 its.) distributions: moments and percentiles

	Employment		Area	
	CBSA	PSA	CBSA	PSA
St. dev.	421818	400,220	0.61	0.87
Kurtosis	122.6	135	33.55	31.57
Skewness	9.5	10	4.46	4.77
Max	6,996,312	6,648,427	6.98	8.76
Min	1,886	1,840	0.011	0.001
P25	13,052	10,726	0.16	0.129
P50	25,811	24,908	0.255	0.197
P75	60,687	74,837	0.525	0.403

1.5 Results

For every industry, our “raw” results consist of the G value, calculated following the specification reported in equation 1.2, for three different spatial classifications: the Core Based Stastical Areas (CBSAs, a single value for each industry), the Pseudo Statistical Areas

(PSAs, random aggregation of contiguous counties, distribution of 1000 values for each industry), and the “Noise” (random aggregation of non contiguous Zip Code Areas, distribution of 1000 values for each industry). Calculations are based on a dataset of 6-digit plant employment at county/ZCA level for the year 2000. More details on data are reported in Appendix C.

1.5.1 Interpretation

CBSA and CBSA minus Noise: the value of the G index measured with the CBSA classification is supposed to capture the maximum effect of all the determinants. The value of the G in the Noise counterfactual quantifies the industry-specific bias that affects the index. The difference between the two values is supposed to be a correct estimate of the degree of spatial agglomeration of industries, significantly improving on other existing methodologies. More specifically, our measure is more precise than standard concentration indices (e.g. Gini or Krugman index) as it controls for the aforementioned "dartboard" and discretisation biases. Similarly, it improves on the Ellison-Glaeser index as the latter controls only for the first bias, and not for the second. To the extent that our results can be matched to other dataset available at the same (discrete) spatial classification, they are also complementary to analyses of industrial agglomeration based on a point pattern analysis (PPA) approach (e.g. Duranton and Overman, 2005).

CBSA minus PSA: The CBSA-PSA difference is uninformative in its absolute value, as we cannot know how much of the labour market effect is absorbed by the PSAs defini-

tion.²⁰ Therefore, two assumptions about this unknown value are required: i) it is a share (strictly smaller than one) of the total (CBSA) effect, and ii) this share is constant across industries. The first assumption has already been discussed and is at the basis of the methodology. The second is weaker than it may appear: the share of the total effect absorbed by the PSAs definition depends on how much the PSAs are *geographically* similar to the CBSAs, over the whole US territory. This unknown “degree of dissimilarity” in turn depends on the spatial matching between the two definitions - the CBSAs and the PSAs - and it is a merely geographical factor. Given that plants of different industries are located in the same counties, it affects different industries in the same way. In the rarer cases in which a few industries account for most of the employment in a small number of areas, the reader should keep in mind that the value we use is the average across 1000 iterations of the PSA algorithm; therefore, even if the “spatial mismatch” between the CBSA and PSA datasets may add an industry-specific disturbance to the value of industry concentration, still the number of iterations of the PSA algorithm should be large enough to make the industry-specific disturbance converge to an “asymptotic” null value. However, a caveat is needed for small industries located in only few counties: for them, results relative to them should be interpreted with caution (number of plants in the industry and number of counties where at least one plant of the industry is present are always reported in the tables of results).

In order to better illustrate the meaning of the CBSA-PSA value, we introduce a simple formalization. Define the total concentration (as measured by a concentration index) of

²⁰ As it is explained earlier in the paper, the PSA definition is *not* assumed to maximize the effect of labour market determinants. However, although not maximized, the effect can still be remarkable, given that the counties are contiguous.

industry k in the CBSAs dataset as X_k , and assume that this is explained by two components - the components of concentration due to labour and non labour market determinants, respectively - plus an idiosyncratic factor. Formally, the relation is the following:

$$X_k = a_k + b_k + \varepsilon_k \quad (1.3)$$

where a_k and b_k are the components of concentration due to labour and non labour market determinants, respectively, for industry k ; and ε_k is an idiosyncratic local factor, independent from a_k or b_k . Symmetrically, define then the total concentration of industry k in the PSAs dataset as

$$Z_k = c_k + d_k + \varepsilon_k + \nu_k \quad (1.4)$$

where c_k and d_k are the components of concentration due to labour and non labour market determinants, respectively, for industry k ; and ν_k is a disturbance due to the imprecision in the PSA algorithm. Following assumption 1 and 2 in section 3.2,

$$c_k = \sigma a_k \quad (1.5)$$

$$\sigma \leq 1 \quad (1.6)$$

$$b_k = d_k \quad (1.7)$$

where σ is the unknown, but constant across industries, share of labour market effect captured by the PSA definition. The two latter equalities therefore imply that i) the effect of labour market determinants in the PSA dataset is always smaller or equal than in the CBSA

one; and ii) the effect of the non labour determinants is the same in the two definitions. Furthermore, v_k converges to zero in a large number of iterations of the PSA algorithm:

$$E[v_k] = 0 \quad (1.8)$$

It follows that the difference between the concentration in the CBSA-PSA dataset, industry-wise, is equal to:

$$X_k - Z_k = (1 - \sigma) a_k \quad (1.9)$$

which is directly proportional to a , the value we are interested in.

1.5.2 Testing the significance

The CBSA value is assumed to be known exactly and without measurement error. This is a strong assumption; a measurement error may arise from an imprecise estimation of commuting flows, or from a suboptimal solution of the CBSA algorithm. However, in the first case the error can hardly be large enough to affect the allocation of counties to CBSA. The second case is more relevant, but – if something – will lead to a lower level of the self-containment of commuting flows (i.e., to a level which is not maximized). Summing up, we expect the CBSA measurement error to be small and pushing the G value downward, which in turn will make our tests more conservative. In the light of that, and considering also the complexity of taking it into account, we decide to ignore it.

On the contrary, the variance around the PSA values generated by individual iterations is expected to be much higher, as i) every random iteration presents a large number

of different solutions, and ii) a few iterations may produce weird geometries. However, to the extent that the measurement error is a zero-mean random component, its effect is cancelled by using the average value across 1000 iterations. Therefore, the exact value of the CBSA-PSA difference is given by the difference between the actual CBSA value and the average PSA value.

Finally, to obtain the p-value of the CBSA-PSA difference we calculate, industry-wise, the share s of the PSA distribution which is smaller than the unique CBSA value, and define the p-value as $1-s$. E.g., if for industry k 950 iterations (out of 1,000) of the PSA algorithm give a G index smaller than the CBSA one, then the p-value for industry k is equal to 0.05.

1.5.3 Discussion of findings

In this section, we discuss the results obtained adopting the different spatial classifications, and the differences among them.

a) Noise: the values of the estimated “Noise” provide extremely useful information for assessing the bias of the concentration index (Table 1.2). As we discussed earlier, if the G index were robust to the bias originating from the lumpiness of plants and the discrete classification of space, the average value for all industries would be zero, as plants are randomly located. Instead, a first, striking result is that the noise is extremely “loud”: the average value of the G across the 473 manufacturing industries is 0.032, which in previous studies based on the value of that index would have been interpreted as a remarkable signal

of concentration (with the caveat, however, that the small dimension of our spatial units and the highly detailed industrial classification partly contributes to generate big values).

b) CBSAs minus Noise: the value represents our "unbiased" estimate of industrial concentration. Results (Table 1.3) show that 298 out of 473 manufacturing industries present a positive CBSA-Noise difference at 5% significance level (215 at 1% level). On the contrary, only three industries exhibit a 5% significant negative value (Magnetic and optical recording media manufacturing; Biological product manufacturing; Quick printing). The most concentrated industries, once the noise is eliminated, are "Jewelers' material & lapidary work manufacturing", "Sugarcane mills" and "Women's, girls' cut & sew dress manufacturing".

c) CBSA minus PSA: the sign of the CBSA-PSA difference²¹ is significantly positive at the 5% level in 125 cases (71 at 1%). It means that for around one quarter of manufacturing industries the level of spatial concentration given by the CBSA classification is significantly bigger (i.e., the industry employment has a more heterogeneous distribution across space) than when using the PSA counterfactual.

Table 1.4 reports the first 50 industries according to the CBSA-PSA value, limited to the sample of industries with a 5% significant difference both in the CBSA-PSA and CBSA-Noise difference. Interestingly, these industries do not show an immediate clear similarity in labour or skill intensity: both hi-tech and labour-intensive industries are in the list. E.g., among the first ten industries in the ranking, we find a few textile industries, as well as space vehicle manufacturing and optic fibre manufacturing. This in turn supports the hypothesis

²¹ From this point on, every mention of Noise and PSA values refers to the average value across 1000 iterations of the zoning algorithm.

of a non monotonic relationship between input intensity and spatial concentration, and confirms the advantages of a non parametric approach.

There are also a few industries for which the CBSA-PSA difference is significantly negative: they are 23 in total, but only nine if the analysis is restricted to the industries with a 5% significant CBSA-Noise difference. These nine industries are: Electrometallurgical ferroalloy product manufacturing; Oil and gas field machinery and equipment manufacturing; Petrochemical manufacturing; Schiffli machine embroidery; Animal (except poultry) slaughtering; All other basic organic chemical manufacturing; Sugarcane mills; Carbon black manufacturing; Softwood veneer and plywood manufacturing; Dried and dehydrated food manufacturing. For them, the employment is more heterogeneously distributed in the PSA counterfactual than in the CBSA dataset. This may apparently depend on the specific pattern of within-CBSA location of these industries. An industry that is systematically located in “outer counties” of CBSAs may be more concentrated in the PSA dataset because it may happen that a few PSAs are composed only by outer counties. However, given that both the employment and area distributions of the PSA dataset mimic the correspondent distribution in the CBSA one, and considering that the outer counties are less densely populated, the number of PSAs composed only by outer counties should be limited, as they have a lower density (thus smaller population and bigger area) than CBSAs. Therefore, to the extent that imposing the CBSA area and employment distribution to PSAs is also avoiding big unbalances in the central/outer counties ratio, the aforementioned bias should be small and infrequent, not being a relevant concern for the interpretation of our results.

The CBSA and PSA difference is an absolute measure of the effect of labour market determinants. In order to assess the labour market effect conditional on the total amount of concentration, we calculate the ratio between the two “de-noised” values:

$$LMDI = (CBSA - Noise)/(PSA - Noise) \quad (1.10)$$

We define this the “Labour Market Determinants Index” (LMDI). This value represents an industry-specific estimate of the importance of the labour market determinants, relative to the whole sample of industries and the effects of the other concentration determinants. The subtraction of the Noise from both the numerator and denominator consents to remove the spurious concentration component and to not underestimate the index when this component is large. The value is reported in the last column of table 1.4,²² and it is generally highly correlated with the CBSA – PSA one.

1.5.4 Comparison of the “CBSA minus Noise” values with the Ellison-Glaeser Index

The values of the CBSA-Noise difference provide some empirical evidence on the limits of the EG index. A similar exercise has already been done by Duranton and Overman (2005), who compared the results from a Point Pattern Analysis (PPA) of concentration of UK manufacturing to the values of the EG index at County level. They find a remarkable difference in the industry ranking, quantifiable in a Spearman rank correlation of 0.41.

Interestingly, by comparing our measure of concentration (CBSA – Noise) with the EG index calculated on the same data and spatial units (CBSA) we find a similar result, i.e.,

²² The table reports only values significant at 5% level in both the CBSA-PSA and CBSA-NOISE difference.

Table 1.2. G Employment Concentration Index in the noise counterfactual, top 20 industries

INDUSTRY NAME	NR. PLANTS	AV. EMP.	RANK HERF.	G NOISE*	G CBSA - G NOISE*	RANK CBSA-NOISE	EG NOISE*
Primary smelting & refining of copper	13	231.3	6	0.381	-0.114	472	0.299
Engineered wood member (exc truss) mfg	12	72.6	1	0.283	0.064	39	0.023
Roasted nuts & peanut butter mfg	133	73.2	127	0.247	-0.191	473	0.282
Cellulose organic fiber mfg	11	253.9	2	0.233	0.07	34	0.024
Other ordnance & accessories mfg	50	152.3	3	0.227	0.049	61	0.024
Alumina refining	11	189.5	4	0.224	0.057	45	0.023
Other missile, space veh parts & aux equip mfg	45	166.9	7	0.197	0.08	25	0.015
Missile, space veh propulsion unit & parts mfg	20	699.7	13	0.188	-0.047	465	0.103
Women's footwear (exc athletic) mfg	90	65.2	19	0.184	-0.027	464	0.135
Magnetic and optical recording media mfg	242	57.1	120	0.169	-0.098	470	0.183
Gum & wood chemical mfg	48	38.7	11	0.156	0.034	95	0.046
Overhead crane, hoist & monorail system mfg	273	55.4	98	0.149	-0.11	471	0.153
House slipper mfg	16	98.6	9	0.138	0.008	294	0.023
Motor home mfg	79	200.5	57	0.128	0.036	88	0.105
Small arms ammunition mfg	110	65.4	18	0.122	0.009	289	0.035
Cane sugar refining	15	223.5	16	0.115	0.064	40	0.02
Household laundry equipment mfg	20	783.4	10	0.114	0.064	41	-0.012
Household vacuum cleaner mfg	39	285.5	17	0.114	0.052	55	0.019
Guided missile & space vehicle mfg	16	3419.3	12	0.11	0.074	30	-0.009
Cigarette mfg	14	1469.2	5	0.109	0.178	8	-0.164

Note: the tables report the decreasing ranking of the top 20 industries, according to the value of the G index in the Noise counterfactual (fourth column).

* Average value across 1,000 iterations.

Table 1.3. CBSA minus NOISE, top 20 industries

INDUSTRY NAME	TOT. EMP.	NR. PLANTS	AV. EMP.	HERF. INDEX	Rank HERF.	EG CBSA	G CBSA	G NOISE*	G CBSA - G NOISE*
Jewelers' material & lapidary work mfg	5776	372	15.5	0.034	146	0.304	0.291	0.030	0.261
Sugarcane mills	3433	35	98.1	0.088	34	0.263	0.317	0.076	0.241
Women's, girls' cut & sew dress mfg	21086	756	27.9	0.010	344	0.311	0.239	0.024	0.214
Oil & gas field machinery & equipment mfg	23636	498	47.5	0.010	331	0.228	0.222	0.011	0.211
Petrochemical mfg	10287	50	205.7	0.051	101	0.267	0.288	0.084	0.203
Carpet & rug mills	40839	417	97.9	0.014	282	0.199	0.212	0.012	0.199
Women's, girls' cut & sew blouse mfg	16538	596	27.7	0.010	341	0.256	0.202	0.010	0.191
Cigarette mfg	20569	14	1469.2	0.267	5	0.024	0.288	0.109	0.178
Women's, girls', infants, cut, sew apparel contr	99024	5873	16.9	0.001	469	0.247	0.179	0.003	0.176
Costume jewelry & novelty mfg	10227	828	12.4	0.011	320	0.198	0.180	0.011	0.169
Motor vehicle air-conditioning mfg	24575	65	378.1	0.196	8	0.041	0.230	0.072	0.158
Ethyl alcohol mfg	1107	27	41.0	0.094	32	0.144	0.229	0.083	0.147
Fiber optic cable mfg	11742	74	158.7	0.116	22	0.089	0.182	0.058	0.125
Tobacco stemming & redrying	4976	31	160.5	0.086	39	0.106	0.193	0.073	0.119
Photographic & photocopying equipment mfg	19317	385	50.2	0.068	62	0.097	0.141	0.023	0.118
Sanitary paper product mfg	23799	118	201.7	0.051	102	0.094	0.145	0.037	0.108
Women's, girls' cut & sew other outerwear mfg	38665	1520	25.4	0.008	368	0.158	0.117	0.016	0.100
Choc & confectionery mfg from cacao beans	9628	142	67.8	0.144	14	0.010	0.153	0.053	0.100
Photo film, paper, plate & chemical mfg	26834	323	83.1	0.139	15	0.014	0.141	0.041	0.099
Other hosiery & sock mills	21129	274	77.1	0.017	256	0.091	0.115	0.016	0.099

Note: the tables report the decreasing ranking of the top 20 industries, according to the difference of the values of the G index in the CBSA dataset and Noise counterfactual, respectively (sixth column).

* Average value across 1,000 iterations.

Table 1.4. CBSA - PSA, top 50 industries, 6-digit

INDUSTRY NAME	TOT. EMP.	NR. PLANTS	NR. COUNTY	G CBSA	G CBSA - G NOISE*	G CBSA - G PSA*	LMDI
Jewelers' material & lapidary work mfg	5776	372	111	0.2914	0.261	0.117	1.81
Women's, girls' cut & sew dress mfg	21086	756	176	0.2385	0.214	0.055	1.35
Ethyl alcohol mfg	1107	27	22	0.2293	0.147	0.047	1.47
Guided missile & space vehicle mfg	54709	16	16	0.1832	0.074	0.041	2.24
Men's & boys' neckwear mfg	3781	100	43	0.1328	0.084	0.037	1.77
Upholstered household furniture mfg	71978	1443	447	0.0822	0.074	0.030	1.69
Women's, girls', infants, cut, sew apparel contr	99024	5873	687	0.1792	0.176	0.030	1.20
Jewelry (exc costume) mfg	32813	2114	401	0.0923	0.073	0.029	1.65
Fiber optic cable mfg	11742	74	57	0.1823	0.125	0.029	1.30
Costume jewelry & novelty mfg	10227	828	269	0.1798	0.169	0.028	1.20
Motor vehicle metal stamping	118041	682	244	0.0762	0.068	0.026	1.61
Fur & leather apparel mfg	2416	256	88	0.1186	0.099	0.023	1.30
Welding & soldering equipment mfg	19687	266	162	0.0829	0.054	0.022	1.70
Other hosiery & sock mills	21129	274	101	0.1151	0.099	0.020	1.25
Other footwear mfg	1459	70	52	0.1054	0.049	0.017	1.51
Women's, girls' cut & sew other outerwear mfg	38665	1520	336	0.1165	0.100	0.015	1.17
Other metalworking machinery mfg	17726	433	201	0.047	0.037	0.015	1.64
Special die, tool, die set, jig & fixture mfg	74585	4117	723	0.0357	0.033	0.014	1.79
Copper wire (except mechanical) drawing	4137	68	60	0.0753	0.025	0.014	2.32
Textile machinery mfg	11080	427	187	0.0573	0.044	0.014	1.49
Other pressed & blown glass & glassware mfg	34393	475	273	0.0485	0.028	0.013	1.84
Cane sugar refining	3352	15	11	0.1781	0.064	0.012	1.22
Aircraft engine & engine parts mfg	80072	371	179	0.066	0.053	0.010	1.24
Machine tool (metal cutting types) mfg	24054	463	213	0.0344	0.017	0.010	2.33
Rolled steel shape mfg	15503	269	167	0.0488	0.030	0.009	1.43
Women's, girls' cut & sew blouse mfg	16538	596	156	0.2015	0.191	0.009	1.05
Electric lamp bulb & part mfg	13257	87	67	0.0662	0.031	0.008	1.38
Other aircraft part & auxiliary equipment mfg	102716	1018	321	0.0636	0.035	0.008	1.31
Flat glass mfg	11681	61	55	0.064	0.027	0.008	1.44
Nonupholstered wood household furniture mfg	108471	3323	846	0.031	0.026	0.008	1.45
Cement mfg	15273	226	168	0.0234	0.013	0.008	2.48
Synthetic organic dye & pigment mfg	7496	121	83	0.0495	0.027	0.008	1.40
Gasoline engine & engine parts mfg	78943	780	375	0.0508	0.037	0.007	1.23
Search, detection & navigation instrument mfg	168093	611	237	0.0447	0.029	0.006	1.28
Metal heat treating	21264	754	266	0.0207	0.016	0.006	1.70
Women's, girls' cut & sew lingerie mfg	11881	208	98	0.0398	0.023	0.006	1.39
Secondary smelting & alloying of aluminium	6200	159	113	0.0381	0.021	0.006	1.43
Photographic & photocopying equipment mfg	19317	385	197	0.1407	0.118	0.006	1.06
Power boiler & heat exchanger mfg	19986	407	248	0.0253	0.017	0.006	1.54
Mineral wool mfg	20031	266	188	0.0356	0.019	0.006	1.45
Other apparel accessories & other apparel mfg	23261	1626	435	0.0418	0.035	0.006	1.19
Industrial & commercial fan & blower mfg	10224	169	121	0.0336	0.014	0.005	1.64
Household vacuum cleaner mfg	11134	39	32	0.1662	0.052	0.005	1.11
Coated & laminated pkg paper & plastics film mfg	5997	98	79	0.0381	0.013	0.005	1.57
All other cut & sew apparel mfg	8330	335	171	0.0194	0.008	0.005	2.49
Electroplating, polish, anodize, coloring	73042	3179	599	0.0162	0.013	0.005	1.56
Industrial mold mfg	44980	2193	514	0.0198	0.016	0.005	1.39
Industrial pattern mfg	7762	648	274	0.028	0.015	0.004	1.43
All other motor vehicle parts mfg	151673	1292	533	0.0197	0.013	0.004	1.47
Inorganic dye & pigment mfg	6959	73	60	0.0545	0.017	0.004	1.31

Note: the table reports the decreasing ranking of the top 50 industries, according to the difference of the values of the G index in the CBSA dataset and PSA counterfactual, respectively (sixth column). The table reports only values significant at 5% level in both the CBSA-PSA and CBSA-NOISE difference.

* Average value across 1,000 iterations.

a Spearman rank correlation of 0.51 across all the 473 industries in the sample. However, while the difference with a Point Pattern Analysis may be attributable to a different concept of space definition – continuous vs. discrete – we calculate the two concentration measures using the same zoning criterion and exactly the same spatial units, thus providing stronger evidence on the limits of the EG approach.

Another evidence of the bias of the EG index – at least in this specific sample – comes from the analysis of its values in the Noise counterfactual (average across 1,000 iterations). Given that it reproduces a completely random location process, we would ideally expect an unbiased index to be very close to zero. Yet, the EG index may absorb some spurious concentration as it is prone to the geographical bias and to the other potential misspecification discussed earlier. This is confirmed by data: we obtain an average of 0.02 across the 473 industries in the sample, which, according to Ellison and Glaeser (1997), should be interpreted as evidence of localization. The distribution appears to be strongly skewed toward the right hand side, which means that the overestimate is particularly high for a few industries (Table 1.5). Overall, 102 industries present a value bigger than 0.02, which implies that the bias is substantial for more than one fifth of industries.

Table 1.5. EG index in the noise counterfactual

Statistics	EG noise
Mean	0.0208
Max	0.2991
P25	0.0177
P50	0.0183
P75	0.0196
P90	0.0288

As we discussed earlier in the paper, the imprecision of the EG index may arise from two sources: an approximate measurement of the plant employment distribution based on the Herfindahl index, and the arbitrary aggregation of plants into given (administrative) spatial units. Our metrics, based on the Noise counterfactual, allows us to overcome both the problems, and can therefore provide the basis for further research on the distribution of the EG index under the null hypothesis of random location pattern.

1.5.5 Analysis at 4-digit level

In this section we present the results obtained using a wider industry classification, i.e., the 86 manufacturing sectors reported at the 4-digit industry level. There are two main reasons why this may be useful: first, the 6-digit is an extremely detailed classification which may produce weird results for a few peculiar branches. Second, the sectors reported in the our dataset (County Business Pattern from US Census) refer to the prevalent activity of the plants; it is likely that some plants are actually multi-product, thus, again, a detailed classification may be misleading.

The results – reported in table 1.6 – are coherent with the 6-digit analysis. The Spearman rank correlation between the G index at 4 and 6 digits is positive and significant for all the three different datasets (CBSAs, PSAs, Noise). 31, out of 86, manufacturing sectors exhibit a positive and significant (at 5%) CBSA – PSA difference, while the CBSA – Noise difference is significantly positive in 74 cases.

Similarly to the 6-digits analysis, the top sectors in the CBSA-PSA ranking do not show a linear dependence on labour input intensity, and a mix between low and high skill

activities emerges. Again, it would be extremely complex to recognize such a pattern with a parametric analysis.

1.5.6 Does the labour market matter for concentration?

In order to assess to what extent the concentration driven by labour market determinants explains the general pattern of concentration, we regress the index reporting the total amount of concentration (CBSA – Noise) on our estimate of the labour market effect (CBSA – PSA). A positive and significant coefficient would reveal a systematic effect of the labour market determinants in explaining the overall pattern of spatial concentration.

At 6-digit level, results are weak: the value of the coefficient is insignificant, and the R^2 is equal to 0.04 (table 1.7, col. 1). Nevertheless, both the Pearson and Spearman correlation coefficients are significant, with a value of 0.18 and 0.26, respectively. Interestingly, when we restrict the sample to observations which show a significant CBSA-PSA difference, i.e., industries for which labour market determinants play a role in explaining agglomeration, the R^2 reach the value of 0.67 and the coefficient is now highly significant (table 1.7, col. 2). This seems to suggest that, for around the 30% of industries for which the labour market is a significant determinant of agglomeration, this latter effect is predominant over the other determinants (e.g. input-output linkages, natural advantages, etc.).

At 4-digit level, the CBSA – PSA difference explains much more of the variations in the CBSA minus Noise values. The regression of the latter variable on the former now produces a robust t statistic equal to 2.69 and a R^2 of 0.28 (table 1.7, col 3). This is

Table 1.6. CBSA - PSA, 4-digit*, positive values

INDUSTRY NAME	TOT. EMP.	NR. PLANTS	G CBSA	G PSA*	G CBSA - G NOISE*	G CBSA - G PSA*	LMDI
Cut & sew apparel mfg	311677	11789	0.078	0.066	0.072	0.0122	1.19
Metalworking machinery mfg	218427	9260	0.022	0.015	0.020	0.0070	1.52
Aerospace product & parts mfg	412944	1691	0.041	0.035	0.032	0.0063	1.25
Motor vehicle parts mfg	740523	5104	0.025	0.019	0.023	0.0060	1.36
Apparel accessories & other apparel mfg	40112	2163	0.031	0.026	0.026	0.0051	1.28
Motor vehicle mfg	255966	378	0.037	0.032	0.019	0.0049	1.35
HH & institutional furniture & kitchen cabinet mfg	339880	12941	0.016	0.011	0.013	0.0048	1.49
Nav, measuring, medical, control instruments mfg	443652	4934	0.017	0.013	0.012	0.0041	1.57
Motor vehicle body & trailer mfg	125491	1748	0.035	0.031	0.025	0.0036	1.15
Steel product mfg from purchased steel	63958	863	0.014	0.011	0.009	0.0028	1.35
Clay product & refractory mfg	69827	1616	0.011	0.008	0.006	0.0027	1.99
Other transportation equipment mfg	38752	757	0.033	0.030	0.022	0.0027	1.12
Coating, engrave, heat treating & oth activity	150012	5917	0.010	0.007	0.009	0.0026	1.52
Glass & glass product mfg	117749	2124	0.013	0.011	0.008	0.0025	1.35
Mach shops, turn prod, screw, nut, bolt mfg	389848	24141	0.006	0.004	0.005	0.0019	1.65
Other nonmetallic mineral product mfg	68044	2286	0.009	0.007	0.006	0.0019	1.62
Cement & concrete product mfg	196681	7739	0.005	0.003	0.004	0.0017	1.75
Ag, construction & mining machinery mfg	161659	2442	0.017	0.015	0.014	0.0016	1.19
Other wood product mfg	258822	8661	0.009	0.007	0.007	0.0015	1.19
Other electrical equipment & component mfg	202114	2378	0.007	0.005	0.002	0.0015	2.12
Animal food mfg	41536	1314	0.011	0.009	0.007	0.0015	1.25
Boiler, tank & shipping container mfg	86680	1603	0.008	0.006	0.005	0.0015	1.30
Other fabricated metal product mfg	290455	6784	0.005	0.004	0.003	0.0015	1.76
Communications equipment mfg	244061	2165	0.022	0.021	0.012	0.0014	1.10
Other miscellaneous mfg	390362	17629	0.005	0.004	0.003	0.0012	1.46
Foundries	197312	2537	0.011	0.010	0.007	0.0010	1.24
Dairy product mfg	113462	1510	0.007	0.006	0.004	0.0010	1.36
Rubber product mfg	190260	2479	0.010	0.009	0.007	0.0006	1.21
Other general purpose machinery mfg	307881	6209	0.005	0.004	0.003	0.0005	1.38
Plastics product mfg	763061	12501	0.003	0.003	0.002	0.0004	1.37
Architectural & structural metals mfg	369395	10911	0.002	0.002	0.001	-0.0002	0.97

Note: the tables report the decreasing ranking of industries according to the difference of the values of the G index in the CBSA dataset and PSA counterfactual, respectively (sixth column). The table reports only values significant at 5% level in both the CBSA-PSA and CBSA-NOISE difference.

* Average value across 1,000 iterations.

definitively a robust association, suggesting that labour market determinants may have a strong effect in shaping the spatial concentration of industries, on one side; on the other side, it suggests that the 6-digit level may be too “noisy” and detailed to investigate the relationship. Again, after restricting the sample to industries with a significant CBSA-PSA difference, the R^2 increases remarkably.

Table 1.7. Regression of total agglomeration on labour market effect

VARIABLES	(2)	(3)	(4)	(5)
Sample	CBSA – Noise 6 d.	CBSA – Noise 6 d., sign. only	CBSA – Noise 4 d.	CBSA – Noise 4 d., sign. only
CBSA – PSA	0.675 (0.667)	2.643*** (0.341)	2.779** (1.149)	4.829*** (0.798)
Constant	0.0323*** (0.00303)	0.0118*** (0.00344)	0.0109*** (0.00184)	-0.00103 (0.00182)
Observations	298	95	74	31
R^2	0.042	0.672	0.289	0.814

Heteroschedasticity robust standard errors in parentheses

*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

1.5.7 Further robustness: regression of the CBSA-PSA difference on industry variables

In Appendix B, we present the results of a regression of the CBSA – PSA difference on a number on industry-specific variables. The only variables significant at 5% are the average plant employment²³ and the number of production workers over total shipment value. Results are only indicative, given the small number of observations and the supposed inadequacy of a linear approach to investigate a data generating process which we expect to

²³ Interestingly, this is in line with findings from Lafourcade and Mion (2007), who showed that in Italy “large plants tend to cluster within narrow geographical areas such as labour market” (p. 48), while smaller plants tend to exhibit colocation at a wider geographical level. The authors comment on their results arguing that large plants are more export oriented, which in turn implies that their location is more sensitive to Marshallian labour market externalities rather than local market potential.

be non monotonic (let alone the non linearity). The refinement of the regression approach is left for future research. Nevertheless, the value of the coefficients is supporting the hypothesis that CBSA-PSA variation across industries is not random and is a good measure of the labour market effect.

1.6 Conclusions

In this paper, we develop a “gerrymandering” approach – which exploits the “bright side” of the Modifiable Areal Unit Problem (MAUP) – to measure industrial concentration and to disentangle the effect of labour market determinants.

We calculate the value of a gross concentration index in two different datasets, obtained by applying different zoning procedures to the same partition of US territory. The first procedure follows the commuting-based Core Based Statistical Areas (CBSA) definition, which is expected to maximize (among all the possible alternatives based on the same building blocks) the within-homogeneity and between-heterogeneity of labour market characteristics across spatial units. In this dataset, the effect of the “labour market” determinants is maximized. The second procedure creates a distribution of 1,000 counterfactuals, each of them composed by randomly aggregating the same counties which form the CBSAs into internally connected “Pseudo Statistical Areas” (PSAs). In this second dataset, all the “non labour market” determinants have the same effects of the previous one, while the “labour market determinants” effect is reduced. The difference from the concentration value found with the first procedure and the average across the 1,000 iterations of the second counter-

factual quantifies the effect of labour market determinants for a given industry. We find this value to be significantly positive in 125, out of 473, manufacturing sectors.

We also propose a new method to measure industrial agglomeration. We empirically estimate the bias which affects traditional gross concentration indices by creating a distribution of “Noise counterfactuals”, generated by i) shuffling plants randomly across plant sites, and ii) randomly aggregating small spatial units (Zip Code Areas) into bigger spatial units – of the same size of the CBSAs – without any contiguity constraint. The first step captures the spurious concentration component given by the industry plant size distribution, while the second step absorbs the geographical bias given by the arbitrary aggregation of events into exogenous spatial units. The amount of industry-specific concentration found in the Noise scenario corresponds to the spurious component comprised in the CBSAs dataset, while the value found in the CBSA dataset minus the Noise is a measure of industrial concentration which satisfies the five benchmark requirements listed in Combes and Overman (2004). A comparison of latter results with the corresponding values of Ellison-Glaser index reveals remarkable differences.

The ranking of industries according to the CBSA – PSA difference suggests that industries which are dependent on labour market characteristics in choosing their location are highly heterogeneous in skill and labour intensity. Furthermore, the CBSA-PSA difference is significantly associated with industry-specific labour variables (average plant employment and number of production workers over total shipment). Finally, results also show that labour market determinants explain around one quarter of the overall pattern of

concentration, although the effect is more easily recognizable at a wider level of industry classification.

1.A The PSA algorithm

Aggregating counties into PSAs is not an easy task. PSAs must be comparable to CBSAs in number, size distribution, and shape, without excessively limiting, at the same time, the degree of “randomness” of the aggregation. This has required a substantial effort in creating the algorithm, which has been developed and compiled by the author in MATLAB. We report a detailed description of its structure below. Original codes are available upon request.

As a first step, a maximum number of counties (30) and a maximum area for every individual PSA are set. These upper limits are high and binding only in exceptional cases. Subsequently, the algorithm goes through the following operations:

- 1) All the counties with a labour force bigger than one million people are identified, and PSAs are composed around them, using the main aggregation loop described below.

- 2) All the counties with a labour force smaller than 600 people are identified, and PSAs are composed around them, again using the main aggregation loop.

Point 1 and 2 are performed at the beginning of the process because both biggest and smallest counties require many “degree of freedom” (i.e., many unassigned counties around the initial county): big counties generally belong to CBSAs composed by many

other counties; small counties must not be left alone in an one-county PSA, otherwise such a PSA would be much smaller than the smallest CBSA (which has around 1,8k workers).

3) 460 one-county PSAs are composed, randomly selecting counties with total employment bigger than 1,8k (approximately the employment of the smallest CBSA) and smaller than 40k. The CBSA size distribution includes 570 one-county CBSAs. However, a number of one-county PSAs - on average around 110 - is created also in the following steps of the algorithm.

4) All the remaining counties are aggregated into PSAs (always using the main aggregation loop)

5) One-county PSAs which are surrounded by counties belonging to the same PSA (i.e., there is an "island" PSA inside another one) are identified, and they are added to the surrounding PSA.

A this point, all the counties are aggregated into PSAs, but their number can be slightly smaller or bigger than the actual number of CBSAs (876). Given that generally the deviation is small (about +/- 20 units), the adjustments are not too sophisticated:

6) i) If the PSAs are more than needed, they are reduced to the targeted number by aggregating one-county PSAs into two-counties PSAs.

ii) If they are less, two-counties PSAs are split into one-county PSAs.

The aggregation process is now complete, and concentration indices can be calculated and stored. The process is then restarted from the beginning, and repeated 1000 times.

The main "aggregation loop"

Below we report a description of the commands used to aggregate counties into PSAs in steps 1,2, and 4 above. Although it is not reported in programming language, the structure is described exactly, in order to facilitate the replication by interested readers familiar with general programming syntax.

FOR each row of the contiguity matrix (i.e., for each county)

IF county *c* is not assigned and it has more than 1 neighbour

 set max pop. = pop. CBSA(*x*) * 0.6 – pop. county *c* (CBSA are listed in decreasing order of employment, *x* starts from 1 and progressively increasing; the adjusting factor of 0.6 has been selected empirically in order to correct for the fact that PSAs tend to overcome the maximum population limit as it is set before the last county is added.)

 set $q=n$ ($n-q$ = counter of counties in PSA *j*, n = counter of total counties assigned so far; every time a county is added to a PSA, n is increased by 1)

 create vector *A* = list of all contiguous counties of county *c*

 chose a random county *r* from *A*

 WHILE county *r* is already assigned and in *A* there are still some unassigned counties

 select another random county *r* from *A*

 END

 IF in *A* there are some unassigned counties (i.e., *r* is not assigned yet)

 add *r* and *c* to PSA *j* and mark them as assigned

 WHILE $n-q$ is less than the set max number of counties for a PSA and PSA *j* has pop. less than the pop. limit

create vector B = list of all contiguous counties of county r (last county added to j)

A = [A B] = list of all neighbours of all the counties assigned to j so far

Add a second column to A, reporting how many times each county is repeated in the vector

Sort the rows of matrix A in decreasing order according to the second column

Chose a row from A, randomly drawing its position from an exponential function

WHILE county r is already assigned and in A there are still some not assigned counties

chose another random county r from A

END

IF in A there are some unassigned county (i.e., r is not assigned yet)

add r to PSA j and mark it as assigned

END

END

ELSEIF county r is already assigned

Add the county c to the PSA j which county r belongs to (no pop. limits for simplicity)

n=n+1

END

ELSEIF county c has only one neighbour b

IF b has not been assigned yet

 Add b and c to a new PSA j

ELSE

 Add c to the PSA j c belongs to

END

END

$x=x+1$

END

1.B OLS regression

Table 1.8 shows the results of an OLS regression of the CBSA-PSA difference at 4-digit level for all the 86 industries in the sample. All the explanatory variables have been obtained from the US Economic Census of 1997, with the exception of the hourly compensation, which comes from the Annual Census of Manufacturing of the year 2000.

As mentioned earlier, results must be interpreted with caution, in the light of the expected non linearity (and probably non monotonicity) of the functional form relating proxies of input intensity with measures of industry concentration. Moreover, coefficients may be biased because of reverse causality and omitted variables. Nevertheless, results show at least some patterns of “partial correlation” which are consistent with the theoretical development of our methodology and the interpretation of the results. In particular, average plant employment and the total number of production workers have a positive and significant coefficient, which may suggest that the most labour intensive industries are the ones discriminating most among different labour markets. On the other side, total capital expenditure has a mild negative effect, while hourly compensation coefficient is positive (although significant only at 11%). This patchy pattern is again coherent with a rather mixed and complex phenomenon, difficult to parameterize with traditional methods.

Table 1.8. Regression output

VARIABLES	(1) CBSA-PSA	(2) CBSA-PSA
Average plant employment	0.00954** (0.00439)	0.00939** (0.00418)
Hourly compensation	0.168 (0.108)	0.154 (0.131)
Non production workers/total value of shipment	-0.303 (0.575)	-0.492 (0.578)
Production workers/total value of shipment	0.553*** (0.192)	0.388* (0.199)
Total capital expenditures/total value of shipment		-0.0453* (0.0248)
Cost of materials/total value of shipment		-0.00630 (0.00524)
Constant	-0.00512** (0.00199)	0.000834 (0.00528)
Observations	86	86
R^2	0.163	0.211

Heteroschedasticity robust standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

1.C Data

We use six-digit North American Industry Classification System (NAICS) employment data for Zip Code Areas (ZCAs) and Counties in the year 2000, freely available from the County Business Patterns (CBP) of US Census Bureau, in the form of the dataset collected by Prof. Thomas Holmes, University of Minnesota, and freely downloadable from his website.²⁴ We also use a shape file from the National Atlas of the United States and Luc Anselin's GeoDa software²⁵ to calculate a first order rook contiguity matrix needed by the PSA algorithm (described in Appendix A). In the map in Figure 2 we report the CBSA borders layer together with a map of populated places.

Because of confidentiality issues, in the CBP database many employment records are reported only in approximated form, i.e., we only know the size class of the plant. There are various ways to overcome this problem (see Isserman and Westervelt, 2006, for a survey). As we do not need a precise locality-specific estimate, we followed the most straightforward route: we ascribed to every plant the average employment of the class it belongs to (as done, for example, also by Holmes and Stevens, 2004). Another minor problem is given by the fact that ZCAs employment data over 1000 employees are merged in only one class, instead of four as it is in counties data. In order to obtain comparable data (and, again, considering that the exact estimate of the employment of each ZCA is not relevant here), we attributed to ZCAs the distribution of employment class size of the counties data (industry-wise).

²⁴ <http://www.econ.umn.edu/~holmes/>

²⁵ Freely available from <http://geodacenter.asu.edu/>

In order to get an illustration of the effectiveness of CBSAs in containing the commuting flows, we used journey-to-work data from the 2000 Census, with the result that only 9% of employees living in a CBSA work outside the same CBSA where they reside. On the other side, 25% of workforce resident in a CBSA commute outside the county they reside in. This confirms that CBSAs are well designed and do contain most of the commuting flows, and, at the same time, there is a significant cross-county commuting activity.

From the 3079 counties composing the Continental US (therefore excluding the States of Alaska, Porto Rico and Hawaii) we selected the records of the 1734 counties which are included in the 2000 standards Core Based Metropolitan and Micropolitan Statistical Areas (CBSAs). From these we eliminated 26 Micropolitan Statistical Areas which are isolated, i.e., do not share any border with other CBSAs and therefore would not show any variability among the random aggregations. We end up with a dataset of 1707 counties which account for the 97% of the total (continental) US population, and form 876 CBSAs, of which 306 are Metropolitan Statistical Areas and 570 are Micropolitan Statistical Areas. The definition procedure of Metropolitan and Micropolitan areas is exactly equivalent, but for the latter the population of the core county has to be smaller than 50,000. However, the overwhelming majority of Micropolitan Statistical Areas are composed by only one county, as a consequence of the fact that the commuting flows with neighbouring counties are limited.

The same selection of the US territory is applied to the ZCAs dataset, using the ZCAs-counties geographical equivalence list, also available from Prof. Holmes website.

Chapter 2

Does urbanisation affect rural poverty? Evidence from Indian Districts

2.1 Introduction

The typical transformation of an economy from agricultural and mainly rural to industrial and predominantly urban in the process of development has long been a well established fact (Lewis, 1954; Kuznets, 1955). However, the direct implications of this transformation on the economic welfare of the population during this process remain less apparent. In particular, what happens to surrounding rural areas when a city grows? Does the area's population receive economic benefit from it and if so, to what extent? In a period of increasing urbanisation in most developing countries, answers to these questions have important implications for development policies.

There is still very little known about the actual economic impact of urbanisation on rural areas. This paper represents one of the first efforts to fill this gap, as it tries to measure the impact of urbanisation on rural poverty in the Indian context. The paper uses district-level panel data between 1981 and 1999 to show that urbanization has been an important determinant of poverty reduction in rural areas. In our preferred estimations, we find that an increase of 100,000 urban residents in the representative district (around 21% increase from the mean) implies a decrease of between 3 and 6 percentage points in the incidence of poverty in the district's rural areas.

This analysis becomes more important when considering that most of the world's poor reside in rural areas, where the incidence of poverty is higher than in urban areas across all developing regions. In 1993 rural areas accounted for 62% of the world population and for 81% of the world's poor at the \$1/day poverty line; in 2002 after a period of intensive urbanisation the same figures stood at 58% and 76% respectively (Ravallion et al., 2007). The process of urbanisation (which mostly concerns the developing world) has been accompanied by an unequal distribution of the global reduction in poverty rates. Between 1993 and 2002 while the number of \$1/day poor in rural areas declined by 100 million, that of urban poor increased by 50 million. Ravallion et al. (2007) explain this "urbanisation of poverty" through two related arguments. First, a large number of rural poor migrated to urban areas, thus ceasing to be rural poor and either they have been lifted out of poverty in the process (through a more productive use of their work) or they have become urban poor. This is a direct (or 'first-round' in Ravallion et al. (2007) terminology) effect of urbanisation on rural poverty. Second, the process of urbanisation also impacts the welfare of those who remain in rural areas through second-round effects. The overall impact of urbanisation on rural poverty is substantial but, in the absence of data on the poverty profile of rural-urban migrants, it is not possible to distinguish between the two effects. We mainly focus on these second-round effects, trying to control for the direct effects of urbanisation on rural poverty.

Distinguishing between first and second-round effects is important. The former involves only a statistical association between urbanization and changes in rural poverty due to the change in residency of some rural poor (who may or may not be lifted out of poverty

in their move to the urban areas). This entails no causal link. On the other hand, second-round effects capture the impact of the urban population growth on the rural rate of poverty. Such a relationship is causal in nature and tells us how good or bad urbanisation is for rural poverty. In a developing country context, understanding this relationship is particularly important because most of the population in these countries will continue to be rural for at least another decade and for another three decades in the LDCs. This figure, along with the recognition that poverty has a higher incidence in rural than urban areas, suggests that it is on this rural non-migrant population that the implications of urbanisation will be most important for global poverty reduction in the near future. The focus on developing countries is essential given that almost the entire future population growth in urban areas (94% in 2005-2030) is predicted to take place in developing countries (UN, 2008).

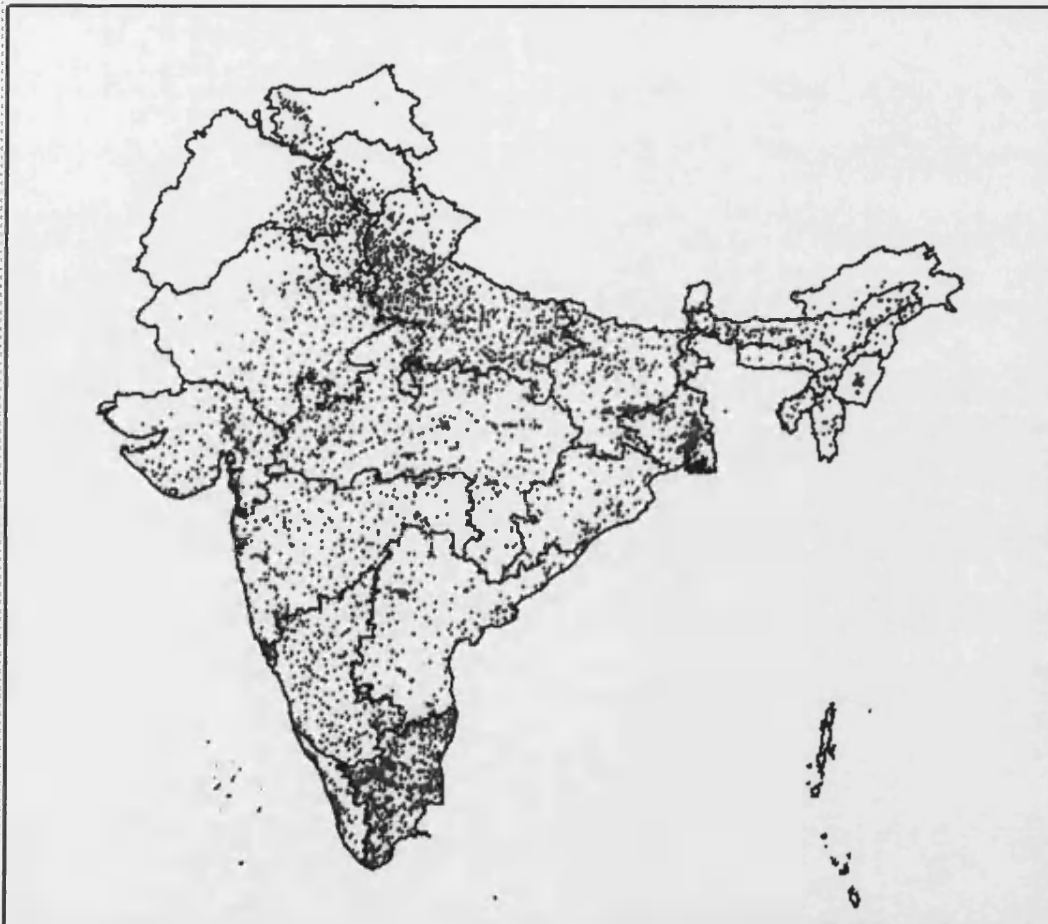
We consider Indian urbanisation at the district-level for the period 1981-1999. During this period the country urbanised at a relatively slow rate: the urban population was 23.3% of the total in 1981 and 27.8% in 2001 (Government of India, 2001). However, given the sheer size of the Indian population, this moderate increase turned into a massive rise in the absolute number of urban dwellers (126 million). This represents an increase of almost 80% in the urban population over this period. These figures mask a large variability in urbanisation patterns at the sub-national level; states have urbanised at very different rates. Among the major states, Tamil Nadu increased its share of urban population from 33% to 44% between 1981 and 2001, while Bihar maintained the same urbanisation rate over this period (13%). The differences are also evident in absolute terms: Uttar Pradesh increased its urban population by 28 million people (+140%); at the other extreme West Bengal

increased its urban population by only 8 million (+56%). Not only are the urbanisation dynamics different, but so is the geographical spread of urban areas. Figure 2.5 shows that the density of towns is concentrated in Northern India, roughly in the area along the Ganges river and in the South-East (Tamil Nadu in particular). Other areas, such as Andhra Pradesh, Madhya Pradesh and the North-West have significantly lower densities.

This variability (both in levels and in changes) is even more remarkable at the district level, as the left hand-side map in figure 2.6 shows. For instance, a district like Idukki in Kerala increased its urban population by 13,000 (+29%) between 1981 and 2001, while the urban population in Rangareddi (Andhra Pradesh) increased by 1.6 million (+416%) and in Pune (Maharashtra) by 2.4 million (+130%) over the same period. In the subsequent analysis we try to exploit this variability to identify the impact of urbanisation on rural poverty.

In this period India also provides an interesting case in terms of the policy environment and economic performance because the country experienced structural changes in economic policy, rate of growth, and poverty levels. After a long period of economic planning and import substitution industrialisation, the government started reforming the economy toward a more liberal regime in 1991. This change was brought about by the external payment crisis due to the government's deficit spending. Possibly helped by the liberalisation of the economy, economic growth took off since the mid-1980s, and more evidently since 1993, having increased more rapidly than in the 1960s and 1970s (Datt and Ravallion, 2002). Despite disagreements on the extent to which economic growth increased the welfare of India's poor, poverty in India declined steadily in the 1990s, particularly in rural

Fig. 2.5. Indian towns (2001 Census)



Note: the State of Delhi is excluded from the map

Source: Authors' elaboration on data from Indian Census 2001, and data on city spatial coordinates from Indian Gazetteer and GPSvisualizer.com.

areas (Kijima and Lanjouw, 2003). The geography of the decrease in the share of poor, however, is extremely variegated, as the right hand side map in figure 2.6 shows. While in many districts more than 30% of rural population was lifted out of poverty between 1983 and 1999, for around a quarter of them the share of poverty has remained roughly constant or has even worsened over the same period.

This paper's geographical focus is particularly important as India is the country with the largest number of both rural and urban poor. Its number of \$1/day rural poor in 2002 was over 316 million, representing 36% of the world's rural poor. Moreover, its urbanisation process is still in its infancy with only 28% of the population being urban in 2000. The country is expected to add a further 280 million urban dwellers by 2030. Thus estimating the impact of urbanisation on rural poverty in India may help identify the potential effects of this expected massive growth of urban population on the world's largest stock of rural poor.

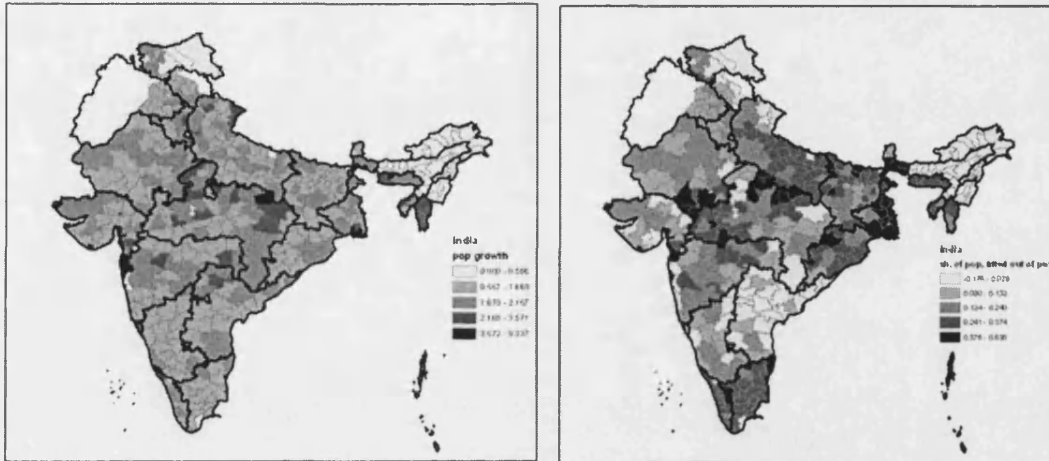
2.2 Urbanization and rural poverty: Channels

Why would the increase in urban population have an impact on poverty in surrounding rural areas? There are various ways in which urbanisation and rural poverty are linked. We can distinguish between a simple composition effect due to migration of poor from rural to urban areas (first round effect), and a spillover effect due to positive externalities of urbanization on surrounding urban areas (second round effect). In the following, we

Fig. 2.6. Urban population growth and poverty reduction, by district 1981-99

(a) Urban population growth (%), 1981-97

(b) % of rural pop. lifted out of poverty, 1983-99



Note: the map (b) reports the difference between the district poverty share in 1983 and 1999. E.g., a value of 0.30 means that in 1983 the share of poor rural population was 0.30 bigger than in 1999. The State of Delhi is excluded from the map
Source: Authors' elaboration on Indian Census and NSS (various rounds)

analyse the main mechanisms through which the latter effect may take place. Then we discuss the way in which we can try to isolate second-round from first-round effects.

2.2.1 Second round effects

There are at least six main indirect channels through which urban population growth may affect rural poverty in surrounding areas: backward linkages, rural non-farm employment, remittances, agricultural productivity, rural land prices and consumer prices.

Backward linkages: an expanding urban area (both in terms of population and income) will generate an increase in the demand for rural goods. For perishable products and in general for those products without spatially integrated markets (e.g. due to high transportation costs), such a demand will typically be met by surrounding rural areas; while the

other agricultural products could be provided by locations farther away. This is linked to an idea that goes back to von Thünen's (1966) theory of concentric circles of agricultural specialisation around cities that is determined by the size of transport costs. Rural locations close to urban areas specialise in high transportation cost goods, while locations farther away specialize in lower transport cost commodities. The farther one moves away from cities the more likely it is for rural communities to be self-subsistent in both agricultural and non-agricultural commodities. This is similar to the pattern found by Fafchamps and Shilpi (2003) for Nepal.

This channel is likely to operate via an income as well as a substitution effect. The former is related to the increased demand for agricultural goods due to higher incomes in urban areas relative to rural areas. Such a higher income is usually explained by urbanisation economies: urban areas have denser markets for products and factors, which raise labour productivity and wages over the level of rural areas (see Fujita et al., 1999). The substitution effect relates to the increased share of higher value added products in total agricultural demand typical of more sophisticated urban consumers. Empirical evidence confirms this composition effect. Parthasarathy Rao et al., 2004 found that Indian districts with an urban population over 1.5 million have a significantly higher share of high value commodities than the other districts. Thanh et al. (2008) show that per capita consumption of high value fruit in Vietnam has increased faster in urban than in rural areas over the nineties.

Rural non-farm employment: expanding urban areas may also favour the diversification of economic activity away from farming, which typically has a positive effect on

incomes (see e.g. Berdegue et al., 2001; Lanjouw and Shariff, 2002). This effect is particularly important in rural areas surrounding the cities. Three concomitant effects may explain such increased diversification. First, proximity to cities may allow part of the peripheral urban workforce to commute to the city to work. This in turn generates suburban non-farm jobs in services, such as consumer services and retail trade, which are needed by the growing commuter population. Second, as cities provide dense markets to trade goods and services more efficiently, rural households close to cities may afford to specialise in certain economic activities (based on their comparative advantage), relying on the market for their other consumption and input needs (Fafchamps and Shilpi, 2005). This more extensive specialisation should boost productivity and income (Becker and Murphy, 1992). Third, proximity to urban areas stimulates non-farm activities instrumental to agricultural trade (which is increased by urbanization), such as transport and marketing. Recent evidence from Asia provides strong support for the effect of cities in stimulating high return non-farm employment in nearby rural areas (see Fafchamps and Shilpi, 2003 on Nepal, Deichmann et al., 2008 on Bangladesh and Thanh et al., 2008 on Vietnam). On the other hand, and consistent with this line of argument, isolated rural communities do not tend to specialise and rely on subsistence activities dominated by farming. The growth of urban areas would raise the share of rural areas that are close enough to cities to develop a substantial non-farm employment base.

Remittances: remittances sent back to rural households of origin by rural-urban migrants constitutes another potentially important second-round effect of urbanization on rural poverty. The vast majority of rural-urban migrants (between 80% and 90%) send re-

mittances home although with varying proportions of income and frequency (Ellis, 1998). To the extent that urbanization is (partly) fuelled by rural-urban migration, this growth may be associated with larger remittance flows to the rural place of origin. The positive effects of remittances in reducing resource constraints for rural households as well as providing insurance against adverse shocks (as their income is uncorrelated with risk factors in agriculture) have been shown by the literature (Stark, 1980, Stark and Lucas, 1988). On the other hand the migrant's family often provides economic supports (monetary or in kind) to the migrant during his initial stay in the urban area. This support aimed at covering the fixed costs of migration can be interpreted as an investment whose main return is the counter urban-to-rural remittances flow which is received afterwards (Stark, 1980). This urban-to-rural remittance flow may somewhat reduce the net resources transferred to rural areas by urban workers.

Agricultural (labour) productivity: urbanisation and rural poverty can also be linked by the changes in rural labour supply that accompany the urbanisation process. To the extent that rural-urban migration reduces the rural labour supply, this may increase (reduce the decrease of) agricultural labour productivity, given the fixed land supply and diminishing marginal returns to land. This may pose some upward pressure on rural wages. There is indeed evidence from some areas of India of out-migration from rural areas being associated to higher wages in sending areas (Jha, 2008).

Rural land prices: the growth of cities may increase agricultural land prices (owned by farmers) in nearby rural areas due to the higher demand for agricultural land for residential purposes. This may generate increased income for landowners through sale or lease, or

through enhanced access to credit markets, where land acts as collateral. Some evidence from the US indicates that expected (urban) development rents are a relatively large component of agricultural land values in US counties which are near or contain urban areas (Plantinga et al., 2002). The impact on rural poverty through this channel depends on the way this increased income is distributed across the rural population. Typically, if land is very concentrated, this channel is likely to benefit a few landowners, potentially restricting access to waged agricultural employment for the landless population. To illustrate, let us assume the extreme case of all rural land concentrated in the hands of one landowner, who employs labour to cultivate it. If the growth of the nearby city pushes the price of the land above the expected value of the discounted stream of profits from cultivating the land, the landowner will sell it. This would leave all the agricultural labourers in the district unemployed. The net effect on poverty will depend on the extent to which the new use of the land will be able to absorb labour (e.g. via construction-related employment). However, given the constraints to the reallocation of agricultural labour across sectors and the high labour intensity of agriculture, we would expect the net effect on rural poverty to be adverse (i.e. increase in rural poverty) when land is highly concentrated (and vice-versa).

Consumer prices: because the growth of a city is associated with lower consumer prices, this may benefit surrounding rural consumers who have access to urban markets. This effect may be due to increased competition among a larger number of producers in the growing urban area as well as to thicker market effects in both factors' and goods' markets (e.g. Fujita et al., 1999).

A further potential channel may relate to early arguments made by Jacobs (1969) and Dore (1987) that agriculture in rural areas surrounding cities also benefits from spillover effects in technology and marketing. However, to the best of our knowledge, no specific evidence has been provided in support of this view yet.

Table 2.9 summarises the expected net effects of these second-round channels on rural poverty as well as their expected reach on rural areas according to the discussion above. The total net effect of urbanization on rural poverty is predicted to be negative (i.e. poverty reducing) with the bulk of the effects being felt at a relatively small distance to the urban area (in surrounding rural areas). The next sections will detail the methodology used to test these hypotheses by measuring this total net effect in the case of Indian districts.

Table 2.9. Urban population growth

	Predicted net effect	Expected reach of the effect
Backward linkages	Negative	Nearby rural
Share of non-farm employment	Negative	Peri-urban
Remittances	Negative	Rural
Changes in agricultural productivity	Negative	Rural
Rural land prices	Pos/Neg (depending on land concentration)	Nearby rural
Consumer prices	Negative	Nearby rural

Note: Reach of the expected effect is defined in descending order of distance from the urban area as: Rural; Nearby rural and Peri-urban.

Source: Authors' elaboration on the basis of main text.

2.2.2 Disentangling first and second round effects

As discussed above, we are particularly interested in estimating the second-round effects of urbanization on rural poverty. To do this we first need to disentangle the two effects and then to identify an appropriate way to control for the first round effects in the empirical analysis. This section deals with the former task. Let us assume N distinct geographical units (districts), each with population P_{it} at time t , split between urban (P_{it}^U) and rural areas (P_{it}^R), with $i \in [1, N]$. We can characterise the incidence of poverty (H_{it}^R) in rural areas in district i at time t as a function of the urban population of the district and a series of other characteristics of the district (such as its total population, specific policies, etc.), represented by the vector X :

$$H_{it}^R = f(P_{it}^U, X_{it}) + \varepsilon_{it} \quad (2.11)$$

Let us assume that natural growth rate is zero and the only changes in the rural-urban split of the population are determined by one (or both) of these two phenomena: intra-district rural-urban migration or rural areas becoming urban (either because they are encompassed by an expanding urban area or because their population has grown sufficiently to upgrade from the status of village to that of town). Define α_t as the number of rural poor divided by rural population (i.e. headcount poverty in rural areas), σ_t as the number of poor rural-urban migrants divided by the number of rural poor and λ_t as the number of rural poor in villages which become urban migrants divided by the number of rural poor. Define also γ_t as the number of rural-urban migrants divided by rural population at time t and φ_t as the

number of rural dwellers residing in villages which become urban at time t divided by rural population at time t (with $\gamma_t \geq \alpha_{t-1}\sigma_t$ and $\varphi \geq \alpha\lambda$). We can then re-write 2.11 as:

$$H_{it}^R = \frac{\overbrace{\alpha_{t-1}P_{it-1}^R}^{\text{Rural poor at } t-1} \times \overbrace{[1 - (\sigma_t + \lambda_t - \sigma_t\lambda_t)]}^{\text{Share rural poor turning urban between } t-1 \text{ and } t}}{\underbrace{P_{it-1}^R}_{\text{Rural pop at } t-1} \times \underbrace{[1 - (\gamma_t + \varphi_t - \gamma_t\varphi_t)]}_{\text{Change in rural pop between } t-1 \text{ and } t}} + g(P_{it}^U, X_{it}) + \varepsilon_{it} \quad (2.12)$$

The first term on the right hand side of 2.12 defines the first-round effects of the growth of urban population on rural poverty. Its numerator represents the number of rural poor at time t as if the change in this number (between t and t-1) were only due to the change of status of those rural poor (at t-1) becoming urban dwellers at t (through parameters σ_t and λ_t). The denominator represents the total rural population at t.

The condition under which this first-round effect decreases rural poverty incidence (*ceteris paribus*) is

$$\frac{\alpha_{t-1}P_{it-1}^R(1 - \sigma_t - \lambda_t + \sigma_t\lambda_t)}{P_{it-1}^R(1 - \gamma_t - \varphi_t + \gamma_t\varphi_t)} < \frac{\alpha_{t-1}P_{it-1}^R}{P_{it-1}^R}$$

Ignoring the terms $\sigma_t\lambda_t$ and $\gamma_t\varphi_t$ as they are likely to be very small and the subscripts to save clutter this condition becomes:

$$\sigma + \lambda > \gamma + \varphi \quad (2.13)$$

The key variables here are the poverty distributions of both rural-urban migrants and dwellers of rural-urban transitional areas relative to the poverty distribution of the rural population. Expression 2.13 states that if the distribution of migrants is skewed towards

low income individuals – i.e, the incidence of poverty is higher among migrants than non migrants – and if the poverty incidence in rural villages that become urban is higher than that in total rural population of the district then rural-urban migration will directly reduce rural poverty. Recent cross-country evidence by Ravallion et al (2007) seem to be consistent with the validity of condition 2.13. They find a sizeable negative effect of urbanisation on the incidence of rural poverty and concomitantly an increase in the number of urban poor with urbanisation. Although they cannot isolate the direct effects of rural-urban migration, their findings would be hard to reconcile without condition 2.13 holding. Although there is no evidence establishing empirically the relative size of the parameters in 2.13, some studies find that those rural areas on the outskirts of (large) urban areas may benefit economically from this vicinity (e.g. Fafchamps and Shilpi, 2003 for Nepal). This may imply lower levels of rural poverty in those peri-urban areas about to be incorporated into urban areas, i.e. $\lambda < \varphi$. This means that the poverty incidence among rural-urban migrants needs to be substantially higher than that among rural non-migrant population for expression 2.13 to be verified, i.e. $\sigma > \gamma + (\varphi - \lambda)$.

As the main aim of this paper is to estimate the size and direction of the second-round effects of urbanization on rural poverty, we can re-arrange 2.12 to control for the direct effects of urbanisation as well as for other covariates of rural poverty:

$$H_{it}^R(P_{it}^U | \sigma_{it}, \gamma_{it}, \lambda_{it}, \varphi_{it}, X_{it}) = h(\sigma_{it}, \gamma_{it}, \lambda_{it}, \varphi_{it}) + g(P_{it}^U, X_{it}) + \varepsilon_{it} \quad (2.14)$$

This expression represents the basis of the empirical analysis described in the next section. Effectively we need to estimate the partial derivative of H_{it}^R with respect to P_{it}^U .

The channels described above should underlie the second-round effects that we are trying to capture through this partial derivative.

2.3 Empirical methods

Using a district-level analysis, we try to systematically assess whether and to what extent urbanisation in Indian districts during the 1981-1997 period has affected rural poverty in those districts. In order to evaluate the eventual effects of urbanization on the people in extreme poverty, we also use specifications of rural poverty which try to isolate changes in the intensity of poverty for the very poor.

We argue that districts are an appropriate spatial scale for such an analysis in India as all of the first and second-round channels described above are likely to display most of their effects within the district's boundaries. This is consistent with the theoretical discussion above, arguing that the effects of city growth are concentrated in surrounding rural areas. Various pieces of specific evidence on India confirm that this is likely to be the case.

First, evidence suggests that intra-district migration in India is a large component of total rural-urban migration. According to the Census (Government of India, 1991), 62% of the total stock of permanent internal migrants was intra-district in 1991, although a share of this stock was composed of women migrating for marriage reasons. However, a consistent part of internal migration in India is not captured by the Census because it does not involve change in residence. This may include various forms of temporary migration, such as seasonal and circular as well as commuting. Such migration may account for an important part of income generation and livelihoods in several rural areas (Deshingkar and

Start, 2003, and Deshingkar, 2005). Due to its temporary nature, this migration is likely to be short-distance. In a recent survey of a number of rural villages in two Indian states, Deshingkar and Start (2003) reported that in a number of villages several households were commuting daily to nearby urban locations (although this movement was not registered in the migration data) and in one village, one entire caste took up casual labouring in the urban sector. This does not deny the existence of long-distance migration in India, which in fact was increasing during the nineties (Jha, 2008). However, long distance rural-urban migration is mainly directed to a few growing metropolitan areas, such as Mumbai, Delhi, Bangalore and Chennai, which are excluded from the analysis. Notwithstanding the importance of intra-district migration, in the empirical section we also test the robustness of the results against the relative size of the intra-district migrant population.

Second, during the period of analysis (1981-1999) most perishable agricultural goods' markets do not appear to be well integrated at the national or even at the state level in India. This is due to relatively poor transport infrastructure networks and lack of appropriate technology (such as cold storage facilities). Agricultural produce is often sold in nearby towns and even most trade in livestock tends to occur at a short distance. This is due to lack of infrastructure, which brings livestock marketing costs to distant markets up to 20-30 percent of the sale price (Chandra Mohan Reddy, 2000). As a result, most transactions in live animals take place within the same district (Birthal, 2005). Thus we would expect a consistent share of agricultural trade to occur at a small distance, making districts a suitable spatial scale to capture a substantial part of the first two channels above as well. In line with these ideas, some studies have performed district level analyses to try to capture

demand-side effects on agriculture. Parthasarathy Rao et al. (2004) for instance analyse the effects of urbanisation on agricultural diversification into high value commodities, such as fruit, vegetables, dairy products, using districts as the unit of analysis.

There is also emerging evidence of increases in land prices in peri-urban and rural areas surrounding urban agglomerates. Land values in those areas may be well above the discounted future stream of income from agricultural activity, inducing several landowners to sell the land (Jha, 2008).

The core idea of the empirical analysis is to assess the effects of urbanization on rural poverty at the district level over time. For that we estimate equation 2.14 trying to control for the direct effects of urbanisation as well as for other determinants of rural poverty. We use the basic specification:

$$H_{dt}^R = \beta_0 + \gamma_d + \beta_1 P_{dt-j}^U + \beta_2 [(\sigma_{dt} + \lambda_{it})/(\gamma_{dt} + \varphi_{it})] + \chi X_{dt} + \varepsilon_{dt} \quad (2.15)$$

where H_{dt}^R is a measure of rural poverty in district d at time t , γ is district fixed effects, P_{dt}^U is the urban population of district d at time $t-j$ (where $j \in [0, 2]$), $[(\sigma_{dt} + \lambda_{it})/(\gamma_{dt} + \varphi_{it})]$ is a term capturing the direct effects of urbanization on rural poverty, i.e. the term $H(\sigma_{it}, \gamma_{it}, \lambda_{it}, \varphi_{it})$ in 2.14, and X is a vector of controls, which include other variables likely to have independent impact on rural poverty. The district's urban population is computed as $\sum P_{it}$, where P_i is the population of town i in district d at time $t-j$ (where $j \in [0, 2]$) and N_d is the number of cities in district d . Given the above discussions, we would expect $\beta_1 < 0$ and $\beta_2 < 0$.

2.3.1 Data and variables

Data to run specification 2.15 comes from three main sources: district level measures of poverty are available from various rounds of the Indian household survey data (National Sample Surveys), which have been appropriately adjusted by Topalova (2005) for the 1983-84, 1987-88, 1993-94 and 1999-2000 rounds of the NSS. Other district level data, such as population composition come from the Indian districts database at the University of Maryland (which has been extrapolated from the original data in the Indian Census). Data on town populations are available from various rounds of the Indian Census. In addition, for crop production volumes and values we use the district level database for India available with International Crops Research Institute for semi-Arid Tropics (ICRISAT) from 1980 to 1994 and recently updated by Parthasarathy Rao et al (2004) up to 1998.

The district classification has been modified during the period of analysis, as some districts have been split into two units. Topalova (2005) created a consistent classification by aggregating the 2001 districts originated from the splitting into the district division of 1987. We conform to this re-aggregation and modify the original population and demographic data accordingly.

Dependent variables

We use two standard Foster Greer Thorbecke (FGT) measures of poverty as dependent variables: the poverty headcount ratio and the poverty gap index, which we define in Appendix A. Both measures are increasing in poverty, i.e. a higher value means a higher level of poverty.

Population variables

The Census 1991 (and 2001) classifies towns as all the statutory places with a municipality, corporation, cantonment board or notified town area committee, or, alternatively, places satisfying simultaneously the following three criteria: i) a minimum population of 5000; ii) at least 75 per cent of male working population engaged in non-agricultural pursuits; and iii) a density of population of at least 400 per sq. Km. This is consistent with the classification of the 1981 Census, except for condition iii), which required a minimum population density of 1000 per sq. Km. The year effects should anyway control for eventual problems of statistical consistency of urban data over time. The NSS uses the Census definition to classify urban vs. rural areas, thus ensuring the consistency of data across sources.

There were 5179 towns that met these criteria in 2001. We calculated the total urban population at the district level, by summing the figures for towns. Due to its peculiar nature, we excluded from the dataset the State of Delhi and the districts of the other megalopolises, Calcutta, Chennai, Bangalore and Mumbai; we also excluded three other districts due to an extraordinary increase in urban population in the period under study, which is extremely likely to be imputable to errors in the data: Anantapur in Andhra Pradesh, Kanniyakumari in Tamil Nadu, and Thane in Maharashtra.

As population data are available only with a ten-year frequency (1971, 1981, etc.), we estimate the values for the year 1997 by non linear interpolation in order to conduct the analysis for three rounds of the NSS. We first estimate the yearly growth rate in the period 1991-1997, calculating a weighted average of the growth rate of the 1981-1991 and

1991-2001 periods; we then calculate the 1997 population applying the estimated growth rate to the 1991 level. In this way we try to reduce the potential endogeneity of the urban population to rural poverty interpolated only using the 1991-2001 growth rate. The main results are also robust to using interpolated 1997 data based only on the 1991-2001 growth rate (results available upon request).

There are 431 districts in Topalova's (2005) original dataset, 409 of which have a positive urban population (at least for one of the three time periods); total population figures are available for only 363 of these, therefore constituting our main sample of analysis; in the year 2001, this sample accounts for a total population of 1,000,053,152 of which 270,153,691 are urban residents, corresponding to 97% and 94% of the Indian total respectively.

Controls

Following the discussion in section 2.2, we would need data on the poverty profile of rural urban-migrants (σ_{dt}) and of dwellers of areas which are rural at t-1 and become urban at time t in order to properly estimate β_2 in expression 2.15, i.e. the direct effects of urbanization on rural poverty. Unfortunately this data is not available, thus we proxy for it by including variables measuring the extent to which migrants (and dwellers of rural areas turning into urban areas) are over- or under-represented among the poor (σ_t) relative to the whole rural population (γ_t). We use two types of such variables.

The first is the district's urban poverty rate H_{dt}^U . To see why, let us re-express H_{dt}^U on the basis of the variables in question. Consider that H_{dt}^U depends on urban poverty at t-1,

on the share of rural-urban migrants at time t whose income in the urban sector is below the urban poverty line and on the change in the poverty rate of previous urban dwellers between t and $t-1$. Dropping the subscript d to save clutter, we have:

$$H_t^U(\pi_t, P_{t-1}^R, \gamma_t, \sigma_t) = \frac{\overbrace{\psi_{t-1} P_{t-1}^U}^{\text{Urban poor at } t-1} + \overbrace{\rho_1(\pi_t)(\gamma_t - \alpha_{t-1}\sigma_t)P_{t-1}^R}^{\text{Non poor rur-urb migrants becoming urban poor between } t \text{ and } t-1} + \overbrace{\rho_2(\pi_t)\alpha_{t-1}\sigma_t P_{t-1}^R}^{\text{Poor rur-urb migrants becoming urban poor between } t \text{ and } t-1} + \overbrace{\Delta\psi_t(\pi_t)P_{t-1}^U}^{\text{Change in poverty of existing urban stock between } t \text{ and } t-1}}{\underbrace{P_{t-1}^U + \gamma_t P_{t-1}^R}_{\text{Urban population at time } t}} \quad (2.16)$$

where ψ_{t-1} is the urban poverty rate at time $t-1$, ρ_1 and ρ_2 are respectively the share of non-poor rural migrant ($\gamma_t - \alpha_{t-1}\sigma_t$) at time t and the share of poor rural migrants $\alpha_{t-1}\sigma_t$ at time t who have become urban poor at time t (both are a function of urbanisation rate at time t , π_t); $\Delta\psi_t$ is the change in poverty rate (between $t-1$ and t) of the existing stock of urban population at $t-1$. From this expression it follows that $\rho_1 \leq \rho_2$ and $\partial\rho_1/\partial\pi_t < 0$, $\partial\rho_2 \leq \partial\pi_t < 0$. For any values of π_t we can compute the condition for which $H_t^U < H_{t-1}^U$ (i.e. a reduction in the urban poverty rate between $t-1$ and t) as:

$$z(\sigma, \gamma|\pi_t) = \alpha\sigma(\rho_1 - \rho_2) + \gamma(\psi - \rho_1) > \Delta\psi P_{t-1}^U (P_{t-1}^R)^{-1} \quad (2.17)$$

with $\partial z/\partial\sigma \leq 0$ (as $\rho_1 \leq \rho_2$) and $\partial z/\partial\gamma < 0$ if $\psi \leq \rho_1$.

Equation 2.16 implies that for any given value of urban economic growth at time t , urban poverty is more likely to have decreased between t and $t-1$ the lower the share of rural poor that migrated to the urban areas during this period (σ_t). This is explained by the fact that the probability of poor rural-urban migrants becoming urban poor (after migrating) is higher than the same probability for non-poor rural-urban migrants. On the other hand a

smaller rural-urban migrant population will decrease urban poverty only if the incidence of poverty in this population, once it becomes urban, is larger than the pre-existing incidence of poverty in the urban area ($\psi \leq \rho_1$). Condition 2.17 therefore implies that the evolution of urban poverty over time should capture the evolution of the parameters γ and σ at time t for any given value of π_t . This means that at any given time urban poverty should capture the combined effect of economic growth and of the direct effects of urbanisation on rural poverty (the term $h(\sigma_{it}, \gamma_{it})$ in 2.14).

We also control for the first-round effects of urbanization on rural poverty through the socio-demographic composition of the rural population (i.e. age and literacy). Again, this is an indirect form of control and is probably less effective than the share of urban poor in capturing first-round effects. The rationale behind it relies on the assumption that the income distribution of migrants can be expressed as a function of the migrants' age composition. Other things being equal, poverty incidence tends to be lower among young adults (i.e. 15-34), as they represent the most productive age class. Therefore the higher the share of young adults in the total migrant population (relative to their share in the rural population) the lower the probability that urbanisation will directly reduce rural poverty. Rewriting expression 2.13 (without considering rural areas becoming urban for ease of exposition) we have: $\frac{\sigma}{\lambda}(\lambda_{15-34}) > 1$, with $\partial \frac{\sigma}{\lambda} / \partial (\lambda_{15-34}) < 0$, where λ_{15-34} is the share of people aged 15-34 in total migrants relative to their share in the rural population. The same argument can be applied to literate migrants. As we do not observe the composition of the migrants' population, we can only control for it indirectly through the composition of the actual rural population. This is based on the plausible assumption that the change

in the number of young adults in the rural population is inversely related to the change in their number in the rural-urban migrant population in the same period.

This assumption is supported by the results of regressing the 1981-91 change in the urban population in the 15-34 age group on the change in the rural population in the same age group (controlling for changes in district's total population and total population in 1981):

$$\Delta P_{15-34}^U = -4954 - 1.038\Delta P_{15-34}^R + 0.2554\Delta P^{tot} + 0.0123P_{t-10}^{tot}$$

(2.57) (29.44) (38.71) (11.93)

N=334 R² = 0.97 (robust t-statistics in parenthesis)

The coefficient of ΔP_{15-34}^R is not statistically different from -1 indicating that changes in the rural population are reflected in mirror changes in the urban population (through either rural-urban migration or rural-to-urban change in status of villages).

Obviously, the incidence of young adults (as well as literates) in the rural population also directly and positively affects rural income and thus has a direct impact on the poverty rate. Therefore this variable will capture two contrasting effects on rural poverty: a first-order poverty reducing effect and a second-order poverty increasing effect (which should capture part of the direct effect of urbanisation on rural poverty). It should be clear that the ability to control for first round effects of these two types of variables (urban poverty rate and socio-demographic characteristics) is only residual to their direct relationship with rural poverty. Thus they are not likely to fully control for the first round effects of urbanisation on rural poverty. However, to the extent that they can control for at least part of

these effects, the direction of change in the urban population coefficient after the inclusion of these variables should provide an idea of the likely intensity of first-round effects.

Aside from the controls of first-round effects, we need to control for any other determinants of rural poverty. We use two variables which should control for the composition of the rural population: the number of people in the age group 15-34, and the proportion of literates in this age group. The latter variable is meant to capture the level of literacy of the most productive part of the population, following the idea that the most powerful influence of education on income and poverty is through its labour market effect. We also include in some specifications the share of rural population which is reported as scheduled castes and scheduled tribes, as this is expected to have an independent (adverse) effect on poverty.

However it is likely that other unobserved factors affect the relationship under scrutiny. We exploit the panel dimension of our dataset to deal with that. First, we include district fixed effects, which absorb any time-invariant component at the district level, such as geographical position, climatic factors, natural resources, etc. Second, we add a whole set of state-year dummies, which control for state-specific time-variant shocks (including economic dynamics and policies). The inclusion of these controls may still not completely account for three other sources of potential bias in the coefficient of interest (capturing the second-order effects of urbanization on rural poverty in 2.15).

First, there may be unobserved time varying district-specific shocks which may affect both rural poverty and urban population. For example there may be a localised shock (e.g. the election of an effective district government) which spurs district's economic growth. As economic growth is generally associated with urbanisation, this may foster urbanization

while reducing rural poverty at the same time. This omitted variable problem would imply a spurious negative association between the two variables. Data on income per capita at the district level is not available to us. However, as economic growth directly affects urban poverty (as described above) the inclusion of the urban poverty rate in the controls should minimise this problem.

Second, unobserved time varying rural specific shocks may affect urbanisation via increases in agricultural productivity. This view is supported by a long-standing argument in development economics that a country's urbanisation (and industrialisation) process is fuelled by increasing agricultural productivity (e.g. Nurske, 1953). In closed economies an expanding urban population requires increases in productivity of the rural sector in order to be sustained. However, Matsuyama (1992) shows that in open economies this need not be the case, as they may rely on agricultural import for their subsistence (as in the case of the East Asian newly industrialised economies). In our case, districts can safely be considered as small open economies (within India), trading across borders in most agricultural markets. Thus this potential source of bias may not be very relevant in the analysis. In line with this Fafchamps and Shilpi (2003) do not find that agricultural productivity of nearby rural areas is an important determinant of city size in Nepal. To be on the safe side, we also control for a measure of agricultural productivity. The variable is constructed as the sum of the total quantities of 22 different crops produced in a given district, multiplied by the average India-wide price of the respective crop in the same year and divided by the district's rural population. We use an India-wide price instead of district specific prices to minimise both the data gaps (which are several for the latter) and the potential endogeneity of districts'

prices to rural poverty. This is in some way an extra control because it may eat up some of the effects of urbanization on rural poverty, which may occur via its effects on agricultural productivity (see channel two above).

Instrumental variable

Finally, there may be a problem of reverse causation to the extent that rural poverty may drive rural-urban migration. It could either act as a push factor (i.e. poorer people migrate in search of an escape out of poverty) or, in the presence of high fixed costs of migration, it may act as a restraint to migration. If the former case prevails (i.e. poverty is mainly a push factor), the coefficient β_1 in 2.15 would have a downward bias; while the opposite is true if the latter effect of poverty on migration dominates. The findings by Ravallion et al (2007) that global rural-urban migration has been associated with large reduction in the number of rural poor lends some credit to the importance of the former case. Kochar (2004) also provides indirect support to this hypothesis, showing that in India landless households have the highest incidence of rural-urban migrants among rural households.

Regardless of the direction of the bias, we need an additional variable to act as a valid instrument, i.e. it must be correlated with district urban population, but must also be exogenous to poverty-induced rural-urban migration flows. A variable which satisfies both prerequisites is the number of people who migrate to urban areas of the district from states other than the one where the district is located. It is plausible to assume that rural poverty in a given district has no effect on migration decisions in other states, which typically do

not share the same rural condition of the district in question. On the other hand, the number of migrants coming to district towns from other states is part of the urban population of the district, thus bearing a positive association with our main explanatory variable.

A concern about the exogeneity of the instrument may arise from the fact that, within a given district, both migration to the cities and rural poverty are likely to be affected by the underlying, unobserved economic trend. However, the first stage of the IV estimation includes all the controls listed in the OLS specification, and particularly the rate of urban poverty, the measure of agricultural productivity, and the interaction of time and States' fixed effects. We argue that these variables would absorb most of the economic trend in the district, thus limiting the potential bias originating from the instrument endogeneity.

Although measurement error is not likely to be a major cause of concern in our analysis, it is worth noticing that the IV estimation may also correct eventual biases arising from errors in the measurement of urban population. This is the case if the measurement error of the instrument and that of the instrumented variable are independent.

2.3.2 Results

Table 2.10 presents the descriptive statistics for the main variables used in the analysis while table 3 presents the results from regression 2.15 a using OLS estimation. Our dataset includes observations of 363 districts for three different time periods: 1983, 1993, and 1999. We run 2.15 applying a two years lag to the measure of urban population and to the other demographic controls for two main reasons. First, in this way we reduce the risk of potential simultaneity bias. Second, the two-year lag allows us to minimise the use of inter-

polation for obtaining the Census variables (both population and socio-demographic variables), which are recorded in 1981, 1991 and 2001. We also include district and state-year fixed effects in all specifications. Standard errors are robust to heteroscedasticity (using the Huber-White correction) and allow for intra-group correlation within individual observations.

Table 2.10. Descriptive statistics of the main variables, 1981-99

	Obs	Mean	Std. Dev.	Min	Max
Rural poverty (share)	1,170	0.321	0.183	0.004	0.81
Poverty gap index, rural	1,170	0.076	0.061	0	0.315
Rural 15-34 age (share)	1,003	0.247	0.025	0.2	0.326
Rural literates 15-34 age (share in 15-34)	1,003	0.485	0.179	0.107	0.997
Rural poors (abs. nr)	1,000	567,725	485,956	320	4,127,495
Rural population	1,003	1,668,426	982,274	15,078	8,247,888
Scheduled caste (share)	1,001	0.177	0.084	0	0.545
Agr. productivity	793	0.216	0.266	0	3.261
Urb. migr. from other states	1,007	31,098	54,077	0	545,521
Urban population	1,200	436,497	550,895	0	4,526,745
Urban poverty (share)	1,131	0.255	0.178	0	0.701

1981-1999 period

We run a number of different specifications in Table 2.11, testing the robustness of the results to the inclusion of a number of controls and the use of different dependent variables. When controlling only for rural population (as well as for the range of fixed effects described above), the result indicates that the growth of urban population exerts a

highly significant poverty reducing effect on rural areas (column 1). This result is robust to the inclusion of socio-demographic controls for the rural population, including the share of scheduled caste, the share of young adults (15-34 age group) in the rural population and the share of literates in the young-adults rural population (column 2). These last two variables are meant to capture a change in the composition of the rural population and therefore should partly absorb the first round effects of urbanization on rural poverty. The inclusion of these controls slightly decreases the urban population coefficient. The signs of the controls are as expected, except for the share of literates: a higher share of young adults decreases poverty, while a higher presence of scheduled caste increases it (although not significantly). This suggests that the direct effect on poverty of the young adult population prevails over their indirect effect which captures the rural-urban migration of young adults. The share of literates has a poverty-increasing, albeit not significant, effect. At a closer inspection, this unexpected effect of literacy is driven by its Post-1993 impact. As shown in column 3, the coefficient of this variable turns negative (but not significant) when we account for the significant poverty increasing impact of literacy in the post-1993 period. In this period a higher incidence of literates in the most productive part of the rural labour force was associated with higher levels of rural poverty. Understanding the rationale of such an unexpected result is beyond the scope of our analysis, but we will suggest a possible reason for this below.

Accounting for this differential impact determines also an increase in the urban population coefficient, as its effect is probably estimated with more precision. This coefficient is slightly above that of column 1, suggesting that rural socio-demographics may be cap-

turing some first-round impact of urbanisation, which in this case increases rural poverty. As discussed above, this would be the case if a high level of urbanization was fuelled by high intra-district migration rates. Considering that young adults are over-represented in the migrant population, and that this is the most productive (and thus least poor) part of the population, there may a positive association between urbanization and poverty via this type of first-round effects. The rest of the direct effects of urbanization on rural poverty should be captured by the inclusion of urban poverty rate as a control. This is significantly and positively correlated with rural poverty (column 4). As urban poverty captures both the effects of district's economic growth (πt) on rural poverty and the direct effects of urbanisation on rural poverty, this suggests that the former are larger than the latter. The inclusion of urban poverty reduces the absolute size of the urban population coefficient, confirming that the rural poor tend to be over-represented in the migrant population. However this reduction is very mild: the coefficient goes from -0.066 to -0.062 (column 3 to column 4). Following the discussion in the preceding section, we interpret this as a clear indication that most of the effect of urbanization on rural poverty is given by "second-round" mechanisms.

The magnitude of the effects of urban population on rural poverty over the 1981-1999 period is not particularly strong although it is robust. An increase in the district's urban population of 200,000 (a 43% increase from the mean value) reduces on average the poverty rate by 1 to 1.4 percentage points according to the specifications. Given that the average share of rural poverty over the period considered is 32%, this effect ranges between 3.2% and 4.2% of the mean poverty rate.

Results using the poverty gap index as the dependent variable are less robust than those using the poverty rate (columns 5 and 6). Urban population exerts a negative effect on the poverty gap with the other controls keeping the same sign as in the preceding regressions. This result appears to be driven by the effects of urbanisation on those poor who are relatively close to the poverty line. When the rural poverty share is included among the explanatory variables, the urban population has a positive albeit not significant effect on the poverty gap (column 6), which suggests that the poor closer to the poverty line are those who benefit most from urbanisation. This category does not include those poor far behind the poverty line. In the absence of more precise data, we could only speculate about why this may be the case. The effects of urbanisation are not likely to concern the very poor much. For example, the increase in demand for agricultural goods may affect those involved in commercial agriculture, specifically those who own capital and/or certain skills not usually available to the very poor. The same can be said about rural-urban migration: the very poor may not have enough capital to cover the fixed costs of migration. For these reasons urbanisation seems to have a fairly neutral effect on the very poor rural dwellers. Interestingly, the presence of rural dwellers from the scheduled caste is negatively associated with severe poverty. Along with the results from the preceding regressions, this suggests that the scheduled caste population tends to be concentrated among the rural poor close to the poverty line, but not among those in severe poverty.

We also test for the effects of urbanisation on the number of rural poor (column 7), obtaining similar results. For every increase in urban population by 100 people the rural population in poverty decreases by 13 people. The other controls are in numbers rather than

Table 2.11. The effects of urbanization on rural poverty across Indian districts, 1981-1999

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Rural pov. (share)	Rural pov. (share)	Rural pov. (share)	Rural pov. (share)	Poverty gap	Poverty gap	Rural poor (millions)
Urban pop. (millions)	-0.0616*** (0.0220)	-0.0522** (0.0206)	-0.0655*** (0.0206)	-0.0615*** (0.0218)	-0.0157** (0.00776)	0.00192 (0.00388)	-0.1220** (0.0517)
Rural pop. (millions)	-0.0126 (0.0163)	-0.0192 (0.0160)	-0.0110 (0.0162)	-0.00758 (0.0149)	-0.00193 (0.00511)	0.000250 (0.00220)	0.9739*** (0.2193)
Scheduled caste (share)		0.194 (0.284)	0.0686 (0.278)	0.314 (0.299)	-0.0417 (0.116)	-0.132** (0.0583)	0.9605 (0.6086)
Rural pop 15- 34 age (share)		-2.920*** (0.770)	-3.881*** (0.825)	-4.103*** (0.826)	-1.330*** (0.271)	-0.151 (0.120)	
Rural lit 15-34 (% in 15-34)		0.0450 (0.179)	-0.112 (0.172)	-0.122 (0.167)	-0.0203 (0.0566)	0.0147 (0.0217)	
Rural lit 15_34 x Post-1993			0.237*** (0.0680)	0.215*** (0.0656)	0.0807*** (0.0200)	0.0189** (0.00821)	
Urban poverty (share)				0.326*** (0.0616)		0.287*** (0.00831)	
Rural poverty (share)					0.106*** (0.0210)	0.0122 (0.00855)	0.3987*** (0.1098)
Rural pop 15- 34 age (mln)							-0.2127* (0.1166)
Rural literates 15-34 (mln)							-0.1706* (0.09710)
Observations	997	996	996	964	964	964	964
No. of districts	363	363	363	354	354	354	354
R-sq. (within)	0.65	0.65	0.66	0.68	0.757	0.949	0.582

Note: All specifications include district and state-year fixed effects. Robust standard errors (Huber-White method) in parentheses; *significant at 10%; **significant at 5%; ***significant at 1%; all explanatory variables are lagged two years except for Agricultural Productivity (1 year lag) and urban poverty (contemporaneous).

in shares (except for scheduled caste). Following the discussion in section 2.4, this represents a different way of controlling for the first round effect of urbanisation on rural poverty. In this way, the urban population variable may capture some of the effects of changes in the remaining rural population (net of the young adult population). The controls maintain the same sign as in the previous regressions, except for the rural population, which is now positive and significant and literates in the 15-34 year group, which is now negative and significant. The former result is expected as, other things being equal, a larger rural population is associated with more rural poor. The latter captures the direct association between literacy and poverty, which is negative. This may differ from the preceding regressions using shares because those may capture second-order effects of literacy on poverty.

1981-1993 period

We now examine the impact of urbanisation on rural poverty using only the first two time periods available, covering the time interval 1981-1993. This is a robustness check for our results with three time periods, as in this case no interpolation of urban population is needed. It is also an interesting analysis focusing only on the pre-liberalisation period. Overall, the effect of urbanisation on rural poverty is stronger than over the entire period (Table 2.12). The coefficient for the urban population ranges between -0.08 (column 1) and -0.11 (column 3) depending on the specification; this is almost twice as large as the range reported in Table 2.11. An increase in the district's urban population of 200,000 reduces on average the poverty rate by between 1.6 and 2.2% of total rural population. The basic specification without controls (except for the fixed and year effects) confirms the negative

relationship between urbanisation and rural poverty, although it is only mildly significant (column 1). The inclusion of socio-demographic controls increases the significance and the size of the coefficient, again confirming that some adverse first-round impacts of urbanisation on rural poverty are taken away by these controls (column 2). Both the share of young adults in the rural population and the share of literates in the young adult population exert a poverty-reducing impact. This supports the hypothesis of a differential impact of literacy on rural poverty over time, i.e. poverty-reducing up to 1993 and then poverty-increasing. The results are robust to the addition of the share of urban poverty (column 3). However, this time the magnitude of the coefficient of urban population increases from 0.099 (column 2) to 0.111 (column 3). This increase suggests that the first-round effects of urbanisation on rural poverty captured by urban poverty may have been poverty-increasing in the eighties. Again this is a very small change, confirming that second-round effects are likely to dominate first-round ones. The impact of urbanisation on the poverty gap index is negative but less significant than for the entire period (column 4), while the impact on severe poverty seems to be neutral again (column 5). Finally, the results also hold when using the number of rural poor as a dependent variable (column 6). Again, the elasticity of poverty reduction is much higher than that considered in the 1981-1999 period.

Further robustness

To control for the possible endogeneity due to the potential effects of agricultural productivity on urbanisation, we add a measure of agricultural productivity to the list of controls. This variable is lagged one year, given that the simultaneity bias should not be

Table 2.12. The effects of urbanization on rural poverty across Indian districts, 1981-1993, OLS

	(1)	(2)	(3)	(4)	(5)	(6)
	Rural pov. (share)	Rural pov. (share)	Rural pov. (share)	Poverty gap	Poverty gap	Rural poor (millions)
Urban pop. (millions)	-0.0791 (0.0592)	-0.0928* (0.0553)	-0.111** (0.0549)	-0.0265 (0.0168)	0.00549 (0.00809)	-0.2814** (0.1114)
Rural pop. (millions)	0.0061 (0.0221)	0.0082 (0.0316)	0.0047 (0.0022)	-0.0028 (0.0082)	-0.0015 (0.0066)	0.1661*** (0.0435)
Scheduled caste (share)		0.0691 (0.398)	0.383 (0.505)	-0.00927 (0.207)(0.	-0.120 114)	0.8171 (1.0103)
Rural pop 15-34 age (share)		-4.619*** (1.306)	-5.313*** (1.408)	-1.739*** (0.473)(0.	-0.207 224)	
Rur. literates (share in 15-34)		-0.700*** (0.216)	-0.835*** (0.255)	-0.179** (0.0845)	0.0620 (0.0408)	
Urban poverty (share)			0.378*** (0.106)	0.140*** (0.0396)	0.0310 (0.0233)	0.4839*** (0.1867)
Rural poverty (share)					0.288*** (0.0116)	
Rural pop 15-34 age (mln)						-0.1202 (0.2152)
Rural lit. 15_34 age (mln)						-0.4245*** (0.1119)
Observations	682	682	659	659	659	659
No. of districts	363	363	354	354	354	354
R-sq. (within)	0.611	0.640	0.660	0.763	0.940	0.589

Note: All specifications include district and state-year fixed effects. Robust standard errors (Huber-White method) in parentheses; *significant at 10%; **significant at 5%; ***significant at 1%; all explanatory variables are lagged two years except for Agricultural Productivity (1 year lag) and urban poverty (contemporaneous).

an issue in this case (but a contemporaneous specification is not possible due to the lack of data for 1999). The main results reported in Table 2.13 appear to be robust to the inclusion of such a measure. Surprisingly, the urban population coefficient for the entire period increases (column 1). However, this effect is mainly due to the restricted sample for which agricultural data is available. When we run the same regression as in Table 2.11 column 4 with the same sample as in Table 2.13 column 1, the increase in the size of the urban coefficient disappears (column 2). To the extent that part of the poverty-reducing effects of urbanisation may operate through increases in agricultural productivity (see section 2 above), the unchanged urbanisation coefficient is a somewhat puzzling result. The key to explain this may be the surprisingly weak (negative) effect of agricultural productivity on rural poverty (column 2). If this is the case, then the effects of urbanisation via productivity increases would be fairly insignificant as well. In fact, when restricting the analysis to the 1981-93 period, the coefficient of agricultural productivity becomes negative (as expected) and the magnitude of the urbanisation impact on rural poverty decreases slightly, although it maintains its significance (column 3 vs. column 4). This suggests that agricultural productivity may have had a different impact on rural poverty in the post-1993 period. Column 5 confirms such a hypothesis, as the post-1993 effect of productivity appears to have been robustly adverse to rural poverty.

Such a surprising finding may be in contradiction with earlier literature on India, which shows the key effect of higher farm yield in poverty reduction only until 1994 (Datt and Ravallion, 1998). Investigating the reasons behind this adverse post-1993 impact is beyond the scope of our analysis, and we only speculate about a possible explanation for

it. This may lie in the (negative) effect of agricultural productivity on rural employment in the non-farm tradable sector (e.g. rural industry). Foster and Rosenzweig (2004) find this pattern for Indian villages and explain it through the negative incentives that agricultural productivity growth provides to capital in the non-farm tradable sector through higher wages. To the extent that non-farm growth is especially pro-poor (as rural industry tends to productively employ the main asset of poor rural households, i.e. low-skilled labour), this negative effect on non-farm growth may dampen that of agricultural productivity growth on rural poverty. This effect may have been particularly strong in the post-liberalisation period (i.e. post-1991), when labour was freer to move in search for lower-wage locations (see Aghion et al., 2008). Incidentally, the same argument may also help explain the adverse impact of literacy on rural poverty in the nineties. Since literate labour has a higher reservation wage than illiterate labour, a high share of literate labour may have acted as a restraint to investments by the non-farm tradable sector.

We already mentioned that to the extent that rural-urban migration occurs across districts, the identification strategy may not enable us to properly capture the channels linking urbanisation to rural poverty. In order to control for this, we need to construct a variable that measures the weight of rural-urban intra-district migration in the total rural emigrant population. By connecting this variable to the urban population, we may control for the fact that the effects of urbanisation on rural poverty are better identified in those districts with a relatively higher share of internal rural-urban migration in total rural emigrants. However, the data available does not allow us to compute such a share; we instead compute a rough approximation of this measure by dividing intra-district rural-urban migration by rural pop-

ulation. Including the interaction between this variable and the urban population leaves the results unaffected (column 6) with the interaction term bearing an expected but insignificant negative coefficient. We also use a different variable, i.e. the ratio of intra-district rural-urban migrants over the urban immigrants from other districts, obtaining similar (negative and non significant) results (not shown here). The lack of significance of these results may be due to the imprecise measure of the importance of intra-district migration.

We also test for the importance of the backward linkage effects of urbanisation on poverty. Considering that urban agricultural demand affects the district's rural sector more intensely in less spatially integrated markets, we need information on the share of urban demand of perishable products in total urban demand. Since we do not have this information, we instead compute a rough approximation based on agricultural data: the share of land cultivated fruits and vegetables (proxy for perishable goods) in total land cultivated. This measure relies on a number of assumptions, i.e. that a district's supply is a good proxy for urban demand and that fruits and vegetables are the main perishable agricultural goods. The interaction term between this share and the urban population variable has an expected negative coefficient (i.e. the higher the share the more poverty-reducing the urbanisation impact) – column 7. Again, this is not significant probably due to the imprecision of the measure. Also, including this interaction term reduces the explanatory power and the significance of the urbanisation variable. This may be due to the high collinearity between the two variables generated by the small variation of the fruit and vegetable share over time.

Given that limiting the spatial extent of the effect of urbanization within the border of single districts may be questionable, we run the same specifications of tables 2.11 - 2.13

Table 2.13. The effects of urbanization on rural poverty across Indian districts, Further robustness

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	1981-99		1981-93		1981-99		
	Rural pov. (share)	Rural pov. (share)	Rural pov. (share)	Rural pov. (share)	Rural pov. (share)	Rural pov. (share)	Rural pov. (share)
Urban pop. (millions)	-0.0684** (0.027)	-0.0678** (0.026)	-0.153** (0.063)	-0.158** (0.065)	-0.074*** (0.027)	-0.075*** (0.027)	-0.065* (0.039)
Rur. pop. (millions)	-0.0137 (0.019)	-0.00989 (0.018)	-0.0131 (0.026)	0.00411 (0.024)	-0.00946 (0.019)	-0.00992 (0.019)	-0.00392 (0.021)
Scheduled caste (share)	0.486 (0.34)	0.488 (0.34)	0.738 (0.56)	0.701 (0.57)	0.555 (0.34)	0.540 (0.35)	0.625* (0.35)
Rural pop 15-34 age (share)	-4.628*** (0.97)	-4.690*** (0.99)	-5.445*** (1.47)	-5.716*** (1.54)	-5.024*** (0.98)	-5.039*** (0.98)	-4.764*** (1.02)
Rural lit. 15_34 age (% in 15-34)	-0.0896 (0.21)	-0.0969 (0.21)	-1.067*** (0.28)	-1.004*** (0.28)	-0.135 (0.21)	-0.133 (0.21)	-0.0631 (0.22)
Rural literates 15_34 x Post- 1993	0.215*** (0.074)	0.218*** (0.074)			0.231*** (0.075)	0.233*** (0.075)	0.227*** (0.077)
Urban poverty (share)	0.327*** (0.074)	0.328*** (0.073)	0.355*** (0.12)	0.380*** (0.11)	0.329*** (0.072)	0.331*** (0.072)	0.371*** (0.072)
Ln Agricultural productivity	-0.0167 (0.020)		-0.0613** (0.030)		-0.0274 (0.020)	-0.0268 (0.020)	-0.0260 (0.021)
Ln Agr. prod. x Post-1993					0.0429*** (0.015)	0.0431*** (0.015)	0.0397** (0.016)
Share Internal migrants						-0.285 (0.58)	
Urban pop x Share fruits and vegetables							0.0201 (0.13)
Observations	753	753	519	519	753	753	707
Nr. of districts	275	275	275	275	275	275	253
R-sq. (within)	0.67	0.67	0.65	0.65	0.67	0.67	0.64

Note: All specifications include district and state-year fixed effects. Robust standard errors (Huber-White method) in parentheses; *significant at 10%; **significant at 5%; ***significant at 1%; all explanatory variables are lagged two years except for Agricultural Productivity (1 year lag) and urban poverty (contemporaneous).

adding a spatially lagged urbanization variable, i.e., the average of the urban population of the contiguous districts. We also try to include the spatial lag of total population. These variables however were never significant, while other coefficients were only minimally affected (Table 2.14, first column).

Finally, a further bias may be due to small villages upgrading to towns in the census definition. To the extent that these growing villages are systematically located in rural areas where poverty is decreasing (increasing) for reasons independent of urbanisation, we may detect a negative (positive) effect of urban population on poverty share which would be spurious. We therefore re-estimate the models excluding from the urban population variable towns with less than 20,000 inhabitants – i.e., the size category which would contain most of the ‘upgraded villages’. Results of this regression are extremely similar, although slightly less precise (see Table 6, second column). In the last two columns in Table 6 we run the same regressions as in the first two but employing IV estimation (using the number of migrants from other states to the urban areas of the district as an instrument). Again neither the spatial lagged variable nor the ‘small villages’ issue seem to affect the main IV results either (as discussed below).

IV estimation

Although the results are neat, we still need to control for the direction of causality in the relationship between urbanisation and rural poverty. As rural poverty declines, the rural-urban migration rate and thus urbanization may slow down as well and vice-versa. This would provide a source of (downward) bias in the coefficient. Without properly con-

Table 2.14. The effects of urbanization on rural poverty across Indian districts, Further robustness

Sample	(1)	(2)	(3)	(4)
	All	Cities >20k	All	Cities >20k
	Rural pov. (share)	Rural pov. (share)	Rural pov. (share)	Rural pov. (share)
Urban pop. (millions)	-0.0496** (0.0222)	-0.0365 (0.0231)	-0.108*** (0.0377)	-0.112*** (0.0408)
Urban pop. of bordering districts (millions)	1.67e-07 (5.79e-07)		2.96e-07 (5.97e-07)	
Rural pop. (millions)	-0.0155 (0.0146)	-0.00851 (0.0145)	-0.0132 (0.0146)	-0.00348 (0.0146)
Urban poverty (share)	0.326*** (0.0637)	0.326*** (0.0626)	0.322*** (0.0629)	0.323*** (0.0621)
Scheduled caste (share)	0.474 (0.301)	0.372 (0.293)	0.483 (0.301)	0.325 (0.298)
Rural literates 15_34 age (share in 15-34)	-3.329*** (0.769)	-3.181*** (0.820)	-3.262*** (0.739)	-3.039*** (0.787)
Rural literates 15_34 age (share in 15-34)	0.0253 (0.162)	-0.118 (0.162)	0.00369 (0.159)	-0.147 (0.160)
Observations	953	952	914	901
R-squared (within)	0.678	0.682		
Number of districts	343	354	306	305
Method	OLS	OLS	IV	IV

Note: all specifications include district and state-year fixed effects. Robust standard errors (Huber-White method) in parentheses; *significant at 10%; **significant at 5%; ***significant at 1%; urban population is instrumented through the number of urban immigrants from other states.

trolling for this potential endogeneity, the coefficient of equation (5) may have a downward bias, which means the estimates in Table 2.11 may be lower in absolute value than the real ones.

We employ IV estimation to deal with this problem, using the number of migrants from other states to the urban areas of the district as an instrument. The first stage regressions, reported in different specifications in Table 2.15, substantiate the strong correlation of the instrument with the instrumented variable, and F- statistics are well above the confidence threshold of Stock and Yogo (2005) test for weak instruments (Table 8-9, last row). In analogy with OLS, standard errors in the IV estimations are robust and allow for intra-group correlation at district level.

Results from the second stage regressions confirm the suspect of a downward bias of the OLS parameters, with new estimates being roughly twice as large as the OLS estimation for the period 1981-1999 (Table 2.16). This in turn implies a fairly substantial impact of urbanisation on rural poverty, with the rural poor decreasing by between 2% and 3% of districts' rural populations as the effect of an increase by 200,000 in urban residents (columns 1-3).

The IV analysis confirms the small first-round relative to second-round effects of urbanisation on rural poverty (column 1 to 2). Again, the results are robust when agricultural productivity variables are included as a control (column 3). We also run the IV estimation using the poverty gap as the dependent variable (column 4). The change in the magnitude of the urban population coefficient compared to the OLS specification in Table 3 is even

Table 2.15. The effects of urbanization on rural poverty across Indian districts, 1983-1999, IV Estimation, first stage

	(1)	(2)	(3)	(4)	(5)	(6)
	1981-99			1981-93		
	Urban pop.	Urban pop.	Urban pop.	Urban pop.	Urban pop.	Urban pop.
Urb. migrants from other states	4.248*** (0.82)	4.177*** (0.91)	4.177*** (0.91)	4.095*** (0.72)	3.814*** (0.72)	3.814*** (0.72)
Rural pop.	0.0558 (0.035)	0.0891** (0.035)	0.0891** (0.035)	0.0320 (0.031)	0.0598* (0.030)	0.0598* (0.030)
Scheduled caste (share)	-129121 (566798)	5459 (742202)	5459 (742202)	470989 (518338)	869828 (786918)	869828 (786918)
Rural pop 15-34 age (share)	3838 (1183220)	-59321 (1316907)	-59321 (1316907)	-1291961 (1152661)	-1125466 (1290325)	-1125466 (1290325)
Rural literates 15-34	-439969** (200881)	-354376 (241649)	-354376 (241649)	-141172 (176400)	-94391 (237122)	-94391 (237122)
Rural literates 15-34 x Post-1993	178712** (70735)	195241** (80831)	195241** (80831)			
Urban poverty (share)		-88634* (52376)	-88634* (52376)		29978 (94211)	29978 (94211)
Ln Agr. productivity		64776 (77841)	64776 (77841)		53597 (72036)	53597 (72036)
Ln Agr. prod. x Post-1993		806.3 (32739)	806.3 (32739)			
Observations	996	779	779	682	520	520
Number of districts	363	280	280	363	275	275
R-squared	0.72	0.74	0.74	0.67	0.68	0.68
F-stat	61.83	30.97	30.97	15.17	40.16	40.16

Note: all specifications include district and state-year fixed effects. Robust standard errors (Huber-White method) in parentheses; *significant at 10%; **significant at 5%; ***significant at 1%

Table 2.16. The effects of urbanization on rural poverty across Indian districts, 1983-1999, IV Estimation

	(1)	(2)	(3)	(4)	(5)	(6)
	Rural pov. (share)	Rural pov. (share)	Rural pov. (share)	Poverty gap	Poverty gap	Rural poor (millions)
Urban pop. (millions)	-0.112*** (0.033)	-0.117*** (0.034)	-0.139*** (0.031)	-0.0393*** (0.012)	0.00105 (0.0052)	-0.1624** (0.0656)
Rural pop. (millions)	-0.00770 (0.016)	-0.00427 (0.015)	0.000761 (0.017)	0.00204 (0.0059)	0.00182 (0.0025)	1.4754*** (0.2544)
Scheduled caste (share)	0.0646 (0.27)	0.292 (0.30)	0.406 (0.32)	-0.0281- (0.12)	0.146** (0.058)	1.0945* (0.6497)
Rural pop 15-34 age (share)	-3.845*** (0.79)	-4.057*** (0.79)	-4.808*** (0.88)	-1.573*** (0.30)	-0.172 (0.14)	
Rural literates 15_34 age (share in 15-34)	-0.139 (0.17)	-0.153 (0.16)	-0.263 (0.20)	-0.0702 (0.069)	0.00652 (0.027)	
Rural literates 15_34 x Post-1993	0.249*** (0.067)	0.230*** (0.064)	0.281*** (0.070)	0.105*** (0.022)	0.0231** (0.0094)	
Urban poverty (share)		0.323*** (0.061)	0.338*** (0.067)	0.116*** (0.023)	0.0173* (0.0095)	0.4001*** (0.1124)
Ln Agr. productivity			-0.128 (0.078)	-0.0236 (0.021)	0.0136 (0.012)	-0.4115*** (0.1462)
Ln Agr. prod. x Post- 1993			0.165*** (0.062)	0.0482*** (0.017)	-0.0000217 (0.0085)	0.3731*** (0.1137)
Rural poverty (share)					0.291*** (0.0090)	
Rural pop 15-34 age (millions)						-0.3695*** (0.1036)
Rural literates 15_34 age (millions)						-0.3068*** (0.1019)
Rural lit. 15_34 age (millions) x post-93						9.128*** (2.925)
Observations	950	914	753	753	753	753
Number of districts	319	306	255	255	255	255
R-squared	0.04	0.11	0.13	0.14	0.82	0.31
Kleibergen- Paark Wald F statistic	27.089	26.068	21.018	21.018	20.861	20.849

Note: all specifications include district and state-year fixed effects. Robust standard errors (Huber-White method) in parentheses; *significant at 10%; **significant at 5%; ***significant at 1%; urban population is instrumented through the number of urban immigrants from other states.

bigger, and it maintains its significance. Again, when the share of rural poor is included as a control, the coefficient of urban population becomes insignificant (column 5). This confirms that urbanisation does not have an independent effect on the poverty gap, and thus on the severity of poverty, other than through the effect induced by the decrease in the share of poor in the rural population. The increase in magnitude of the coefficient is confirmed even when using the absolute number of rural poor as a dependent variable (column 6), although in this case the coefficient is only 1.5 larger than in the OLS (cf. Table 3, column 7).

We also run the same regressions for the period 1981-93, obtaining similar results (Table 2.17). The coefficient of urban population is magnified by a factor of between 3 and 5 relative to its OLS value (cf. Table 2.12, columns 1-3), although it is estimated fairly imprecisely in the specifications with few control variables (column 1 and 2). The same increase in size is also true for the specification using the poverty gap as a dependent variable (column 4). However the inclusion of the share of rural poor as a control eliminates any effect of the urban population on poverty gap (column 5). This is also the case for the estimation run with the number of rural poor as a dependent variable: the increase of the urban coefficient is 4-fold. The robustness checks examined in the previous section, including the spatially lagged variable and the population of towns with more than 20,000 inhabitants, do not affect our results when applied to the IV setting (Table 2.13, columns 3 and 4).

Finally, the substantial downward bias of the OLS estimates implied by the IV results suggests that an increase in poverty may be an important push factor for rural-urban migra-

tion. This could indicate that the poverty incidence is higher among migrants than among non-migrants (thus $\sigma > \gamma$). At the same time, our results suggest that first-round effects are quite small, i.e. condition $[\sigma > \gamma + (\varphi - \lambda)]$ does not hold in its strong form. This would imply, consistently with the discussion in section 2.2, that the poverty incidence is lower in rural areas that are about to become urban than in the other rural areas (thus $\lambda < \varphi$), and interestingly this difference is similar to that of poverty rates between rural-urban migrants and rural non-migrants, i.e. $[(\sigma - \gamma) \simeq (\varphi - \lambda)]$. Obviously the evidence provided here is not strong enough to make this more than an interesting speculation. And further research would be necessary to provide more direct empirical testing of such a hypothesis.

2.4 Conclusions

Do the poor in rural areas benefit from population growth of urban areas? And if so, what is the size of the benefits? Answers to these questions could help clarify whether trade-offs exist between urban investment and rural poverty and may help shed new light on the old debate on urban bias in developing countries. Notwithstanding the importance of these questions, little empirical evidence is available to provide adequate answers. We have tried to address this gap, by analysing the effects of urbanization on rural poverty. Using data on Indian districts between 1981 and 1999, we find that urbanization has a significantly poverty reducing effect on surrounding rural areas. Results are robust to the inclusion of a number of controls and to the use of different types of specification. The findings suggest that most of the poverty reducing impact of urbanization occurs through second-round ef-

Table 2.17. The effects of urbanization on rural poverty across Indian districts, 1981-1993, IV Estimation

	(1)	(2)	(3)	(4)	(5)	(6)
	Rural pov. (share)	Rural pov. (share)	Rural pov. (share)	Poverty gap	Poverty gap	Rural poor (millions)
Urban pop. (millions)	-0.268 (0.20)	-0.315 (0.20)	-0.506** (0.21)	-0.147** (0.058)	0.00143 (0.0157)	-0.8431** (0.3710)
Rural pop. (millions)	0.00030 (0.0024)	0.00111 (0.0024)	0.00248 (0.0031)	0.000471 (0.0011)	-0.00256 (0.00432)	0.7259 (0.5135)
Scheduled caste (share)	0.174 (0.44)	0.556 (0.56)	0.877 (0.67)	0.133 (0.26)	-0.125 (0.122)	1.1043 (1.3225)
Rural pop 15-34 age (share)	-4.754*** (1.31)	-5.535*** (1.41)	-5.628*** (1.52)	-1.889*** (0.55)	-0.237 (0.259)	
Rural literates 15_34 age (share in 15-34)	-0.738*** (0.21)	-0.867*** (0.25)	-1.073*** (0.29)	-0.257** (0.10)	0.0582 (0.0460)	
Urban poverty (share)		0.390*** (0.11)	0.400*** (0.12)	0.164*** (0.046)	0.0465* (0.0260)	0.5081** (0.2006)
Ln Agr. productivity			-0.0984 (0.078)	-0.0163 (0.024)	0.0126 (0.0134)	-0.3782** (0.1494)
Rural poverty (share)					0.294*** (0.0132)	
Rural pop 15-34 age (millions)						0.7514 (2.418)
Rural literates 15_34 age (millions)						-0.4973*** (0.1141)
Observations	636	608	488	488	488	488
Number of districts	318	304	244	244	244	244
R-squared	0.06	0.10	0.04	0.03	0.823	0.306
Kleibergen- Paark Wald F statistic	31.941	32.260	27.910	27.910	20.861	20.939

Note: all specifications include district and state-year fixed effects. Robust standard errors (Huber-White method) in parentheses; *significant at 10%; **significant at 5%; ***significant at 1%; urban population is instrumented through the number of urban immigrants from other states.

fects rather than through the direct movement of rural poor to urban areas. We resort to IV estimation to test for causality. The results suggest that the effect is causal (from urbanisation to poverty reduction), and that failure to control for causality bias the coefficient of urbanisation downwardly. In our preferred estimations, we find that an increase of urban population by one fifth determines a decrease of between 3 and 6 percentage points in the share of rural poverty. These poverty reducing effects appear to apply mostly to rural poor relatively closer to the poverty line. Although the very poor do not seem to be negatively affected by urbanization, they are not able to reap the benefits of such a growth.

These findings may have a number of potentially important policy implications. First, they may help re-consider the role of public investment in urban areas for poverty reduction. In fact it is a popular tenet that investments in developing countries need to be concentrated in rural areas in order to reduce poverty, as the poor in developing countries are mainly concentrated there (see for instance World Bank, 2008). However, investments in rural areas are often very onerous as substantial resources are needed to reach a population which is scattered around vast territories. To the extent that urbanization may have substantial poverty reducing effects on rural areas, urban investments may become an important complement to rural ones in poverty reduction strategies.

Second, our findings run counter to the popular myth that rural-urban migration may deplete rural areas causing them to fall further behind. The relatively low rate of urbanisation of India itself may also be due to public policies which have not facilitated (and in certain instance even constrained) rural-urban migration (Deshingkar and Start, 2005).

At the very least, this paper questions the appropriateness of this bias against rural-urban migration.

Third, to the extent that the benefits from urbanisation do not spill over to the very poor in rural areas, specific actions may be needed to facilitate these rural dwellers to enjoy the benefits of urbanisation. Examples of these may include developing the types of skills useful for an expanding urban sector; or the provision of capital to cover the fixed costs of rural-urban migration.

Although this paper has not touched upon the issue of urban poverty, rising urban populations may imply that urban poverty could become in the future the main issue in its own right (Ravallion et al., 2007). Further research is needed to assess whether the growth of urban population entails a trade-off between rural and urban poverty reduction.

2.A Methodological note to the construction of poverty measures

The poverty headcount ratio and the poverty gap index are two standard Foster Greer Thorbecke (FGT) measures of poverty. FGT poverty measure for a given population is defined as:

$$H_{\alpha}^i = \int_0^{z_i} \left(\frac{z_i - y}{z_i} \right)^{\alpha} f(y) dy \quad (2.18)$$

where z_i is the poverty line in the area i (with $i = [\text{rural, urban}]$), and $f(y)$ is the distribution function of monthly per capita expenditure (in this case), with the population ordered in ascending order of y (i.e. starting from the poorest).

The **Headcount Ratio** is computed by setting $\alpha=0$, thus it represents the proportion of the population below the poverty line in a certain geographical unit (poverty rate). The poverty lines used by the dataset are those recommended by the Planning Commission (1993) and are as follows. The rural poverty line is given by a per capita monthly expenditure of Rs. 49 at October 1973-June 1974 all-India rural prices. The urban poverty line is given by a per capita monthly expenditure of Rs. 57 at October 1973-June 1974 all-India urban prices (see Datt (1995) for further details on the rural and urban cost of living indices and the estimation of poverty measures).

The **Poverty Gap Index** is computed by setting $\alpha=1$ and is defined as the mean distance below the poverty line as a proportion of the poverty line where the mean is taken over the whole population, counting the non-poor as having zero poverty gap. That is the mean shortfall from the poverty line (counting the non poor as having zero shortfall), expressed as a percentage of the poverty line.

Chapter 3

Stars and Comets: an Exploration of the Patent Universe

3.1 Introduction

The analysis of patent and citation data has become a key source of evidence on localized knowledge spillovers and innovation. Nevertheless, one aspect has been generally overlooked: the patent distribution across inventors is extremely skewed, as many inventors register one or a few patents, while a small number of inventors register many patents. Innovations developed by inventors at the opposite extremes of the distribution are unlikely to be the outcome of an homogeneous innovation “black box”. Interestingly, this peculiar characteristic of the patenting activity recalls the more general “innovation paradox” highlighted in the innovation literature (e.g. Acs and Audretsch, 1990): while big companies massively invest in formal R&D activities, many new products and processes are generated by small and medium firms, with little or no reported investments in R&D. The latter kind of innovation process is therefore more likely to be based on learning-by-doing and informal innovation, being thus intrinsically different from the activity of “professional scientists”.

To our knowledge, none of the previous empirical literature on “local innovation” based on patent data has discussed the different “innovation scales” patents may originate from. A first contribution of this paper is therefore to document the issue. A second con-

tribution is to investigate whether patents originating from different categories of inventors are located in different cities. A third contribution - which constitutes the main aim of the paper - is to test whether the concentration of the activity of star inventors is beneficial to the local productivity of more occasional, and less prolific, inventors.

In order to achieve that, using the USPTO/NBER database we identify two illustrative categories of inventors situated in the tails of the distribution: we define as *stars* those inventors who are highly productive in a time window of 8 years - while we define as *comets* those inventors that develop only one or two patents in same time window. A preliminary data inspection at MSA level shows how the association with establishment births and other MSA structural characteristics and number of patents is significantly different for the two patent categories. This confirms that the categorization is not trivial, and suggests that i) the two categories may relate to different innovation processes, and ii) stars and comets are concentrated in different cities, especially after controlling for the general distribution of the patenting activity.

The location of investments of big companies is increasingly influenced, directly or indirectly, by local policy makers: the attraction of "million dollar plants" is seen as a successful policy targeted at increasing the productivity of incumbent (small) firms through technological spillovers (Greenstone et al, 2008). Similarly, local policy makers may be keen to attract R&D labs of big companies within their jurisdiction. Our results do not seem to support the effectiveness of these policies: we find some evidence suggesting that the direct impact of stars on the local economy is negligible; however, the lack of direct effects might be compensated by indirect effects operating through an increase of the activity

of comet inventors, which in turn may justify the provision of public money to place-marketing policies.

Therefore, in the second part of our empirical analysis we assess whether the activity of star inventors is beneficial to the production of comet patents, and try to quantify this effect. More specifically, using the NBER/USPTO patent database we estimate a model where the number of comet patents produced in a given city, time period, and technological category is a function of the number of star patents developed in the same city, period, and category. We exploit the panel dimension of our dataset to account for various fixed effects, and adopt an instrumental variable approach to avoid a potential endogeneity bias. In our preferred estimation, we find that, on average, 10% more patents developed by star inventors lead to 2-3% more patents authored by comet inventors.

3.2 Patents, localized knowledge spillovers, and the size of innovation

Patent data have become extremely popular in the economic literature in the last two decades, as they represent an easy and accessible way to proxy for an economic activity which is generally very hard to measure, i.e., innovation. Furthermore, the availability of citation linkages has added even more interest in patents data: for the first time, researchers had a tool to "trace" knowledge spillovers, which previously had been considered as one of the most intangible concepts in economic theory. A popular book by Jaffe and Trajtenberg (2005), and the free availability of the USPTO dataset from the NBER website, further contributed to multiply the empirical applications based on patent data.

A significant part of this literature has focused on the geographic component of innovation, with a particular interest in the spatial decay of knowledge spillovers. A seminal contribution by Jaffe et al (1993) showed that a cited-citing patent couple is twice as likely to be in the same US metropolitan area than a couple of technologically similar patents with no citation links. Similarly, Peri (2005) examined the flows of citations among 147 European and US regions to find that "only 20% of average knowledge is learned outside the average region of origin", and Jaffe (1989) demonstrated that academic research has large effects on the number of private patents developed in the same US state. Finally, Carlino et al (2007) used patent data for a cross-section of US metropolitan areas to investigate the relationship between urban density and innovation intensity (as measured by patents per capita) finding a positive and robust association, with the caveat that many omitted variables might explain the positive correlation.²⁶ All these contributions (and many similar which we omit for brevity) highlight that knowledge spillover have a geographically limited distance decay.

It is also important to stress that the nature and causes of knowledge spillovers are still debated. For instance, Breschi and Lissoni (2009), building on previous contributions by Breschi and Lissoni (2001), Zucker et al (1998), and Almeida and Kogut (1999), highlighted how defining localized knowledge spillovers as an *externality* can be misleading, as most of the knowledge diffusion may take place through market interactions - namely the spatially-bounded mobility of inventors among workplaces - rather than through in-

²⁶ The authors include a robustness test based on IV estimation, but, in our opinion, the exogeneity of all the instruments is questionable, as they may affect patenting through e.g. productivity. Also, it is not very clear how sorting of very productive inventors and companies into denser cities may influence their results.

formal contacts. Using data on US inventors' application to the European Patent Office, they were able to show that after controlling for inventors' labour mobility and the related professional network, the role of proximity in explaining knowledge diffusion is greatly reduced.

Previous contributions, however, did not take into consideration an important feature of patent data, i.e., the skewness of the distribution of patents across inventors.²⁷ This is in part due to the fact that until very recently a unique identifier for inventors was not available in the NBER/USPTO database and therefore calculating the distribution of patents by inventors was infeasible. Thanks to the efforts of Trajtenberg et al (2006), who "estimated" a unique inventor identifier using an ad-hoc algorithm,²⁸ we know that out of 1,600,000 inventors listed in the NBER dataset in the period 1975-99, 60% of them registered just one patent, 30% from 2 to 5, and only 0,15% (2,402 inventors) more than 50 patents.

The peculiar distribution of patents by inventors reveals that the innovation process which patenting is a proxy for is an extremely composite phenomenon. On one side, a large number of patents is developed by "comets", i.e., individuals who apply for a patent only once or twice over a long period. On the other side, a small group of "stars" develop individually a huge number of patents. This, beyond being an interesting fact per se, poses

²⁷ Among the closest contributions we could find, we mention: Silverberg and Verspagen (2007), who analysed in depth the skewness of the distribution of citations across patents; Zucker and Darby (2007) looked at the linkages with private companies of a small sample of star inventors.

²⁸ The authors needed to face two orders of problems: first, the same author may appear in the database with different names due to spelling errors; second, different authors may have the same name (the "John Smith problem"). The complex algorithm they developed exploits all the available accessory information (dates, locations, technological fields, etc.), together with word sound matching routines. The validity of the procedure is confirmed by a test on a dataset of Israeli inventors.

a number of questions related to the geography of innovation: do different categories of inventors interact with the local economic environment in the same way? Do they respond similarly to the same location determinants? Are they equally distributed over space or do they tend to concentrate? Is spatial proximity beneficial for their activity?

These questions are related to the growing interest in peer effects in science and in the spillovers originating from star scientists. Among the most interesting recent contributions, Azoulay et al. (2008) exploit the exogenous variation in the number of "superstar scientists" in US university originated by the sudden death of these individuals to estimate the loss in productivity of their collaborators. They find an average 5-10% decline in their average publication rates, starting 3-4 years after the superstars' death and enduring over time, but no differential effect for co-located collaborators. Waldinger (2009) estimates the effect of the dismissal of scientists from Germany Universities during Nazism. Similarly to Azoulay et al., he finds a strong effect on coauthors (13-18%), but no significant effects at department level. Therefore, both the studies challenge the existence of localized positive spillovers originating from stars in academic environments.

Equally on the "skeptical" side, there are the advocates of the "death of distance" theory, who argue for a decreasing importance of the role of spatial proximity following the progress of communication technologies (e.g., Friedman, 2005; Quah, 1999; Cairncross, 1997). On the other side, other economists argue that the technological progress has actually increased the scope for proximity for innovative activities due to the higher importance of face-to-face contacts and agglomeration externalities (e.g. Coyle, 1999). The few empirical assessments of the issue seem to support the "death of distance" hypothesis (Griffith

et al, 2007; Ioannides et al, 2008), indeed suggesting that localized knowledge spillovers are fading over time.

Turning to industrial innovation literature, the skewed patent distribution recalls the well-known difference between the innovative activity of small and big companies. In particular, robust evidence on two distinct aspects of small firm innovation poses a challenging "innovation paradox". First, small firms have a much higher ratio of patents developed to R&D expenditures (Griliches, 1990) than big companies. If we substitute patents with innovations introduced to the market and R&D with employment, the result is equivalent: the ratio is much higher for small firms (Acs and Audretsch, 1990). The authors argue that this can be due to the higher permeability of small companies to local public R&D inputs (e.g., university research) (Acs et al, 1992). An alternative explanation could be that small companies rely on alternative innovation inputs, based on learning-by-doing and applied innovation, rather than formal scientific research. Second, small firm innovation is all but a residual phenomenon, accounting for most of the innovative activity in many sectors (Acs and Audretsch, 1990). In passing, it is also worth mentioning that small firms account for most of the employment growth in the US in the last decades (Audretsch, 2002). Furthermore, Balasubramanian and Sivadasan (2008) in a recent working paper link patent data with Census firm data for the US, being able to assess the impact of patents on firm performance. They focus in particular on firms that patent for the first time, and find a significant and large effect of the first patent on firm growth (but, interestingly, little change in factor productivity). This would suggest that "occasional" patents have a relevant market value, although further research based on patent-firm matched datasets is needed to explore the is-

sue. As we cannot access this kind of data, in this paper we focus only on patents and their inventors; the "innovation paradox" could be an interesting way to generalize the results whenever patent-firm matched data will be accessible to all the interested researchers.

3.3 Stars and Comets

Our analysis is based on the NBER/USPTO database, which lists all the patents granted in the United States from 1969 to 1999. We added to this dataset the inventors' unique ID developed by Trajtenberg et al (2006). As the latter is available only since the 1975, our period of analysis is restricted accordingly. More details on the data, including the geocoding process, are reported in Appendix A.

At a first glance, the abundance of data makes a micro analysis at inventor level the most appealing alternative. A deeper view of the data, however, suggests that this is unfeasible, in light of the simple fact that the dataset is about patents, not inventors, which implies that individual inventors are observed only when they patent. When an inventor is not patenting, we do not know their location, their possible employer (i.e., the assignee of their patents), etc. The problem would be perhaps negligible if we focused only on very productive inventors; but given we are interested also in comets, the issue is crucial.

We therefore opt for an analysis at city level, focusing on the number of *patents* produced by each group of inventors, rather than on the number of *inventors* themselves. Ideally, this would require that, for every time interval, we knew how many comet patents, star patents, and other patents are developed in a given locality. However, the data we use are rather imprecise in the time dimension, for the following reasons: first, we use the

year when the patent is granted,²⁹ which is generally 2-3 years later than the year of application. Second, we do not know how long an inventor has been working on a patent before applying for it. Equally difficult is to time when local knowledge spillovers may have effect - it could be while the source and destination inventors are both working on their respective patents, but it could equally happen a few years after the star has applied (or has been granted) for it. By inspecting the data we found that the median and mean value of the citation lag of patents in the same MSA is four years, and we therefore choose to adopt periods of the same length.³⁰ This seems a reasonable choice in order to "average out" some of the measurement error in the temporal dimension. We thus identify five time periods of four years each, which are listed in table 3.18.

We then need to identify those inventors which we define as stars or comets. The task necessarily entails a degree of arbitrariness, which makes our quantification of the number of star and comet patents relatively noisy. However, the estimations we present in the paper (namely in section 4) are robust to measurement errors,³¹ and we also check whether our results are consistent with other variable definitions, finding very little variation. We describe these alternative specifications and results in Appendix B. Therefore, although we of course aim for the most precise definition, the reader should not be excessively worried

²⁹ The reason why we use the grant year, rather than the application one, is to avoid the bias given by data truncation. More precisely, using the application year we would automatically exclude all the patents not granted (but applied for) before the 1999, as they are not included in the dataset. This subsample could easily be non-random, e.g. better patents may take longer to be examined, etc.

³⁰ We restricted the calculation to patent couples with a maximum citation lag of ten years, as longer lags are unlikely to be related to knowledge spillovers. The citation lag is calculated as the difference between the grant year of the citing and cited patents.

³¹ The number of star and comet patents are used as dependent and independent variables, respectively. In the first case, the measurement error does not affect the consistency of the estimates; in the second case, we rely on 2SLS estimates to obtain consistent coefficients.

about the exact definition: we just need to define two good proxies of the quantity of star and comet patents in a given city, technological category, and period.

Potentially, we could observe inventors for their whole career, and then classify them as stars or comets according to their propensity to patent. There are, however, two problems, one conceptual and one due to data truncation. First, to the extent that we aim at assessing the effect of productivity spillovers, a definition of stars based on their whole career can be imprecise, as productivity may be highly variable along it. Second, given that our data cover the 1975-99 period, we cannot observe the whole career of the large majority of the inventors in the sample. We therefore adopt a definition that takes into account the productivity of inventors for a shorter period of time, but still long enough to approximate the average productivity of individual inventors in that stage of their career, and to smooth short term disturbances. We follow the same approach for comets as well, in order to avoid including in the category inventors who do not satisfy the requirements in the years immediately before, or after, a given period.

Table 3.18. Period classification

Period	Years	Obs. window
1	1978-1981	1976-1983
2	1982-1985	1980-1987
3	1986-1989	1984-1991
4	1990-1993	1988-1995
5	1994-1997	1992-1999

Therefore, for each of the five periods, we define an 8-years long, overlapping *observational window* - they are reported in the third column of table 3.18. In each period, a patent is defined as the outcome of a “star inventor” if its first author has developed five

other patents or more (as first author) in the relative observational window, and it is therefore defined as a star patent. The threshold has been chosen as it approximately limits the top 5% of the inventors' distribution in term of patents per-capita. Similarly, we define "comet inventors" patent (first) authors who developed less than three patents in the relative observational window, and less than six till that point in time (the latter condition excludes the possibility that a star becomes a comet); the patents they develop are defined as comet patents. As a further restriction, comet patents must not have as assignee a company which is assignee of 50 patents or more in the whole dataset, in order to avoid defining as comets those inventors working for companies where many stars are potentially employed. The threshold has been chosen because 80% of star patents are assigned to an assignee which has more than 50 patents assigned. This restriction is important for our analysis, for two reasons: first, it allows us to better identify local knowledge externalities, disentangling them from co-located increases in productivity due to market mediated workplace contacts. The recent literature has indeed highlighted the risk to overestimate the positive effects of externalities by ignoring the "priced" component of the professional network of inventors, as we discussed in the previous section (e.g. Breschi and Lissoni, 2009; Zucker et al, 1998; and Almeida and Kogut, 1999). Second, our definition of comets entails inventors working for firms for which the primary activity is not the production of patented innovation. Without a patent-firm matched dataset this is hard to detect precisely, but the restriction is our best approximation. Furthermore, in order to focus on patents with a direct market application, a comet patent must be assigned to an US corporation: this leaves out around 10% of comets which are unassigned, or assigned to individuals. These

latter restrictions are instead unnecessary for stars, as they are satisfied in the large majority of the cases and, in the few cases in which they are not satisfied, it is likely to be due to spelling errors in the assignee name. A summary of the definition requirements for stars and comets are reported in 3.19.

Table 3.19. Definition requirements

Inventor group	Stars	Comets
Number of patents in the relative obs window	≥ 5	≤ 2
Total number of patents of the assignee		≤ 50
Total number of patents granted to the inventor till that point in time		≤ 5
Kind of assignee		US corporation

The analysis is generally limited to the last three periods, as MSA controls are unavailable for period 1 and 2. We define five periods, however, as the first two are used to build the instrumental variables.

Star patents account for the 26% of the total patents granted in the period 1986 - 1997, while the corresponding share of comet patents is equal to 11%. On the inventors' side, among all the unique inventors listed in the five periods (534,120), around 5% of them are listed as stars at least once, while for comets the same share is equal to 15%. Looking at single periods, star inventors are 7-9% of the total, while comets are 14-16%. It is worth noticing, therefore, that the majority of patents and inventors do not belong to the two categories. The "star" status appears to be quite persistent across time: around 40% of stars in given period were stars also in the previous period. The share goes down to 15% with a two periods lag. Individual inventors listed as stars cannot become comets in following periods by construction, while a comet can potentially become a star; this, however, happens for only 1% of comet inventors listed in the dataset.

Interesting facts emerge also from the analysis of citation data. Table 3.20 reports the flows of citations across groups, expressed as a share of the total citations originating from each group. Compared to patents that are neither comets nor stars (third row), comets (first row) are more likely to cite comets, and less likely to cite stars. The opposite is true for stars: they are more likely to cite stars, and less likely to cite comets. The pattern is similar also when looking at citations within technological categories (not shown). We interpret this as further evidence that the stars/comets categorization, although stylized and somehow arbitrary, do identify different groups of patents. On the other hand, we notice that comets do cite stars, although at a smaller rate than other patents; this in turn suggests that comets might benefit from knowledge spillovers from stars. We will explore this hypothesis in depth in the rest of the paper.

Table 3.20. citations' shares, comets and stars

		Cited		
		Comets	Stars	Other patents
Citing	Comets	16.2	16.8	67.0
	Stars	7.5	34.7	57.8
	Other patents	9.7	19.8	70.5

Citations may also be useful to inspect the average "value" of different categories of patents. Although quite debatable and noisy, the association of number of received citations with the market value of the patents has been convincingly argued (Hall et al, 2001). We use citation data to explore whether patents and comets significantly differ from other patents in this dimension, by regressing the number of received citations on "comet" and "star" dummies, over the whole sample of patents in period 3, 4, and 5. We also include time and technological category dummies, and a variable reporting the number of

Table 3.21. Regression of citations received

Dep. var.	Citations received (standardized)
Nr. citations made	0.00763*** (0.00017)
Star patent dummy	0.176*** (0.0042)
Comet patent dummy	0.0974*** (0.0051)
Other patent dummy	0.0745*** (0.0039)
Period F.E.	YES
Tech. cat. F.E.	YES
Observations	590953
R^2	0.12

Heteroskedasticity robust standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

citations made to control for the heterogeneous propensity to cite among different kinds of patents (within categories and time periods). The dependent variable is de-meaned and standardized, and thus the constant is excluded. We also run the same specification with technological *subcategory* dummies and MSA dummies, and excluding the top 5% cited patents. In both cases, we obtain very similar results (reported in Appendix D).

Results - reported in table 3.21 - show that stars are on average more cited than comets, and comets are more cited than patents which are neither stars nor comets (all the pairwise differences between the three coefficients are statistically significant). A star patent is receiving, on average, 0.87 citations more than "other patents" (0.10 time 8.7, i.e., the difference of the two coefficients multiplied by the standard deviation of the dependent variable, the number of citations received). Comets, on the other side, are receiving just around one fifth of citation more (0.02 time 8.7). Results therefore suggest that star patents have a higher scientific and market value than the average patent. However, the effect is

positive also for comet patents: this is important as it confirms that even comet patents have some scientific value (in other words, they are not just useless "garage patents" made for hobby).

3.3.1 Preliminary evidence on location of stars and comets

In this section, we present some descriptive statistics which i) show how stars and comets are located in different places, and ii) substantiate the validity of stars and comets as good proxies for the output of different innovation processes.

If we look at the distribution of comet, star, and other patents over total employment across MSAs,³² we can see that there is a sizeable correlation (Table 3.22, Figure 3.7), which implies that innovative activity is overall spatially concentrated. When plotting the shares of comets and stars on the total of patents, however, there is a fair degree of dispersion in both the distributions, driven by a long right tail (Figure 3.8, 3.9).

Table 3.22. Patents by MSAs over total employment, rank correlation

	comets	stars	other patents
comets	1	0.42	0.59
stars	0.42	1	0.61
other patents	0.59	0.61	1

We can go further by looking at patterns of partial correlation with MSAs structural characteristics, setting up a simple panel regression for periods 3-4-5 based on the following equations:

$$Share(Comets)_{it} = \beta_1 X_{it} + \delta_t + \epsilon_{it} \quad (3.19)$$

³² Counties are grouped into MSAs according to the 1993 definition, based on 1990 Census data. Counties not included into MSAs are also individually included in the sample. The analysis, therefore, covers the whole US territory.

$$Share(Stars)_{it} = \beta_1 X_{it} + \delta_t + \epsilon_{it} \quad (3.20)$$

where i indexes MSAs and t periods, X_{it} is a matrix of MSA-specific covariates, β_1 and β_2 are vectors of coefficients, and δ_t is a time fixed effect. The aim of these regressions is to assess whether stars and comets show two distinctive location patterns, depending on the industrial structure of cities. The variables included in X , therefore, are a list of simple proxies of the industrial structure of the MSA. In the detail, these variables are the following:

a) the (log of) the total patents in the MSA which are neither stars or comets, in order to control for the size of the patenting sector in the city (we excluded stars and comets to avoid circularity). We included this variable as the absolute size of the patenting sector may impact differently the production of stars and comets.

b) log of total employment (totemp), to control for agglomeration economies and size effects; we expect MSAs with larger employment to produce proportionally more patents, in line with the findings of Carlino et al. (2007), but, again, we do not have any strong a-priori on the association of city size with the different kinds of invention.

c) the share of employment in manufacturing (manuf. share), in order to assess whether comets are associated with specialization in manufacturing. To the extent that comets are linked to production phases through learning-by-doing mechanisms, this variable should also have a positive effect on the number of comets.

d) the Herfindahl diversity index (Herfindahl, calculated as the sum of the squares of the share over the total of employment of 2-digit SIC sectors), as a proxy of the diversity

of the economic structure. This variable can have two opposite effects: on one side, the literature has emphasized the positive effect of diversity on innovation due to Jacobian externalities (e.g., Glaeser et al., 1992; Duranton and Puga, 2005). On the other side, we do not exclude that MAR externalities,³³ rather than Jacobian, might be more beneficial for the kind of innovation which underlies the development of comets. In fact, to the extent that comets are the outcome of a "learning-by-doing" innovation process, we may expect them to be more frequently developed where there are within-industry knowledge spillovers, as well as other economy of scale, i.e., in specialized cities.

vi) log of the number of plants with less than 500 employees (n. plants <500 emp.) as these are defined as "small plants" in the US; to the extent that comets represent a proxy for occasional and less codified innovation, we hypothesize that their number is positively affected by the presence of small plants. Conversely, we expect star inventors to work for big companies, thus the number of star patents should be negatively associated with this variable, once controlling for total MSA employment.

The sample is restricted to the last three periods and to all the MSAs or counties where at least 100 patents have been developed in the same interval of time. The equations are estimated by OLS regressions on the pooled samples, with standard errors clustered at MSA level.³⁴ The results - reported in table 3.23 - clearly show how the two vectors of coefficient are different (as confirmed by the Hausman test: the null hypothesis of equality of the coefficients of column 1-2, and 3-4, is rejected at 1% confidence level). In particular, comet

³³ MAR (an acronym for Marshall-Arrow-Romer) externalities are those based on within-industry knowledge spillovers, and are associated with an high degree of sectoral specialisation.

³⁴ We also estimated a SUR model to account for correlation across errors in the two equations. The significance of the regressors, however, is not affected. Results are reported in Appendix D.

Table 3.23. Regression of comets/stars shares at MSA level

COEFFICIENT	(1) Comets (share)	(2) Stars (share)	(3) Comets (share)	(4) Stars (share)
Tot. emp. (log)	-0.0237*** (0.0057)	0.0116 (0.011)	0.00291 (0.0060)	-0.00365 (0.012)
Herfindahl	-0.276** (0.13)	0.672** (0.33)	-0.284** (0.13)	0.677** (0.33)
Manuf. share	0.0904* (0.046)	0.0503 (0.090)	0.0573 (0.045)	0.0694 (0.090)
N. plant <500 emp. (log)	0.0286*** (0.0064)	-0.00404 (0.013)	0.0351*** (0.0062)	-0.00776 (0.013)
Other patents (log)			-0.0412*** (0.0036)	0.0237*** (0.0082)
Period dummies	YES	YES	YES	YES
Observations	1289	1289	1289	1289
R ²	0.11	0.03	0.23	0.04

Heteroskedasticity robust standard errors clustered at MSA level in parentheses

*** p<0.01, ** p<0.05, * p<0.1

patents are positively associated with the number of small firms, while the total number of other patents and the Herfindahl index have a negative coefficient (which means that a more diversified city is associated with more comets). Conversely, star patents are positively associated with both the number of other patents and the Herfindahl index, suggesting that star patents are more frequently located in specialized cities.

Our (speculative) interpretation of these results is the following: comet patents are associated with more general innovation activities, and therefore are more likely to be located in innovative hotspots with a diversified economy and many small firms; in such cities the pool of patents is not necessarily large, as innovations may be introduced to the market in other forms. On the other hand, the activity of stars is more strongly associated with formal R&D and patenting, thus it is more frequently located where the pool of patents is large, and the structure of the local economy is specialized and dominated by big companies.

Table 3.24. Regression of establishment births at MSA level

COEFFICIENT	(1)	(2)
	Estab. births (log)	Estab. births (log)
Total comets (log)	0.304*** (0.062)	0.151*** (0.048)
Total stars (log)	-0.119*** (0.037)	-0.0818*** (0.027)
Total oth. patents (log)	0.487*** (0.072)	0.230*** (0.054)
Herfindahl Index t-1		-3.297 (2.30)
Tot. emp. t-1 (log)		0.500*** (0.058)
Manuf. share t-1		-0.473 (0.46)
N. plant <500 emp. t-1 (log)		-0.105* (0.063)
Constant	5.302*** (0.16)	4.984*** (0.18)
Period dummies	YES	YES
Observations	418	418
R^2	0.71	0.85

Heteroskedasticity robust standard errors clustered at MSA level in parentheses

*** p<0.01, ** p<0.05, * p<0.1

We also look at the association with establishment births, by regressing the latter variable on the (log of the) number of star and comet patents developed in the same MSA, plus some other controls (log of total employment, Herfindahl index, and log of average establishment employment - all lagged by one period to avoid simultaneity bias), for period 4 and 5 (period 3 is dropped due to data restrictions). The sample is composed of the 209 MSAs for which data are available, and the model is estimated by OLS on the pooled sample, with standard errors clustered at MSA level. Again, the results (table 3.24) show a differentiated pattern for stars and comets: while comets have a significant effect, comparable to the effect of other patents, star patents have a negative coefficient.

We do not claim causality at this stage - many variables are potentially omitted and we cannot exclude a reverse causality bias. Nevertheless, the associations we have analysed support two statements: first, once controlling for the general distribution of patenting activities, comet and star patents are developed in different places; second, star patents seem to have a much weaker connection with the local economy than comet patents. To the extent that the former are developed in R&D labs of big companies, while the latter are the by-product of the innovative activity of small firms, the finding is not surprising.

3.3.2 Why should stars positively affect comets?

Even though we assume comet and star patents are the outcome of substantially different innovation processes, still the activity of stars could generate positive externalities increasing the productivity of comets. We identify four main mechanisms through which the externalities may occur:

a) Informal knowledge spillovers: star inventors and comet inventors develop informal contacts due to residential proximity, which in turn facilitate the activity of the latter (e.g., they may obtain hints on their work).

b) Formal knowledge spillovers: star inventors may transfer their expertise to comet inventors in more formal ways, e.g. during seminars, conferences, and the like.

c) Workplace contacts: (future) comet inventors may have the opportunity to work in an institution where stars are employed, without necessarily becoming stars themselves (they may be employed in different duties, or they may leave the institution at an early stage of their career).

d) Display/attraction effects: the presence of many labs of big companies may attract comets to a locality, as they may expect to enjoy the effects of points a, b, and c.

Although all the mechanisms may, in theory, work also in the opposite direction (from comets to stars), we expect that the main direction of the knowledge spillovers to go from the star scientist to the "occasional" inventor. However, we understand that this may not be true *a priori* and we consequently design our empirical methodology to be robust to reverse causality.

On the other side, we mentioned earlier that a few recent contributions are downsizing the role of localized knowledge spillovers, either arguing for the weakness of local peer effects (Azoulay et al., 2008; Waldinger, 2009), or for the fading of these effects over time in the light of the "death of distance" hypothesis. Thus, the aforementioned mechanisms - and especially a, b, and c - may also play a negligible role in our context.

We therefore test whether the activity of star inventors leads to higher production of comet patents. Unfortunately, the data do not allow us to disentangle the different mechanisms (e.g., a citation may be output of a, b, or c), thus in the following analysis we will generally test for positive spillovers from stars to comets. The definition and empirical identification of the channels thorough which knowledge spillovers take place is probably one of the most challenging and interesting topics in urban economics research agenda, and we hope that the increasing availability of microgeographic data may lead to some progress in the field.

3.4 Analysis

In the present section we investigate whether the production of star patents in a city affects the production of comet patents in the same city and period, and try to quantify this effect.

We therefore estimate the following model:

$$Comets_{ikt} = \beta \cdot Stars_{ikt} + \gamma X_{it} + \delta_k + \tau_t + \phi_i + \delta\tau_{kt} + \varepsilon_{ikt} \quad (3.21)$$

where i , k , and t index MSAs, categories, and periods, respectively; Stars and Comets are the number of patents in the respective group, X is a set of MSA time-variant controls, and δ , τ , ϕ are category, time, and MSA fixed effects. The six technological categories are the following: Chemical (excluding Drugs); Computers and Communications (C&C); Drugs and Medical (D&M); Electrical and Electronics (E&E); Mechanical; and Others.

The unit of observation is the MSA-category pair; the choice is motivated by the assumption that knowledge flows in the patenting activity are mostly contained within the same technological category. This is confirmed by citation data: 80% of citation linkages are bounded within the same category. Furthermore, this allows us to exploit a useful source of variation within MSA and period. The analysis is limited to periods 3-4-5, as MSA controls are not available for previous periods, and the sample is restricted to the MSA-category pairs in which at least 25 patents have been granted in the given period.³⁵

We opt for a log-linear specification because the dependent variable is an extended count variable (with a long right tail and skewed to the left), which approximates the normal distribution after the log transformation. The side effect of the log transformation is the loss

³⁵ The restriction is made in order to exclude small counties where only a few patents are developed, which are likely to act as outliers. This also brings the advantage of reducing drastically the number of zeros and to speed calculations. Robustness tests show that the sample selection is not affecting the results.

of the zeros, which, however, are less than 5% of observations. In the following section, we perform some robustness tests on the whole sample based on a Negative Binomial model with the natural count variable and we find compatible results.

We suppose that there are two groups of time-variant variables which may potentially affect the number of comet patents produced in a given city, technological category, and time period.

The first group of variables is specific to the patenting activity, and it includes i) the relative size of the given technological category - as the number of comets may increase because the category as a whole is growing; and ii) the total number of patents in the given city, as the number of comets in a given category may increase because the city patent sector is expanding. Omitting these two variables will introduce an important source of spurious positive correlation between the number of comets and stars, which in turn will lead to an overestimate of the main coefficient of interest.

The second group of variables relate to general city characteristics, and includes a few variables measuring the total employment and the industrial structure of the MSA. This group of variables is motivated by the findings we presented previously, namely the strong association of comet patents with a few specific MSA structural characteristics, and from theoretical insights suggesting that comets are more likely to be associated with small companies and a high share of manufacturing employment. We anticipate, however, that this group of variables is rarely significant in our regressions. This is due to the inclusion of the MSA fixed effects, which absorb most of the effect of variables with small variations across time.

In detail, the variables included in the matrix X are the following:

i) number of other patents (neither stars or comets) in the technological category, over the other patents in the other five categories (*share other patents cat.*); this variable controls for the relative size of the given technological category, and for idiosyncratic (i.e., specific to the category/city pair) productivity shocks. We expect it to be positively correlated with the number of comets.

ii) total number of patents developed in the MSA - excluding all comets to avoid circularity, and stars of the given category to avoid double counting - as a control for the size of the patenting activity (*tot. MSA patents*) in the whole city. Again, we expect a positive coefficient on this variable.

We then include four MSA-specific variables, as proxies for the industrial structure of the city. These variables are exactly the same as in equations 3.19 and 3.20:

iii) Log of total employment (*totemp*), to control for agglomeration economies and size effects.

iv) The share of employment in manufacturing (*manuf. share*).

v) The Herfindahl diversity index (*Herfindahl*, calculated as the sum of the squares of the share over the total of employment of 2-digit SIC sectors), as a proxy of the diversity of the economic structure.

vi) Log of the number of plants with less than 500 employees (*n. plants <500 emp.*).

Finally, we include a number of fixed effects, controlling for technological category and MSA time invariant factors, for time-specific shocks, and for technological category shocks. In a few specification, we include also a MSA-period fixed effect. Potentially,

we could also include a MSA-category fixed effect but in this case identification will arise only from within MSA-category pairs variation, which is too limited in the data to give significant results. Standard errors are clustered at the MSA-category pair level (i.e., at every cross-sectional unit of observation). Alternative estimates based on clustering at the State-year pairwise combination gave almost identical standard errors.

3.4.1 Instrumental Variable Estimation

Estimates of equation 3.21 can be inconsistent due to reverse causality or omitted variable biases, especially for the main variable of interest (the number of star patents). We therefore create two different instrumental variables for the number of star patents to deal with the issue. The two instruments share a similar intuition: an exogenous variation in the productivity of star inventors in a given MSA and period may arise from the interaction of two factors: i) an historical presence of inventors working in a given technological category or for given companies in that MSA, and ii) an US-wide increase of productivity of these sectors or companies in the given period. To the extent that the first factor is path-dependent and exhibits some inertia over time, it is exogenous to contemporaneous MSA-specific factors once MSA fixed effects are introduced in the specification. At the same time, we expect the productivity of stars inventors working in the same subcategories or companies (but in different cities) to be correlated, due to sharing a similar competition pressure, regulatory framework, market demand, etc. Therefore, we presuppose that US-wide productivity shift in a given sector or company will translate into MSA-specific productivity shocks in proportion to the number of inventors working in that sector or company in the given MSA.

For example, we assume that the total number of star patents developed in the MSA of New York in the year 1994-97 entails an exogenous component due to the interaction of

- a) the historical presence in New York of many R&D labs in semiconductor devices, and
- b) the US-wide growth in (patent) productivity of the semiconductor devices sector in the period 1994-97, relatively to other sectors.

The IV strategy is close in spirit to the approach of Bartik (1991) and Blanchard and Katz (1992), among others, who instrumented regional economic growth interacting the lagged sectoral structure of a region with the contemporaneous national sectoral trend. In what follows, the construction of the instruments is explained in detail.

First instrument

The first instrument is calculated through the following steps:

- a) For each period, we calculate the total number of star inventors active in a given MSA and technological subcategory (patents are classified into 6 categories and 36 subcategories). If an inventor developed patents classified into different subcategories, he/she is assigned corresponding weights summing to one, accordingly to the subcategories' shares. If they have been recorded as resident in several MSAs, the modal one is chosen.
- b) For each period, each subcategory, and each MSA, we calculate the average number of patents produced by star inventors in the whole US, excluding the given MSA.
- c) For each MSA, each period, and each subcategory, we multiply the number of inventors in period $n-2$ at point a) by the productivity in the respective technological

subcategories in period n calculated in b). Subsequently, we sum the outcome by MSA, period, and technological category. The result is the instrumental variable for total number of star patents in period 3-4-5, by MSA and category.

Formally, it can be expressed with the following equation:

$$IV1_{ikt} = \sum_s (StarsInv_{ikst-2} \cdot AvPat_{ikst}) \quad (3.22)$$

where i indexes MSAs, t periods, k technological categories, and s technological subcategories within the category k . The first element of the product is calculated at point a), and the second one at point b).

The validity of the IV relies on an assumption of excludability for point a), i.e., once MSA fixed effects and the share of patents in a given category are controlled for, the number of star inventors active in a given MSA/category in period $n-2$ (on average ten years before) has no independent effect on the number of comet patents developed in period n in the same MSA/category; and on an assumption of exogeneity for b), i.e., the average productivity in the whole US is exogenous to MSA-specific unobserved factors.

There is, however, a reason of concern about the exogeneity assumption for point b). To the extent that comets in a given MSA are specialized in the same subcategories of stars, the US-wide variation in productivity in a subcategory can be correlated with the error term of equation 3.21. This in turn will compromise the validity of the instrument. We therefore build a second IV in order to improve the robustness of our estimate.

Second instrument

The second instrument follows a methodology similar to the first one, but the technological subcategories are substituted with the assignees of the patents. The steps are the following:

a) For the first period, we calculate the total number of star inventors active in a given MSA and with a given assignee. In case of star inventors with multiple MSAs or assignees in the same period, the modal one is chosen.

b) For each period, each assignee, and each MSA, we calculate the average number of patents produced by star inventors in that period in the whole US, excluding the given MSA.

c) For each MSA, period, and assignee, we multiply the number of inventors in the first period calculated at point a) by the average number of patents produced by star inventors sharing the same assignee in period t calculated in b). Subsequently, we sum the outcome by MSA, period, and technological category (if an inventor has patented in different categories in the same period, the modal one is chosen). The result is the second instrumental variable for total number of star patents in period t , by MSA and category.

Formally, it can be summarized by the following equation:

$$IV2_{ikt} = \sum_a (StarsInv_{ika1} \cdot AvPat_{iat}) \quad (3.23)$$

where i indexes MSAs, t periods, k technological categories, and a the assignees. In the few cases in which the value of point b was missing (because there were not other stars with the

same assignee in other MSAs), it was replaced with the contemporaneous US-wide average productivity of stars in the same technological category.

The excludability condition is identical to the one for the first instrument, while the exogeneity assumption is similar: given that stars and comets generally have different assignees (the assignee is very often the employer of the inventor, and comets have, by definition, assignees which less than 50 patents assigned in total - while, on average, assignees of stars have 4010 assigned patents) we assume that the average productivity of an assignee in the whole US (calculated excluding the given MSA) has no independent effect on the productivity of comets of that MSA.

3.5 Results

In table 3.25 we report mean and standard deviation of the patent variables for the 2113 MSA/category pairs which compose our sample. As it is possible to see, the distribution of the variables in natural form (first two rows) is very skewed. All the count variables (number of patents, number of firms) and total employment enter the regression equations in logarithmic form, thus the coefficients can be interpreted as elasticities. The variables which express continuous shares (the share of other patents in the same category, the Herfindahl index, and the share of manufacturing employment) are reported in natural form (thus the coefficients reflects percentage changes in the dependent variable following unit changes in the regressors).

Results from the OLS estimation are reported in col. 1, 2, and 4 in table 3.26. The effect of star patents on comets is always positive, but overall quite small: when the MSA

Table 3.25. Summary statistics of stars and comets

Variable	Obs	Mean	Std. Dev.	Min	Max
comets	2113	27.165	53.46	1	626
stars	2113	68.80	159.71	1	2125
log(comets)	2113	2.38	1.29	0	6.43
log(stars)	2113	3.07	1.46	0	7.66

fixed effect is included, the coefficient ranges from 0.03 to 0.11. Among the other controls, the share of patents in the category have a positive sign, as expected, although the latter is significant only in the specification without MSA fixed effects (col. 1). The same is true for the small plants variable. The total MSA employment is positive but significant in only one specification, while the Herfindahl index and the manufacturing share are always insignificant. The inclusion of the MSA-period fixed effects reduces the size of the star coefficient, which becomes insignificant (col. 4), and magnify the effect of the share of patents in the category. This is due to the fact that now the only variation left is within-MSA (i.e., across different technological categories) in the same period; which is probably too small to allow us to identify precisely any significant effect of stars (at least with OLS), considering also the strong collinearity of the two explanatory variables included (once other factors are controlled for).

Results from 2SLS regressions are reported in col. 3 and 5 of table 3.26. For brevity, here we report only the results obtained with the second instrument, as it is assumed to be the most exogenous. In Appendix C we report more specifications using also the first instrument, together with first stage estimates and other diagnostics; all the tests reported there confirm the validity of the IV specification and the strength of the instruments.

Table 3.26. regression of comet patents

VARIABLES	(1)	(2)	(3)	(4)	(5)
	comets (log)	comets (log)	comets (log)	comets (log)	comets (log)
	OLS	OLS	IV2	OLS	IV2
stars (log)	0.114*** (0.0215)	0.0980*** (0.0189)	0.273*** (0.0680)	0.0334 (0.0286)	0.303*** (0.0949)
Share other patents cat.	0.289*** (0.0797)	0.437*** (0.0909)	0.114 (0.144)	1.242*** (0.168)	0.635*** (0.241)
Tot. MSA patents (log)	0.369*** (0.0362)	0.0397 (0.0874)	0.00107 (0.0880)		
Total MSA empl. (log)	0.0755 (0.0529)	0.400* (0.242)	0.384 (0.238)		
Plants <500 emp. (log)	0.411*** (0.0609)	0.0444 (0.193)	0.0454 (0.185)		
herfindahl	-1.030 (2.023)	2.676 (3.070)	3.048 (3.079)		
Manuf. share	0.401 (0.432)	0.0214 (0.569)	-0.253 (0.557)		
Constant	-3.239*** (0.180)	-1.462 (1.208)	-0.981 (1.951)	3.532*** (0.378)	-1.475** (0.732)
MSA f.e.	NO	YES	YES	YES	YES
Tech. cat.*Period f.e.	YES	YES	YES	YES	YES
MSA*period f.e.	NO	NO	NO	YES	YES
Observations	2113	2113	2113	2113	2113
R^2	0.764	0.861	0.852	0.834	0.817

Heteroskedasticity robust standard errors clustered at MSA-category level in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Instrumented coefficients are still positive and significant, and now the elasticity of comet to star patents is around 0.3. The value is around three times bigger than OLS estimates (and ten times bigger when the MSA-time fixed effect is included). We explain the downward bias of the OLS as originating from negative selection: it is likely that, in general, patenting activities in a given city and category are specialized in one of the two groups of patents (comets or stars) for unobserved reasons, which in turn creates a negative, spurious association between the number of comet and star patents developed in a given MSA-category pair, thus leading to the downward bias in the OLS estimates. Another plausible explanation for the downward bias is the presence of a measurement error in the star variable: we proxy the intensity of activity of star inventors in a locality with the number of patents they produce, but the measure is clearly noisy, as patents are heterogeneous in quality. To the extent that the measurement error of the instrumental variable is independent from the one in the endogenous variable, IV estimates eliminate the "attenuation bias" of the OLS coefficient. The independence of the two errors is actually plausible as the variables are measured using patents in different localities (in the specific city and in the whole US excluding that city, respectively). We instead rule out that the increase in the coefficients may be due to weak instruments, as the hypothesis is rejected by first-stage results reported in Appendix C.

3.5.1 Robustness tests

We run a series of robustness tests to check the validity of our estimates. In table 3.27, we report the estimates of the model reported in equation 3.21 applying a Negative Binomial

count model to different selection of the sample: the OLS one, the OLS one plus the observation with zero comets, the OLS one plus the observation with zero comets and less than 25 patents in the MSA/technological category pair, and all the observation (thus adding also the observation with zero stars; to easy comparability, this is done by applying the logarithmic transformation to the natural variable augmented by one). We opted for a Negative Binomial, rather than a Poisson model, as the dependent variable shows a remarkable degree of overdispersion.

Results show that the coefficient of star patents is substantially unaffected by the different sample selections. Furthermore, its size is almost identical to the OLS one. We therefore exclude sample selection biases in our OLS estimations, due to either the exclusion of observations with zero comets or the threshold of 25 patents.

A further robustness test involves the inclusion of spatially lagged variables. Although the empirical literature on patents and knowledge spillovers has argued that urban agglomerations are a good approximation of the relevant spatial decay, we cannot exclude *a priori* that some of the effects we are looking at may go beyond the MSA borders. On the other hand, the exact identification of true spatial effects is complex in this context, as unobserved local factors may, in fact, create spurious evidence of spatial dependence. For instance, two contiguous cities may have similar numbers of comet patents because they share other, unobserved attributes, but failing to recognize that would lead to conclude that the number of comet patents in contiguous cities has a *causal* effect on city comets (this is a classic and well known identification problem in spatial economics, and more generally in social sciences, as discussed by Manski, 1999). Nevertheless, totally ignoring spatial

Table 3.27. Negative Binomial count regressions

VARIABLES	(1) comets (count)	(2) comets (count)	(3) comets (count)	(4) comets (count)
Sample	OLS	OLS + 0s	OLS + 0s+ <25 pat.	All
stars (log) [†]	0.132*** (0.0167)	0.129*** (0.0169)	0.147*** (0.0140)	0.153*** (0.0135)
Share other patents cat	0.558*** (0.0830)	0.524*** (0.0815)	0.410*** (0.0517)	0.401*** (0.0460)
Tot. MSA patents (log)	0.129** (0.0655)	0.131** (0.0667)	0.0592 (0.0515)	0.0274 (0.0441)
Tot. MSA patents (log)	0.300* (0.172)	0.345* (0.176)	0.683*** (0.163)	0.822*** (0.145)
Plants <500 emp. (log)	0.158 (0.136)	0.196 (0.140)	-0.0480 (0.117)	-0.149 (0.0990)
herfindahl	2.842 (2.535)	3.306 (2.503)	2.841 (2.096)	-0.0433 (1.535)
Manuf. share	-0.164 (0.474)	0.157 (0.520)	-0.306 (0.467)	0.0317 (0.375)
Constant	-1.408 (1.339)	-2.201 (1.406)	-2.832** (1.346)	-3.150*** (0.446)
MSA f.e.	YES	YES	YES	YES
Tech. cat.*Period f.e.	YES	YES	YES	YES
MSA*period f.e.	NO	NO	NO	NO
Observations	2113	2202	4191	7589

Heteroskedasticity robust standard errors clustered at MSA-category level in parentheses

*** p<0.01, ** p<0.05, * p<0.1

[†]This variable is equal to [log (stars +1)] in the regression of column 4

effects might be also an important omission. In this section, we apply some standard spatial econometrics tools, in order to check whether our results are robust to the inclusion of spatially lagged variables.

We therefore create a set of spatially lagged variables - namely the number of stars, comets, and other patents - calculated by weighting neighbouring observations - within a radius of 300 miles - by the inverse of their distance. Results are reported in 3.28. The inclusion of the spatial variables leaves the other coefficients almost unaffected, while the spatially lagged variables have generally significant coefficients, especially the "other patents" one. Including the spatial lag of the comets makes OLS estimations inconsistent as a spatial lag of the dependent variable is endogenous by construction (Anselin, 1988). Therefore, we opt for an IV estimation, instrumenting both the endogenous variables, i.e., the number of star patents and the spatial lag of comets. Regarding the choice of the instrument for the latter variable, a popular option in spatial econometrics literature is the spatial lag of one or a few independent variables, as long as one assumes that they do not have any independent effect on the dependent variable. However, in this case we have a better candidate promptly available, i.e, the spatial lag of the instrument. The fourth and fifth columns of 3.28 therefore report the results of an IV regression where stars and the spatial lag of comets are the endogenous variables, and the second IV and its spatial lag are the instruments. The stars coefficient is extremely similar to the previous IV regressions; the lagged comets have a sizeable coefficient when they are the only spatially lagged variable included in the specification (col. 4), although it is barely significant. Once the other spatial variables are included, it becomes insignificant; similarly, also the other spatial lags are not

statistically different from zero. One possible reason for that can be the high correlation among these three variables, which may introduce problem of collinearity; another source of concern can be the weakness of the instrument for the lagged comets once the other lagged variables are included, although standard first stage diagnostic seems to exclude the problem.³⁶ However, given that our main concern is to assess whether omitted spatially lagged variables may affect our results, rather than obtaining precise point estimates for these variables, for simplicity we decide not to complicate the specification any further; for the same reason, we omit the calculation of a spatial error model (robust to spatial correlation in the error term), as we believe that the large number of fixed effects included in the specifications, as well as the clustered structure of the estimated standard errors, make unlikely this kind of problem - which, however, would affect only the efficiency, and not the consistency, of our estimates. We therefore conclude that we cannot reject the presence of spatial effects in the context under analysis, but, at the same time, their omission is not affecting our main results.

Finally, we run two other robustness tests, which are:

- i) the exclusion of the sixth category, which includes all the patents not classifiable under the other five categories;
- ii) allowing for different effects of stars in each of the three time periods.

³⁶ The Kleibergen-Paap rk Wald F statistic for regression of col. 5 is equal to 5.86, which corresponds to a bias lower than 15% of the coefficient according to Stock-Yogo critical values. The same statistic for the regression of column 4 is equal to 45.947.

Table 3.28. Regressions with spatially lagged variables

VARIABLES	(1)	(2)	(3)	(4)	(5)
	comets (log)	comets (log)	comets (log)	comets (log)	comets (log)
	OLS	OLS	OLS	IV	IV
stars (log)	0.0877*** (0.0190)	0.0856*** (0.0189)	0.0854*** (0.0190)	0.252*** (0.0696)	0.251*** (0.0695)
Share other patents cat.	0.432*** (0.0906)	0.421*** (0.0901)	0.421*** (0.0902)	0.127 (0.141)	0.137 (0.142)
Tot. MSA patents (log)	0.0363 (0.0882)	0.0368 (0.0883)	0.0362 (0.0882)	0.00594 (0.0874)	0.00415 (0.0884)
sp. lag stars (log)	0.119*** (0.0305)		0.0156 (0.0512)		0.0176 (0.0536)
sp. lag oth. pat. (log)		0.193*** (0.0452)	0.175** (0.0769)		0.297 (0.231)
sp. lag comets (log)				0.141* (0.0734)	-0.248 (0.333)
Total MSA empl. (log)	0.369 (0.245)	0.387 (0.243)	0.382 (0.243)	0.355 (0.237)	0.460* (0.254)
Plants <500 emp. (log)	0.0499 (0.188)	0.0513 (0.187)	0.0513 (0.187)	0.0519 (0.178)	0.0519 (0.180)
herfindahl	2.182 (3.057)	1.952 (3.057)	1.963 (3.056)	2.513 (3.045)	2.464 (3.081)
Manuf. share	0.0344 (0.563)	0.0342 (0.560)	0.0344 (0.560)	-0.228 (0.547)	-0.205 (0.552)
Constant	-1.427 (1.241)	-1.734 (1.233)	-1.690 (1.232)	-1.460*** (0.529)	-2.233*** (0.776)
MSA f.e.	YES	YES	YES	YES	YES
Tech. cat.*Period f.e.	YES	YES	YES	YES	YES
MSA*period f.e.	NO	NO	NO	NO	NO
Observations	2096	2096	2096	2096	2096
R ²	0.863	0.864	0.864	0.856	0.854

Heteroskedasticity robust standard errors clustered at MSA level in parentheses

*** p<0.01, ** p<0.05, * p<0.1

In the first case, results are unaffected. In the second case, coefficients are not significantly different across periods, although the last one is generally slightly bigger. The hypothesis of a fading effect over time is therefore rejected.

Given the close similarity of these results with the ones already presented, they are not reported for brevity (they are however available from the author upon request).

3.6 Conclusions

This paper builds on the analysis of a very peculiar aspect of the patent data, i.e., the skewness of the distribution of patents among inventors. We therefore identify two illustrative categories of patents - stars and comets - based on the average productivity of their inventors. Two main conclusions emerge from the analysis: first, once controlling for the overall concentration of patenting activity, stars and comets are associated with cities with different structural characteristics. In particular, comets are associated with a diversified economic structure, concentration of small plants, and establishment births; while stars are more likely to be found in metropolitan areas with a large pool of patents and a specialized economic structure. Second, we show that the activity of star inventors is beneficial to the activity of comet inventors: in our preferred specifications, we find that the elasticity of comet patents to star patents is approximately equal to 0.3, which means that, on average, a 10% increase in the number of star patents leads approximately to a 3% increase in the number of comets.

More research is needed to expand both the conclusions we reached, in order to better identify the characteristics of cities associated with concentrations of the two cate-

gories of inventors; regarding the second, and to investigate the channels through which the spillovers take place. Also, the availability of a patent-firm matched dataset will allow i) to check our speculative hypothesis that comets are more likely to be employed by small firms, while stars work for the R&D labs of big companies; and ii) to assess more in depth the impact of the different categories of patents on the local economy.

The policy recommendations are not one-way. On one side, given the strong effect of stars on the productivity of comets, the attraction of stars to a city may be highly beneficial to the local economic environment: stars will benefit comets, which in turn will foster the birth of new plants, the innovation output of small businesses, and the generation of new employment. Thus, even though R&D labs of big corporations may have only a limited direct effect on the local economy, as most of the employment and value added is located elsewhere, they may be highly beneficial in the light of the aforementioned indirect effect.

On the other side, we know that stars and comets are concentrating in different places, which might imply that attracting stars where comets are might not be a successful policy, as stars in "comets' places" may be less productive. In other words, the same location for comets and stars will end up to be sub-optimal for (at least) one of the two categories. Therefore, interfering on the location choice of stars (or comets) in order to increase the spatial proximity may introduce perverse incentives and lead to much weaker effect than expected.

3.A Data

Patent data come from the United States Patent and Trademark Office (USTPO) database as processed by the National Bureau of Economic Research (NBER), described in Hall et al, 2001. To the original dataset we added the inventors' unique identifier developed by Trajtenberg et al (2006) and the standardized assignee name available in the Prof. Bronwyn H. Hall website.³⁷ We are aware that the latter is not always reliable as i) the complex ownership structure of companies may imply that differently named assignees correspond, in fact, to the same company, and ii) the same company name can be spelled in different ways (and the standardization routines cannot completely solve the problem).

We eliminated patents granted to inventors residing outside US and geolocated all the cities of residence of inventors through the ArcGis geolocator tool (based on the 2000 gazetter of US places from US Census) and the Yahoo! Maps Web Services. In case more authors are listed for the same patents and they live in different cities, the city of residence of the first author is chosen; this is a standard procedure in patent literature, and Carlino et al. (2007) show that the approximation is substantially innocuous. The geocoding operation was successful for 1,161,650 patents, which correspond to 97% of the database. We then assigned cities to counties using the ArcGis spatial join tool, and subsequently counties into MSAs (1993 definition). Those counties which are not included in the MSAs dataset are reported singularly - the geographical units are therefore a mix of counties and MSAs (for simplicity in the paper we do not distinguish between the two entities and call all the spatial units "MSAs"). This is a sensible choice to the extent that

³⁷ <http://elsa.berkeley.edu/~bhall/>

small counties not included in the MSAs definition do not exhibit strong commuting flows and are therefore self-contained functional entities. To our knowledge, is the first time that patent data are geocoded (almost) entirely, without ignoring small counties.

Other County and MSA specific variables for employment and industrial structure are calculated from the County Business Pattern dataset, while data on establishment births come from Company Statistics. Both the databases are freely available from the US Census webpage.

3.B Alternative definitions of comets and stars

In this appendix we present various alternative definitions of the patent variables, and we briefly discuss how the main results of the paper are affected.

The first concern about the definition we adopted regards the choice of considering only the first author of the patent. Looking at table 3.29, we can see that authors whose surname is starting with one of the first letters of the alphabet are only slightly more likely to be reported as first author, as compared to second or third authors. However, as a further robustness test, we followed a different procedure, defining a patent as a "star patent" if at least one inventor satisfies the requirements listed in section 3, and as a "comet patent" if all the inventors satisfy the relative requirements. The new variables are highly correlated with the single-author ones (99% pairwise, and 98% partial correlation when including also the total number of patents in the same MSA and category), and lead to extremely similar results: coefficients are only slightly (20-30%) smaller (table 3.30). Therefore, to the extent that the first author is generally the project leader, defining comet and star patents based only on her/him probably increase the precision of the estimates.

We then build three other definitions of the comet variable. They are the following:

- 1) Standard definition (described in Section 3) but including patents assigned to all the assignee types (not only to US corporations), or not assigned.
- 2) Same as in 1, but excluding not assigned patents.
- 3) Same as in section 3, but relaxing the constraint on the maximum number of 50 patents for assignee.

Table 3.29. Inventors' surname initial and patent authors' sequence

Initial	first author		second author		third author	
	Freq.	Percent	Freq.	Percent	Freq.	Percent
A	42,942	3.58	14,683	2.69	5,697	2.45
B	115,093	9.6	43,904	8.03	16,242	6.99
C	86,866	7.25	36,552	6.69	13,911	5.99
D	57,310	4.78	24,773	4.53	9,614	4.14
E	23,823	1.99	10,272	1.88	3,941	1.7
F	45,165	3.77	20,096	3.68	7,891	3.4
G	63,038	5.26	28,161	5.15	11,123	4.79
H	85,751	7.16	39,656	7.26	16,097	6.93
I	5,838	0.49	2,606	0.48	1,087	0.47
J	28,038	2.34	12,922	2.36	5,387	2.32
K	63,828	5.33	30,438	5.57	12,917	5.56
L	63,088	5.26	30,152	5.52	13,138	5.65
M	98,633	8.23	47,858	8.76	20,944	9.01
N	24,425	2.04	11,712	2.14	5,365	2.31
O	16,422	1.37	7,974	1.46	3,541	1.52
P	55,056	4.59	27,231	4.98	12,197	5.25
Q	1,854	0.15	970	0.18	386	0.17
R	55,828	4.66	26,368	4.82	12,045	5.18
S	124,636	10.4	60,864	11.14	27,666	11.9
T	37,138	3.1	18,570	3.4	8,690	3.74
U	3,582	0.3	1,769	0.32	928	0.4
V	17,480	1.46	8,525	1.56	4,342	1.87
W	63,419	5.29	30,428	5.57	14,356	6.18
X	304	0.03	247	0.05	120	0.05
Y	9,540	0.8	5,055	0.92	2,481	1.07
Z	9,282	0.77	4,735	0.87	2,297	0.99
Total	1,198,379	100	546,521	100	232,403	100

Table 3.30. regression of comet patents, multi-author

VARIABLES	(1) Comets (log) OLS	(2) Comets (log) OLS	(3) Comets (log) IV2
stars (log)	0.0795*** (0.0205)	0.0830*** (0.0181)	0.228*** (0.0645)
Share other patents cat.	0.355*** (0.0950)	0.543*** (0.0969)	0.249* (0.148)
Tot. MSA patents (log)	0.410*** (0.0372)	0.137 (0.0936)	0.0954 (0.0913)
Total MSA empl. (log)	0.120** (0.0536)	0.358 (0.243)	0.368 (0.239)
Plants <500 emp. (log)	0.352*** (0.0620)	0.0712 (0.178)	0.0574 (0.174)
herfindahl	-2.151 (1.931)	4.111 (3.057)	5.577* (3.037)
Manuf. share	0.516 (0.442)	-0.178 (0.596)	-0.521 (0.598)
Constant	-3.431*** (0.194)	-2.036 (1.273)	-1.875*** (0.558)
MSA f.e.	YES	YES	YES
Tech. cat.*Period f.e.	YES	YES	YES
MSA*period f.e.	NO	NO	NO
Observations	2088	2088	2088
R^2	0.763	0.864	0.857

Heteroskedasticity robust standard errors clustered at MSA level in parentheses

*** p<0.01, ** p<0.05, * p<0.1

We then calculated the results of the specification 3.21 with both OLS and IV (reported in table 3.31 and 3.32, respectively) and checked whether the results were affected. In the first case, the coefficients are reduced by around 50%, although they keep their significance. This is explained by the inclusion in the comet group of many patents not assigned or assigned to individuals, which are likely to bear less scientific and market value than other patents, and therefore they should benefit less from spillovers from stars (assuming that if the quality of patents is lower, there are less points of contact with excellent patents). The second definition gives coefficients that are around 20% smaller than the adopted definition; the difference is therefore small and due to similar reasons. The third comet variable gives a coefficient around twice as higher in the OLS, and similar to the one obtained with the standard comet variable in the IV specification. Again, this is not surprising, as comets defined in this way are more likely to work for the same employers of stars, which in turn leaves room for spurious positive correlation which pushes OLS estimates upward (which reduces the downward bias in the specific case).

To conclude, results are always qualitatively similar to the ones obtained with the standard definition of comets, and none of the (small) quantitative differences is unexpected.

Table 3.31. regression of comet patents, alternative definitions, OLS

VARIABLES	(1)	(2)	(3)
	comets def 1 (log)	comets def 2 (log)	comets def 3 (log)
	OLS	OLS	OLS
stars (log)	0.0488*** (0.0137)	0.0821*** (0.0190)	0.153*** (0.0147)
Share other patents cat.	0.465*** (0.0664)	0.566*** (0.0933)	1.164*** (0.102)
Tot. MSA patents (log)	0.210*** (0.0619)	0.163** (0.0824)	0.344*** (0.0458)
Total MSA empl. (log)	-0.0964 (0.163)	-0.0614 (0.232)	0.429*** (0.123)
Plants <500 emp. (log)	0.194 (0.124)	0.291 (0.185)	0.0382 (0.127)
herfindahl	3.460 (2.232)	2.867 (3.000)	1.735 (1.868)
Manuf. share	-0.950* (0.517)	-0.777 (0.747)	-0.409 (0.401)
Constant	0.673 (0.849)	-0.265 (1.120)	-2.039*** (0.678)
MSA f.e.	YES	YES	YES
Tech. cat.*Period f.e.	YES	YES	YES
MSA*period f.e.	NO	NO	NO
Observations	2113	2113	2113
R ²	0.764	0.861	0.852

Heteroskedasticity robust standard errors clustered at MSA level in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Table 3.32. regression of comet patents, alternative definitions, IV

VARIABLES	(1)	(2)	(3)
	comets def 1 (log)	comets def 2 (log)	comets def 3 (log)
	IV2	IV2	IV2
stars (log)	0.156*** (0.0495)	0.275*** (0.0718)	0.361*** (0.0427)
Share other patents cat.	0.275*** (0.0983)	0.218 (0.143)	0.806*** (0.101)
Tot. MSA patents (log)	0.191*** (0.0616)	0.126 (0.0817)	0.306*** (0.0505)
Total MSA empl. (log)	-0.114 (0.160)	-0.0843 (0.225)	0.386*** (0.135)
Plants <500 emp. (log)	0.189 (0.118)	0.292 (0.181)	0.0435 (0.116)
herfindahl	3.292 (2.078)	3.125 (2.997)	1.492 (1.767)
Manuf. share	-1.015** (0.493)	-1.045 (0.732)	-0.566 (0.365)
Constant	2.498* (1.365)	-1.049** (0.488)	-2.886** (1.239)
MSA f.e.	YES	YES	YES
Tech. cat.*Period f.e.	YES	YES	YES
MSA*period f.e.	NO	NO	NO
Observations	2113	2113	2113
R ²	0.764	0.861	0.852

Heteroskedasticity robust standard errors clustered at MSA level in parentheses

*** p<0.01, ** p<0.05, * p<0.1

3.C IV estimation diagnostics

In this appendix we present the results from alternative specifications relative to the IV estimations, following the recommendations reported in Angrist and Pischke (2008, p. 212). We report the results of the first stage regressions, in order to test the strength of the excluded instruments; subsequently, we test the exogeneity of the instruments, by comparing our main results presented in table 3.26 with the overidentified specifications (thus including also the first instrument) estimated through 2-stages least squares (2SLS) and Limited Information Maximum Likelihood models (LIML).

In table 3.33 we report the results from the first stage regressions; columns 1-2 are the specifications correspondent to our preferred IV estimations reported in columns 3 and 5 of table 3.26, respectively. In both the specifications, the coefficients on the instrument are highly significant, and the F statistics is well above the rule-of-thumb value of 10. In columns 3 and 4 we calculate two other first-stage regressions which do not have any direct correspondence to any of the 2SLS estimates we presented in the paper, but are meant to be a further test on the strength and exogeneity of the instrument: specifically, we added a MSA-category fixed effect (which we did not include in the main model), which absorbs every time-invariant component specific to a given MSA-category pair. As it is possible to see, the coefficient is less precise (but this is not surprising, given the little variability left) but it is still significant, and its size is even bigger than in columns 1 and 2. Column 5, instead, reports the result from the first-stage regression including both the excluded instruments. Again, the F statistic confirms the strength of the instruments. We

also calculated a F-test on null hypothesis that the instrumental variables are jointly equal to zero, which is clearly rejected ($F=46.21$).

In table 3.34 we report some diagnostics on the exogeneity of the instruments. Specifically, we estimate the overidentified regression (thus including also the first instrument) by means of Limited Information Maximum Likelihood models (LIML) as well as Two Stages Least Squares (2SLS). As Angrist and Pischke (2008) argue, LIML models are less precise but also less biased, thus sizeable differences in the point estimates with 2SLS equivalent specifications should be a reason of concern. However, in this case the coefficient values are very close either among them, and to the ones estimated in the main model of table 3.26. Therefore, we can conclude that the validity of the IV estimation is not a concern in our case.

Table 3.33. First stage regression

VARIABLES	(1) stars (log)	(2) stars (log)	(3) stars (log)	(4) stars (log)	(5) stars (log)
IV1 (logs)					0.102*** (0.019)
IV2 (logs)	0.268*** (0.0327)	0.243*** (0.0291)	0.362* (0.215)	0.528*** (0.149)	0.192*** (0.032)
Share other patents cat.	1.352*** (0.216)	1.656*** (0.127)	1.038*** (0.261)	0.0288 (0.161)	1.243*** (0.21)
Tot. MSA patents (log)	0.287** (0.115)			0.812*** (0.106)	0.298** (0.12)
Total MSA empl. (log)	-0.0352 (0.311)			-0.223 (0.265)	-0.153 (0.32)
Plants <500 emp. (log)	0.0190 (0.247)			-0.119 (0.210)	-0.0315 (0.25)
herfindahl	-2.627 (4.135)			-1.747 (3.885)	-3.691 (4.36)
Manuf. share	1.597* (0.920)			1.125 (0.747)	1.588* (0.90)
Constant	0.298 (1.758)	6.376*** (0.260)	1.748*** (0.506)	-1.366 (1.520)	1.091 (1.82)
MSA f.e.	YES	YES	YES	YES	YES
Tech. cat.*Period f.e.	YES	YES	YES	YES	YES
MSA*Period f.e.	NO	YES	YES	NO	NO
MSA*cat. f.e.	NO	NO	YES	YES	NO
Observations	2113	2113	2113	2113	2113
R ²	0.790	0.831	0.664	0.642	0.79
F-stat	27.04	24.60	N.A.	532.08	29.06

Heteroskedasticity robust standard errors clustered at MSA level in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Table 3.34. IV, overidentified regressions, 2SLS and LIML

VARIABLES	(1)	(2)	(3)	(4)
	Comets (log)	Comets (log)	Comets (log)	Comets (log)
	LIML	LIML	2SLS	2SLS
Stars (log)	0.310*** (0.0606)	0.322*** (0.0776)	0.308*** (0.0600)	0.322*** (0.0775)
Share other patents cat.	0.0456 (0.133)	0.592*** (0.216)	0.0495 (0.132)	0.593*** (0.215)
Tot. MSA patents (log)	-0.00715 (0.0897)		-0.00669 (0.0896)	
Total MSA empl. (log)	0.381 (0.242)		0.381 (0.242)	
Plants <500 emp. (log)	0.0456 (0.188)		0.0456 (0.188)	
herfindahl	3.127 (3.148)		3.123 (3.144)	
Manuf. share	-0.311 (0.569)		-0.308 (0.568)	
Constant	-1.113 (1.985)	-1.358** (0.679)	-1.106 (1.983)	-1.361** (0.679)
Observations	2113	2113	2113	2113
R^2	0.848	0.814	0.848	0.814

Heteroskedasticity robust standard errors clustered at MSA level in parentheses

*** p<0.01, ** p<0.05, * p<0.1

3.D Ancillary tables and results

Table 3.35. citations' shares, comets and stars, within tech. category

	Cited			
	Comets	Stars	Other patents	
Citing	Comets	16.5	16.8	66.7
	Stars	7.3	35.6	57.5
	Other patents	9.7	19.8	70.5

Table 3.36. Regression of citations received with tech. subcat. fixed effects

Dep. var.	Citations received (standardized)
Nr. citations made	0.00244*** (0.00014)
Star patent dummy	-0.125*** (0.012)
Comet patent dummy	-0.183*** (0.012)
Other patent dummy	-0.206*** (0.012)
Period F.E.	YES
Tech. subcat. F.E.	YES
Observations	590953
R^2	0.06

Heteroskedasticity robust standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Table 3.37. Regression of citations received excluding the top 5 per cent cited patents

Dep. var.	Citations received (standardized)
Nr. citations made	0.00817*** (0.00018)
Star patent dummy	0.243*** (0.0045)
Comet patent dummy	0.151*** (0.0053)
Other patent dummy	0.135*** (0.0041)
Period F.E.	YES
Tech. subcat. F.E.	YES
Observations	564339
R^2	0.13

Heteroskedasticity robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table 3.38. Regression of comets/stars shares at MSA level, SUR

COEFFICIENT	(1) Comets (share)	(2) Stars (share)	(3) Comets (share)	(4) Stars (share)
Tot. emp. (log)	-0.0237*** (0.0049)	0.0116 (0.0088)	0.00291 (0.0049)	-0.00365 (0.0094)
Herfindahl	-0.276** (0.12)	0.672*** (0.21)	-0.284*** (0.11)	0.677*** (0.21)
Manuf. share	0.0904** (0.036)	0.0503 (0.064)	0.0573* (0.033)	0.0694 (0.064)
N. plant <500 emp. ((log)	0.0286*** (0.0061)	-0.00404 (0.011)	0.0351*** (0.0056)	-0.00776 (0.011)
Other patents (log)			-0.0412*** (0.0029)	0.0237*** (0.0055)
Period dummies	YES	YES	YES	YES
Observations	1289	1289	1289	1289
R^2	0.11	0.03	0.23	0.04

Heteroskedasticity robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Fig. 3.7. Star and comet patents over employment, MSAs

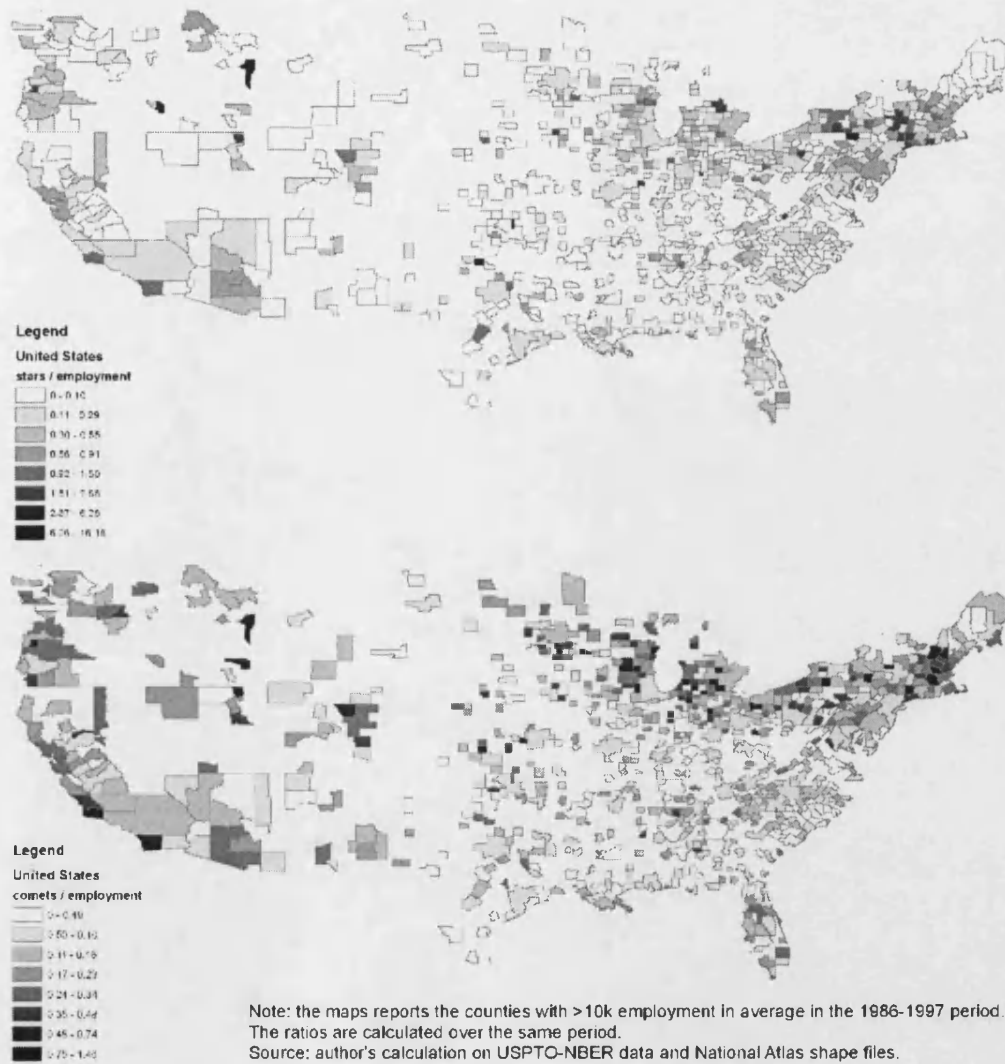


Fig. 3.8. Share of comets by MSAs

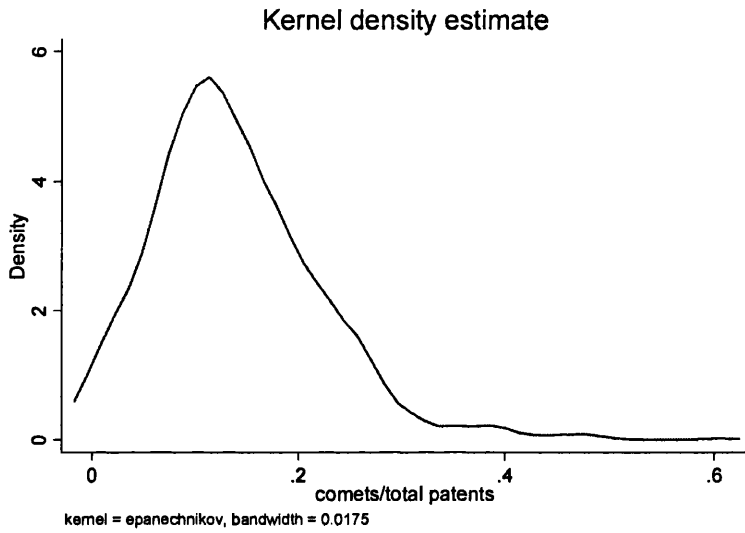
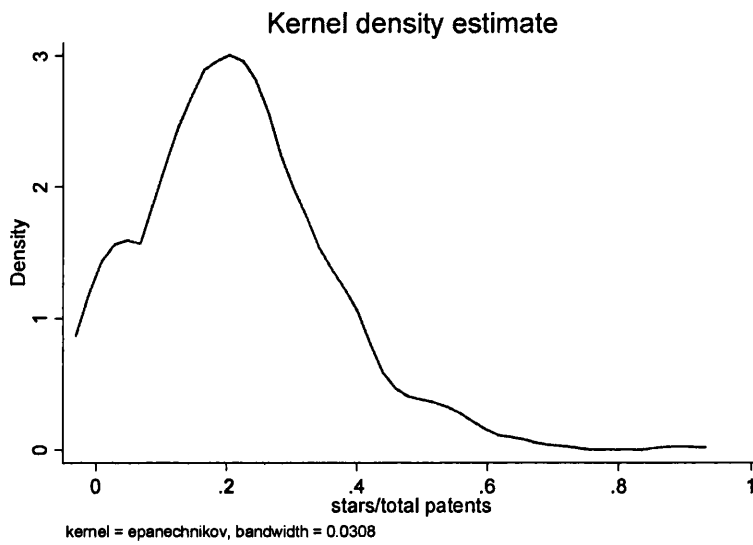


Fig. 3.9. Share of stars by MSAs



Chapter 4

Bank Location in the city of Antwerp: Evidence from Microgeographic Data

4.1 Introduction

This paper aims to analyse the effects of the structural changes in the retail banking industry in Belgium between 1991 and 2006 on patterns of branch location.³⁸ While most of the focus in the economics literature has been on analysing the pace and extent of liberalization at the level of a nation state or a larger geographical region, this study looks at these issues at a more disaggregated level, shifting focus to the very local scale, i.e., the most relevant to the everyday consumer of bank services.

Using detailed neighbourhood-level data for the metropolitan area of Antwerp, changes in the patterns of bank location are examined in three periods, corresponding to distinct stages in the deregulation process, namely 1991-1996, 1996-2001 and 2001-2006. We investigate how geographical branch penetration and potential choice among different bank groups have developed over time, and how these developments relate to neighbourhood characteristics. In particular, we are interested in assessing whether the change in bank

³⁸ The paper constitutes the empirical part of a bigger project, currently in progress, which will comprehend also a detailed theoretical analysis. This version has been produced by the author, which, nevertheless, greatly benefited from guidance, comments, and help of the two other colleagues involved in the research project: Marieke Huysentruyt (based at the dept. of Management at the London School of Economics and at the SITE at the Stockholm School of Economics), and Eva Lefevere, based at the Herman Deleeck Centre for Social Policy at the University of Antwerp. The reported author is the only responsible for any error or omission.

presence is significantly affected by the average income level in the neighbourhood, in the light of the potentially high social costs which this may bring.

Indeed, local access to finance services providers is particularly important for low income people. In fact, those are presumably less mobile than high income people, and therefore more affected by the local provision of financial services. On the other hand, high income people generally face a higher (opportunity) cost of time, and higher marginal utility of free time, than poor people. At the same time, rich people have a larger amount of saving, and therefore higher marginal utility from choosing the bank which offers them the highest interest rate (and higher monetary cost of opting for the sub-optimal one). Thus, when the stock of saving is large, the marginal benefit of traveling an extra kilometer to reach a bank offering better conditions may greatly offset the cost of the traveled distance. Therefore, to the extent that rich people have a lower (and potentially negative) incentive to become customers of local branches, the fact that a few branches are closing in the neighbourhood can be irrelevant to them. On the other side, poor people may have a small private return on opening a bank account; after taking into account for the cost of distance, the net benefit may become negative, which would imply that they may decide not to open a bank account if the first bank is too distant. However, this individual decision may bring a number of social costs. Generally, this component of the population has low within-city mobility, and, at the same time, scarce financial education and high needs of guidance for saving and budgeting. Very often, having a bank account is a prerequisite to access social benefits. Furthermore, the exit of mainstream banking from poor neighbourhoods may leave room to alternative service providers, or “fringe banking” services (Smith et al,

2008, test this hypothesis for the Philadelphia region, US), which are more expensive, and probably less likely to promote virtuous practices in saving and budgeting.

The study makes important contributions to the literature in at least three ways. First, only very few studies, in the economic literature, have been able to tackle geographical aspects of the banking industry (notable exceptions are e.g. Damar, 2007 for Turkey and Avery et al., 1999 for the United States) and even less have done so at a more disaggregated level than that of the nation state. The topic is far more popular in economic geography, where studies on "financial exclusion" or "financial segregation" abound (e.g., Leyshon et al, 2008; Chakravarty, 2006; Leyshon and Thrift, 1996), although with a completely different methodological approach than the one we follow. The urban economics literature also offers some relevant contributions related to redlining, i.e., the statistical discrimination which people living in poor neighbourhoods are prone to, for instance in accessing job vacancies or mortgages (e.g. Zenou and Boccoard, 2000). However, to the best of our knowledge no contribution has studied these phenomena in relation to bank geographic penetration. The main cause of this is a severe lack of appropriate data (for a comment on this see e.g. Udell, 1999). We circumvent this problem by constructing our own original dataset.

Second, the analysis of geographical phenomena requires the use of appropriate methodology in order to accurately deal with the specificities of spatial data. In this paper, we pay particular attention to developing the right empirical tools in order to provide precise evidence on the phenomena investigated.

Third, the intense and fast process of deregulation and concentration which took place in the late '90s makes Belgium an interesting quasi-natural experiment for assessing the impact of these processes on branch geographic penetration. At the same time, the city of Antwerp is also particularly well-suited to the analysis, as it is a city with very high income disparities, and strong pattern of neighbourhood segregation, thus making easier to detect whether changes in bank location are associated with differentials in neighbourhood wealth.

4.2 Background and Data

4.2.1 Deregulation in the Belgian retail banking industry, 1991 - 2006

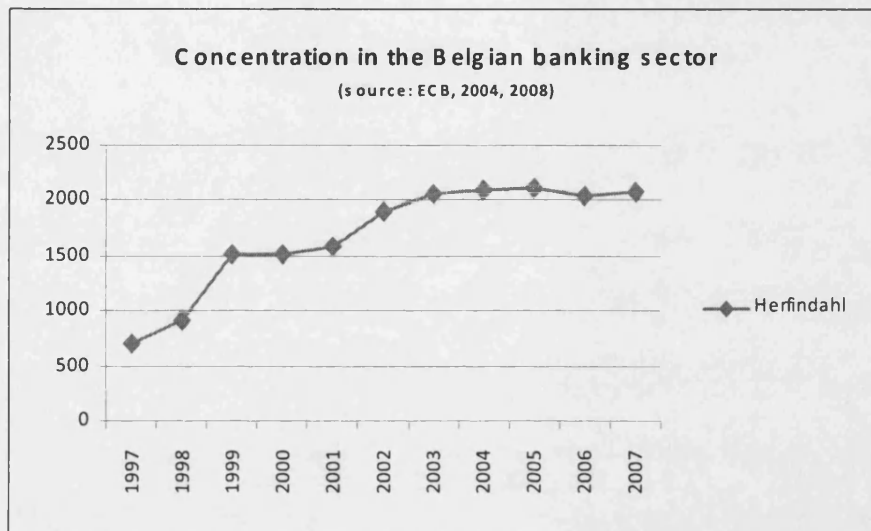
The European banking sector has liberalized dramatically during the past twenty years. Indeed, although the roots for the liberalization of the market in financial services can already be found in the Treaty of Rome (1958), it is generally acknowledged that the deregulation of the financial market remained very limited until the early nineties. From that time on, deregulation and European integration, together with globalization and technological change, have fundamentally changed the environment for the delivery of retail banking services. Although the degree to which the idea of a fully integrated European banking market has yet been realized remains open to debate (Goddard et al., 2007), no one can doubt the fact that far-reaching structural changes in the banking sector have taken place.

In our paper we study the combined effects of deregulation, liberalization and technological change on banking location patterns. We thereby look at three five-year periods that correspond to three rather distinct stages in these processes.

The first period, from 1991 to 1996, corresponds to the early starting years of the deregulation process in most West-European countries. In this period the two Banking Directives - and in particular the Second Banking Directive, issued in the '80s by the European Community - were implemented at national level (Cerasi et al., 2002). In Belgium the Directive was implemented between 1990 and 1994 (Gual, 1999). It established the principle of the single license allowing banks and other credit institutions to set up branches and to offer services throughout Europe (Ernoult et al., 2008). However, despite this and several other developments, to a large extent markets remained segmented, and often the measures were superseded by other political and economical developments.

In the second period of our analysis then, from 1996 to 2001, the deregulation speeded up substantially, and the European and Belgian banking landscape profoundly changed. The Financial Services Action Plan in 1999 gave a major impetus to financial services liberalization in Europe, together with the introduction of the Euro in the same year, evolutions in electronic commerce and improved technology (European Commission, 2003). It is in this period that the mergers and acquisitions that still determine the Belgian landscape for retail banking took place. Market concentration, measured for example by the Herfindahl index, almost doubled (European Central Bank, 2004 & 2008), causing Belgium to go from a moderately concentrated banking market in 1998 (Alegria and Schaeck, 2008) to one of the most concentrated in Europe nowadays (International Monetary Fund, 2006).

Fig. 4.10. Herfindahl index



In the last period in our analysis, 2001-2006, the deregulation process was further continued. Although there were some evolutions, these were less pronounced compared to the previous period. At the European Level, the introduction of the SEPA (Single European Payment Area) was implemented. In Belgium the five major players in the banking market continued to acquire other smaller banking institutions.

4.2.2 The city of Antwerp

Antwerp is the capital city of the province of Antwerp, located in the north of Belgium. With its 461,000 inhabitants it is the second largest city in Belgium (after Brussels). It covers an area of 20,151 hectare, 7,239 of which are harbour areas. The city is located on the right shore of the Scheldt River. In the heart of the city lies the 19th century city centre, surrounded by a Ring road. The harbour region is located in the north (dashed area in figure

4.11). The city is administratively divided in 9 so-called districts (indicated with bold lines in the figure). The central district, the district of Antwerp, roughly corresponds to the city centre. The other districts are located in the north, the east and the south of this central district. They all are governed both by their own council and by the council of Antwerp. All of the districts are divided into neighbourhoods (indicated with thin lines in figure 2), which constitute our unit of analysis.

Neighbourhoods³⁹ are the smallest spatial units for which statistical data can be obtained in Belgium. They were created for the Census in 1970, and revised in 1981 and 2001. Originally they corresponded to areas with uniform social, economic and morphological characteristics. Over time this within-neighbourhood uniformity somewhat diminished: although the neighbourhood borders were revised in 2001, changes due to evolutions in social, economic and morphological characteristics remained limited, in order to easy comparison over time. However, because the neighbourhoods are rather small (average area of 366,388 m², which - if they were circular - would correspond to a radius of 341 m) very large within-neighbourhood differences are rare.

Nowadays there are 277 neighbourhoods in the city of Antwerp. 29 of them were excluded from the analysis, because they were neighbourhoods with very specific characteristics and extremely low population density (ten or less dwellers in a vast area): we excluded all harbour neighbourhoods (7), all neighbourhoods next to the harbour (8), the neighbourhoods that corresponded to a green zone (8) and 3 very lowly populated (a few individuals) and remote areas (3). Only two of the excluded neighbourhoods had a bank

³⁹ The original denomination is 'statistische sectoren' or 'secteurs statistiques' (statistical sectors).

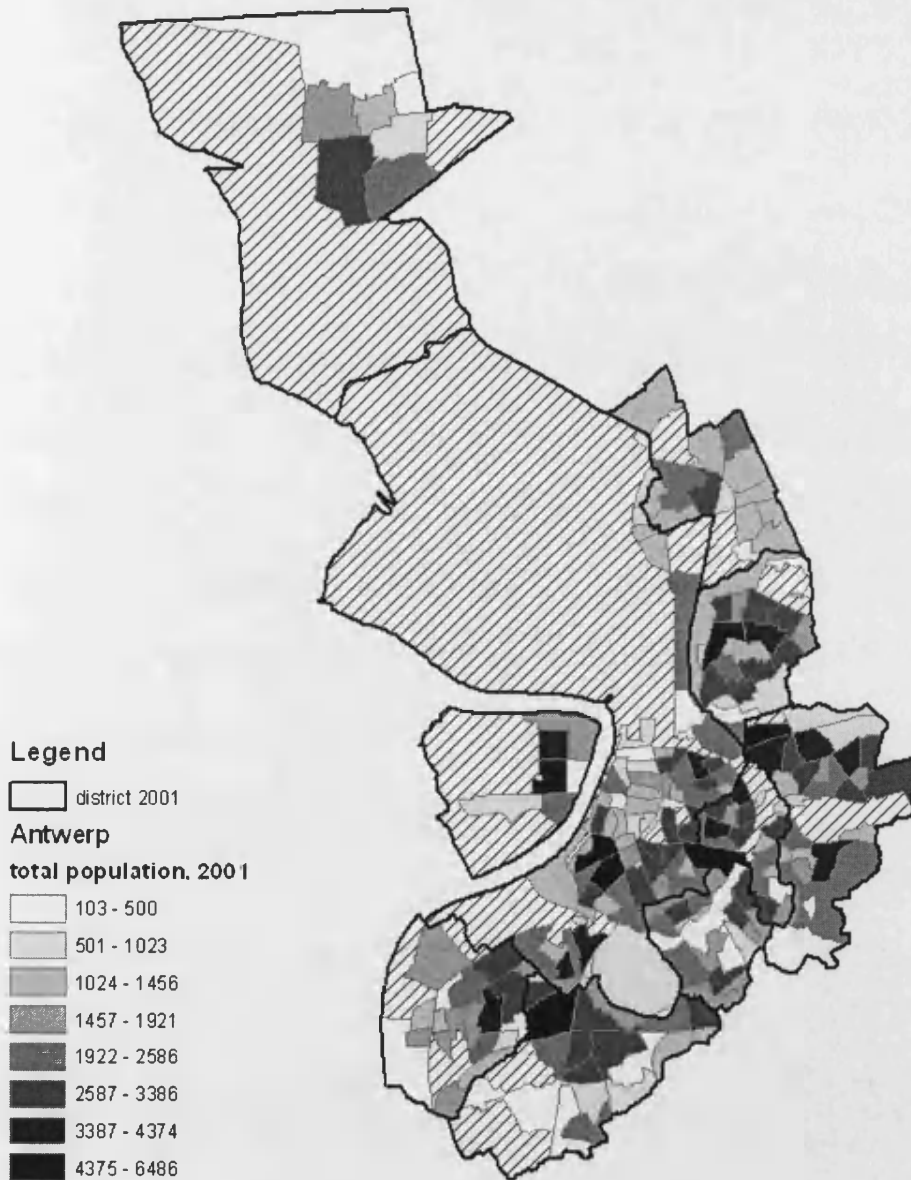
within them. Furthermore, 14 small and very lowly populated neighbourhoods (with less than 30 inhabitants) were merged to neighbouring neighbourhoods. We also merged together all the contiguous neighbourhoods which were affected by boundary changes over time, in order to create a classification invariant across time. This left us with 233 spatial units which constitute the base for our analysis.

Antwerp is particularly suited to our analysis for two reasons. First, it is a good example of a large urban region in Belgium and Europe. Second, and more importantly, it is characterized by a high degree of residential segregation and large income and ethnic disparities, which are much more distinct than in other Flemish cities (Kesteloot et al., 2006). In the year 2001, the average nominal income was equal to less than 11,000 Euro in the poorest neighbourhood (based on 432 income declarations), while in the richest one the same value was equal to more than 37,000 Euro (based on 401 declarations). The median income shows a similar dispersion, ranging from 10,000 to 33,000. These disparities are very persistent across time. To illustrate the overall pattern of geographic distribution of wealth, in figure 4.12 we report a map of average income by neighbourhood in the year 2001.

4.3 Data

Our 'raw data' consisted of the addresses of all bank branches operating in the metropolitan area of Antwerp from the telephone directories for the years 1991, 1996, 2001 and 2006. Using CRAB software, we converted these addresses into x-y coordinates. Based on these

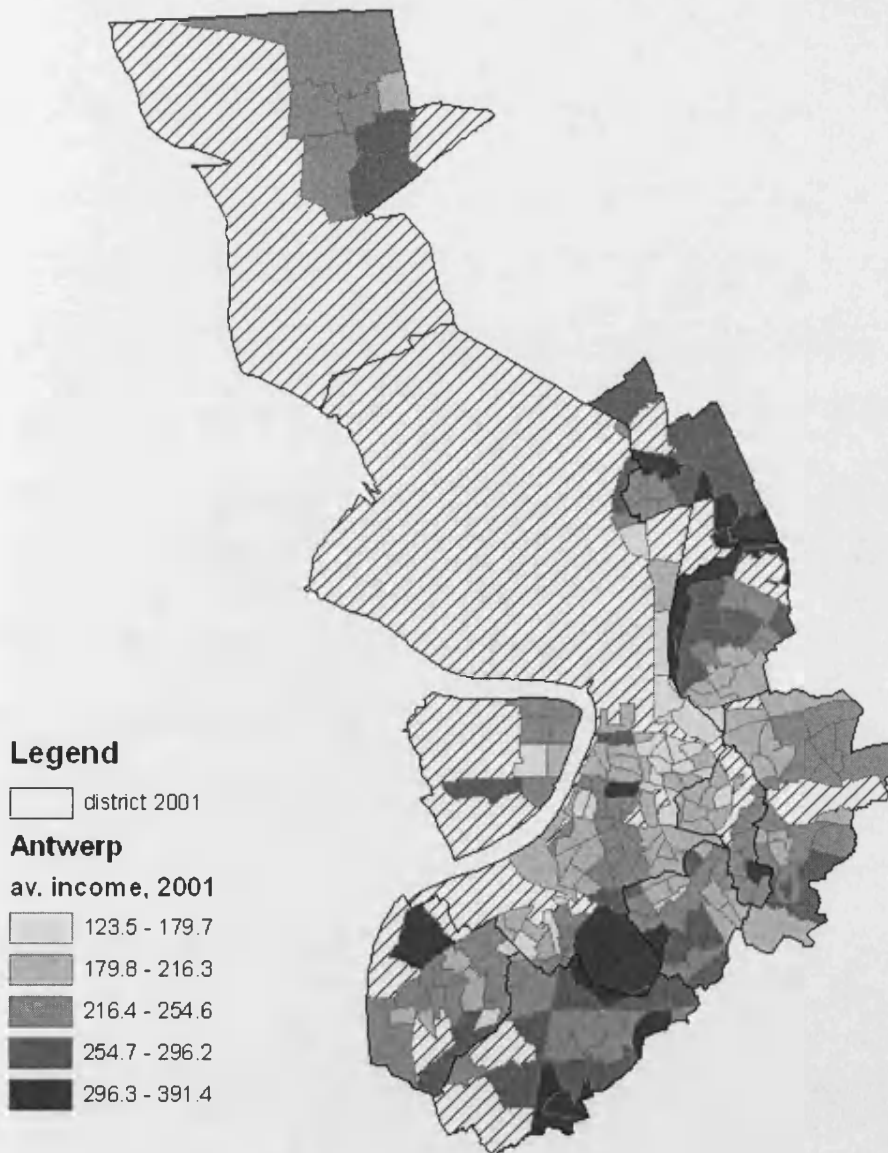
Fig. 4.11. The city of Antwerp: total polulation



Note: the figure reports the total population in the city of Antwerp in the year 2001. The thin lines are neighbourhood borders, and the bold lines are district borders. The dashed areas are lowly populated neighbourhoods, or harbour areas, which are excluded from the analysis.

Source: FPS Economy, Directorate General Statistics Belgium.

Fig. 4.12. The city of Antwerp: total polulation



Note: the figure reports the average income in the city of Antwerp in the year 2001. The thin lines are neighbourhood borders, and the bold lines are district borders. The dashed areas are lowly populated neighbourhoods, or harbour areas, which are excluded from the analysis.

Source: FPS Economy, Directorate General Statistics Belgium.

data several measures of bank presence and choice were composed, as we will explain below. These measures were then combined with data on neighbourhood characteristics.

4.3.1 Measure of bank presence, entry, and exit

Our measures of bank presence, entries, and exits are based on the full list of all banks, in which every observation is characterized by unique X-Y coordinate, a unique bank group,⁴⁰ and all the years at which the bank was found at the same address. The measures were composed in two stages. First, we classified all the banks as exiting, entering or continuing in a certain year. Second, we developed an original measure of bank location aimed at minimizing the bias originating from imposing a discretization to point events in the space.

We classified all the banks as exiting, entering, or continuing, according to the following criteria: if a bank was present in a certain year, but disappeared five years later, we considered this bank as exiting within this 5-year period; in case that a bank was not present in a certain year, but appeared five years later, we considered this bank as entering within this 5 years period; finally, a bank which is present in both the periods is defined as "continuing" in that interval. In order not to overestimate the number of exits and entries, we performed three corrections. First, we corrected our data for banks with a "discontinuous" time pattern. Whenever the presence of a bank was 'interrupted' in one of the years considered for the analysis, but then reappeared at the same address with the same bank group in the following period, we assumed that this was due to an inaccuracy in our data,

⁴⁰ In the following, with the term "bank" we mean a branch of a commercial bank, unless otherwise specified. We use the term "branch" with exactly the same meaning. With the term "bank group" we refer to the consortium to which the bank office belonged.

and therefore treated this bank as being present without interruption. This was the case for 8 banks. If the intermittence of a bank lasted for more than one period, no correction was carried out, and we recorded an exit and an entry, respectively. Second, when a bank simply 'moved' within a distance of 100 meters (this means: when within a certain period we recorded an exit and an entry of banks of the same bank group within 100 meters of distance), we did not consider this as exit or entry, but as an uninterrupted presence of one bank. This was the case for 132 banks.⁴¹ We assumed that the moves within a very small distance were driven by other forces than the ones we are interested in, e.g., banks could decide to move to a better suited building nearby; they can also be due to changes or inconsistencies in the addresses.

Third, the massive wave of mergers and acquisitions which involved many Belgian bank groups needed also to be considered while measuring bank presence. Indeed, the change of bank group of a branch due to a merger may be erroneously recorded as an exit and an entry of two different banks. In order to avoid that, whenever a bank changed bank group consistently with the wave of M&As which involved Belgium in the same period, we treated that bank as continuing.

The process we are analysing - the location of branches of commercial banks - is essentially a collection of "point events" in the space, i.e., phenomena which are not defined by a meaningful spatial extension. Most of the time data of this kind are aggregated into arbitrarily chosen spatial units, often by means of a simple count or of a density ratio. The spatial economics literature, indeed, has recently emphasized the *bias* originating from

⁴¹ In these cases we kept the original spatial coordinates.

“taking points on a map and allocating them to units in a box” (Duranton and Overman, 2005), especially if the “boxes” (i.e., the spatial units) are heterogeneously shaped and sized. This bias stems from the fact that distance is reduced to a binary variable (in/out) and arbitrary boundaries, which may not match real discontinuities in the spatial process under study, are simply imposed. Furthermore, since many banks in our study are located along a street which lies on the border of two neighbourhoods, allocating all the banks to one or the other of the two would yield only a rough approximation of bank presence.

Recently, a few contributions (Marcon and Puech, 2003; Duranton and Overman, 2005) have stressed the advantages of following a "Point Pattern Analysis" approach: following the seminal contribution by Ripley (1976), various statistics based on a continuous definition of space are proposed, with applications to the location of manufacturing plants and patterns of industrial agglomeration. This approach has been shown to provide a more precise evidence on the phenomena investigated, significantly improving on comparable statistics based on discrete spatial units (see for example the comparison of the Duranton and Overman (2005) metric with the Ellison and Gleaser index).

However, in our context a "pure" Point Pattern Analysis approach is limiting, as its metric is difficult to interact with socio-economic variables measured using discrete spatial units (neighbourhoods). Therefore, we develop a measure of bank presence at the local level which is much more precise than a simple neighbourhood count of bank events, but, at the same time, is neighbourhood-specific, and easy to match with other variables measured in the discrete space. In detail, we proceeded as follows. Around each "bank event" in the map, we calculated a probability density function which is highest at the location of

the event and reaches zero at a distance of 600 meters. We choose the distance of 600 meters because it is approximately correspondent to the hypothetical diameter of the average neighbourhood, and also because it seems a plausible distance to identify the maximum sphere of influence of banks. Different distances, however, only slightly affect the value of the statistic: we tried to use a distance of 300 meters and we obtained values that are correlated at 95% on levels, and at 93% on flows.⁴² We then calculated the kernel-smoothed sum of these values for each cell (a square of 65 meters width) in the map, applying the quadratic kernel function described in Silverman (1986, p. 76, eq. 4.5). This allowed us to draw a continuous surface of bank intensity covering the whole area under scrutiny. To compute a neighbourhood level statistic, we simply took the sum of the value of each cell which lay inside the neighbourhood's border. Our approach thus eliminates the "discretisation bias", given that the zonal statistic of a spatial unit now depends also on the presence of banks in the contiguous neighbourhoods, and generally increased the level of spatial precision. The methodology was easily applicable using GIS software.⁴³ We applied the same method to compute our final entry and exit measure: in this case, the "point events" we measured were the locations where banks were appearing or disappearing, respectively, in a given period.

We illustrate the calculation of the zonal statistic in figure 4.13. The map reports the area of the city of Antwerp we analyse (the empty zones are the harbour and the extremely

⁴² We also recalculated the results we discuss in the following of the paper and we obtained extremely similar results. This is in line with the software user's guide, which states that "increasing the radius will not greatly change the calculated density values. Although more points will fall inside the larger neighbourhood, this number will be divided by a larger area when calculating density."

⁴³ More precisely, we used two tools available in ESRI ArcInfo: the kernel smoothing tool, and the zonal statistic.

low populated neighbourhoods we excluded from the sample), the points represent banks, the polygons are the neighbourhoods, and the dark surface is the zonal statistic.

The econometric analysis we describe in the remainder of the paper significantly benefit from the higher precision of the zonal statistic, as compared to a simple count of banks by neighbourhood; we show that in Appendix A, where we re-estimate some of the main models substituting the zonal statistic metrics with the bank count.

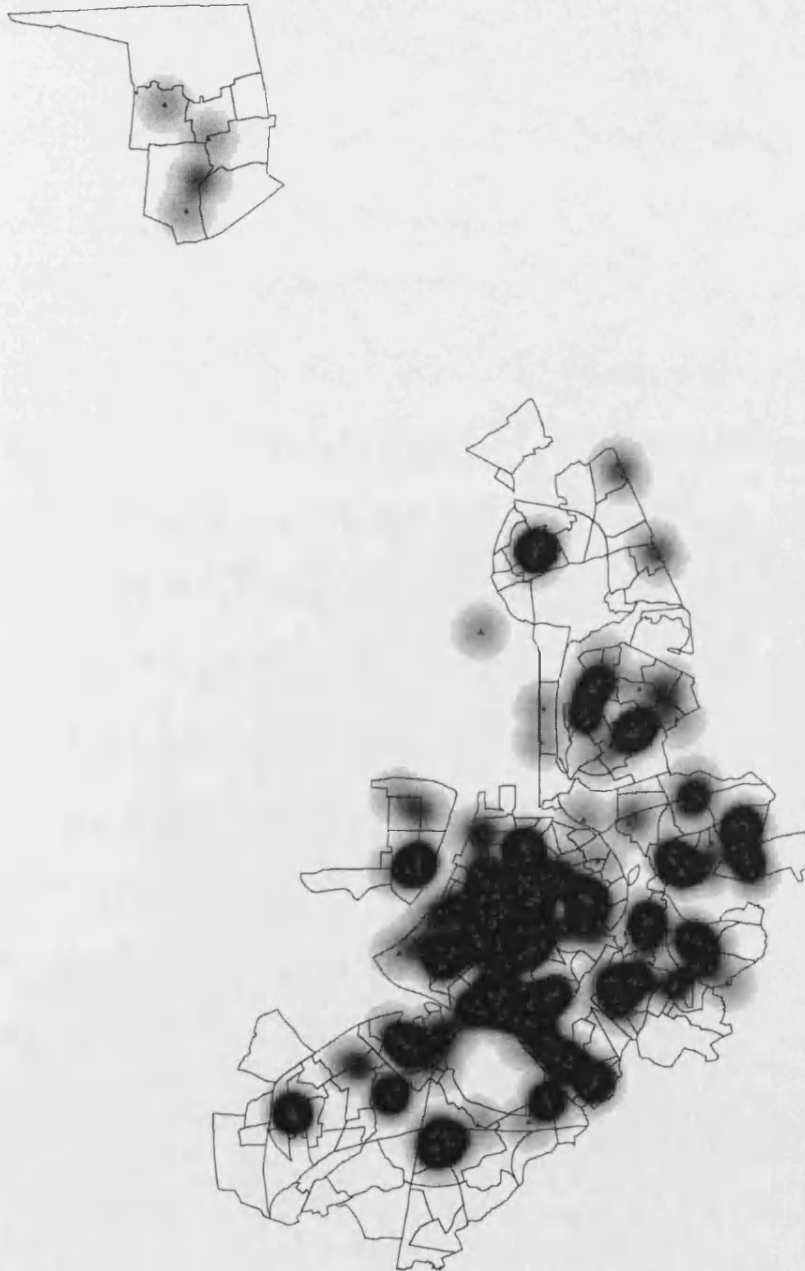
4.3.2 Measures of bank "desert" and bank choice

We calculated two other measures related to the location of banks in the period under scrutiny.

The first measure is aimed at assessing the level of "financial void", or "financial desert", in a given neighbourhood, and, essentially, tells us how long one should walk for to find the first bank, starting from a random point in the neighbourhood. More precisely, it quantifies the average area of the map that needs to be covered to find the first bank. It is calculated through the following steps:

- 1) we create a regular grid of points distant 100 meters each other, which covers the whole area under analysis
- 2) we draw progressively larger circles around each of them, until the first bank is met
- 3) we then calculate the total number of points of the grid that are within the largest circle; this gives us a statistics of "bank desert" for every point of the grid

Fig. 4.13. The zonal statistic



Note: the figure reports the city of Antwerp under analysis, in year 2001. The small triangles represent banks, the polygons are neighbourhoods, and the dark surface is the zonal statistic.
Source: Authors' elaboration.

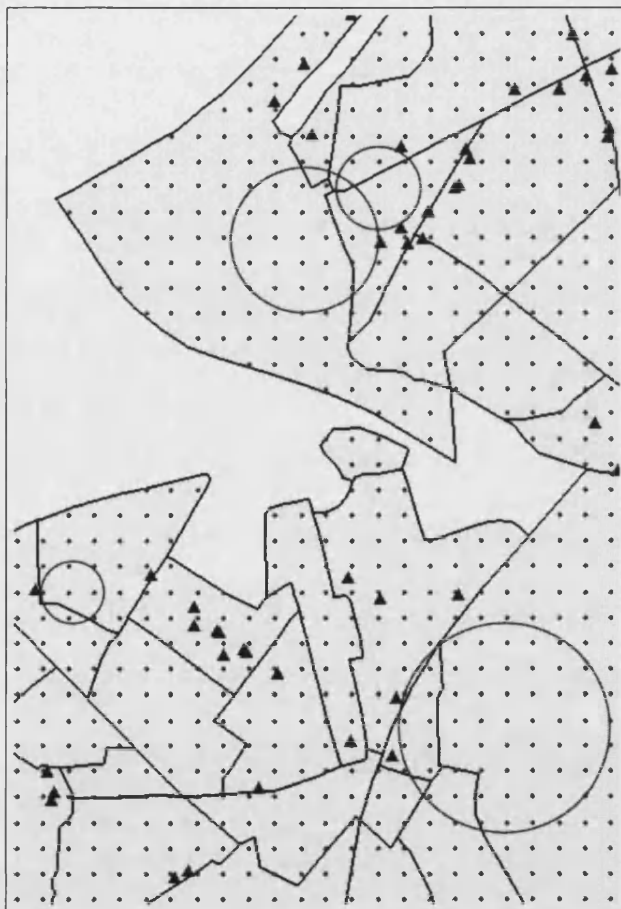
4) For each neighbourhood, we calculate the average value across all the points which fall inside it, which constitutes our final statistics.

The measure is similar to the average shortest distance to the first bank, with a couple of advantages: i) it is a quadratic function of the distance to the first bank, rather than a linear one; and ii) by using the number of grid points within the circle, rather than its area, we control for edge effects, i.e., for the fact that neighbourhoods close to the border of the map need a bigger distance to find a given number of banks within a given distance than do central neighbourhoods (given that the point grid cover only the area of the map, the number of points approximate the number of potential locations of banks).⁴⁴ An illustration of the measure is shown in figure 4.14, which reports a selection of the area under analysis, jointly with the point grid and the banks located in that area (black triangles). We also reported the neighbourhood borders, and a few examples of the circles drawn around each point of the grid. The neighbourhood average of the total number of points within each circle around each points of the grid constitutes our measure of bank desert.

The second measure is aimed at quantifying the level of bank diversity, or choice available to customers, in a given neighbourhood. Probably the most immediate solution would be to calculate a standard diversity index (e.g. the Herfindahl) at the neighbourhood level, but there are a few reasons to think that this is inappropriate. Neighbourhoods of a city are not randomly drawn, and spatial economic phenomena depend on their location and on the general spatial structure. E.g., a low level of competition in a neighbourhood may be explained by a high level of competition in the contiguous neighbourhoods. There-

⁴⁴ Marcon and Puech (2003) contain a detailed discussion of edge effects in point pattern analysis.

Fig. 4.14. Measure of "bank desert"



fore - similarly to our measure of bank location - a neighbourhood-level statistic which completely ignores the characteristics of the contiguous observations loses a significant bit of information.

Thus, to correctly measure the level of competition in the retail banking sector and its change over time, we need to develop an index which is not prone to a spatial aggregation bias and control for the spatial structure of the observations. On the other hand, we ideally need the measure to be specific for every neighbourhood, in order to match it with other available data.

The measure we propose is conceptually simple and similar to the one we adopted for the "bank desert": we use the same grid of points distant 100 meters each other, drawing progressively larger circles around each point until *banks belonging to three different bank groups* (at least) are met. The rest of the calculation is identical: we calculate the total number of points of the grid that are within the largest circle and we calculate the average across neighbours. Again, the statistics is easy to interpret: a bigger average circle corresponds to a longer hypothetical multi-directional walk from a point of the grid to enjoy a satisfactory differentiation of the supply of retail banking services. The correlation with the Herfindahl index calculated at neighbourhood level is significant but small (the Pearson's linear correlation is equal to 0.2, and the Spearman's rank correlation to 0.17), suggesting that our measure is indeed capturing different information as opposed to more traditional competition measure.

It is worth noticing that both the measures are neighbourhood-specific, but at the same time they depend non-parametrically on the spatial structure of the data. They are easily

comparable across spatial units and time periods. Technically, the measures have been calculated with a simple function in Matlab language developed by the author (available upon request).

4.3.3 Neighbourhood characteristics

Data on neighbourhood characteristics come from different sources. The population data (total population, percentage elderly, percentage people on working age, percentage foreigners) are official population statistics that come from the Federal Public Service (FPS) Economy, Directorate General Statistics Belgium. The data on income (average net taxable income per inhabitant) are official tax data. They were corrected for purchasing power by means of the consumer price index from the same source, and they are all expressed in Euro. The area of the neighbourhoods was calculated using GIS. We also calculated a centre dummy: neighbourhoods were considered as belonging to the centre if they were located within the main Ringway surrounding the centre of Antwerp.

In order to create a spatially lagged version of the neighbourhood characteristics, we create an inverse distance matrix including all the neighbourhoods within 2 km from the neighbourhood under consideration (distance is calculated at the centroid of the neighbourhood). We then create the lagged variables by pre-multiplying the matrix of explanatory variables by the row-standardized inverse distance matrix. Therefore, the spatially lagged version of a variable is equal to the average of the values of this variable in the neighbouring neighbourhoods, weighted by distance. In the tables, these variables are reported with a "W" before the variable name.

4.4 Empirical Analysis

4.4.1 Patterns of Bank Location

Methods

Our interest focuses on the detailed patterns of bank location *before, during* and *after* the banking service reformation of the late nineties. Did this liberalization outbreak (as well as deregulation and technological change) produced a discrete break in bank presence, and if so, which neighbourhoods benefited most from this? In particular, we are keen to see whether intensified competitive pressure in effect led banks to rationalize their bank presence, and to check if the intensity of this effect systematically varies according to other socio-economic characteristics of the neighbourhood (especially income). To examine this relationship, we estimate a regression of the following form:

$$\Delta y_{i,t+5}^s = \sum_t \beta_t X_t^s + \sum_t \gamma_t W X_t^s + \tau_t + \delta^s + \tau \delta_t^s + \varepsilon_{s,t} \quad (4.24)$$

where $\Delta y_{i,t+5}^s$ is the change in (the zonal statistic of) bank presence over a five-year period, starting at year t , in neighbourhood s , X_t^s is a vector of variables that characterize the neighbourhood s at time t , WX is a vector of those same variables but now spatially lagged, δ is a fixed effect for each of the nine districts in the city of Antwerp, τ is a time fixed effect, and ε is a neighbourhood-time specific error term. The coefficients of interest are elements of both β and γ , which we allow to change over time.

Given that we are especially interested in assessing how the *flows* of bank presence relate to the *stocks* of the explanatory variables in different time periods, we do not include a neighbourhood fixed effects as it would absorb most of the effect of the regressors, which generally show little variability across time. The model is estimated on the pooled sample by OLS with standard errors clustered at neighbourhood level. We also test whether different spatial specifications - namely the spatial error or spatial lag models - outperform the chosen model, and the answer is negative.⁴⁵

Variable specification

We regress the model reported in equation 4.24 using three dependent variables, namely the zonal statistic of i) the net flow of banks (entries minus exits), ii) entries, and iii) exits. However, given that the net flows is equal to entries minus exits, the coefficients of the first specification could also be calculated by the difference of the coefficients of the second and the third specifications.

We expect the banking groups to adapt their location strategy according to the socio-economic characteristics - especially average income - of the inhabitants of a neighbourhood, as well as to its size and to the presence of other branches. We therefore include the following, neighbourhood specific, explanatory variables:

a) The log of the average net taxable income, as a proxy for the wealth of the inhabitants and potential customers

⁴⁵ Given that the dependent variable is spatially smoothed, we would expect some spatial correlation in the error, and the fact that spatial error models do not outperform OLS is a bit surprising. However, an explanation for that could be that most of the spatial effect is absorbed by the spatially lagged independent variables and by the district dummies.

b) The log of total population, accounting for the size of the neighbourhood

c) The proportion of elderly (people over the age of 65), proportion of people in active age (people between 15 and 65), over the total population, and the proportion of non-Belgian nationals. These variables were used in order to test the hypothesis that banks target a specific kind of customer, namely the wealthy, Belgian, economically active customer.

d) The area of the neighbourhoods and a centre-dummy identifying all the neighbourhoods located within the main Ringway surrounding the centre of Antwerp, capturing the fact that the dynamics in the city centre are different from those in the surrounding neighbourhoods.

e) The level of zonal statistic, to control for both the pre-existing location of banks, existing competition, and also to proxy for unobserved factors which may make a neighbourhood more suitable for bank location (in natural form as it can be null).

f) A set of time, district (there are 9 districts in the city of Antwerp), and time-district fixed effects. The latter set of fixed effects is meant to control for unobserved shocks at the city and district level, as well as for other unobserved factors, like, for instance, the number of local businesses, the road and public transport network, etc.

To minimize the simultaneity bias, all the socio-economic variables refer to the beginning of the relative period, e.g., the zonal statistic flow for the period 1996-2001 is regressed on variables relative to 1996.

Table 4.39. Summary statistics

Variable	Year	Mean	Std. Dev.	Min	Max
active/tot. pop.	1991	0.63	0.06	0.45	0.79
active/tot. pop.	1996	0.61	0.06	0.43	0.80
active/tot. pop.	2001	0.61	0.07	0.29	0.79
average income*	1991	255.37	46.84	164.54	505.83
average income*	1996	242.07	45.69	135.20	436.99
average income*	2001	231.77	46.77	123.47	391.41
non Belgian/tot. pop.	1991	0.11	0.10	0.00	0.46
non Belgian/tot. pop.	1996	0.12	0.11	0.00	0.48
non Belgian/tot. pop.	2001	0.12	0.09	0.00	0.38
old/tot. pop.	1991	0.18	0.07	0.02	0.47
old/tot. pop.	1996	0.19	0.07	0.02	0.48
old/tot. pop.	2001	0.20	0.08	0.02	0.69
zonal statistic entry	1996	0.08	0.10	0.00	0.74
zonal statistic entry	2001	0.10	0.13	0.00	0.69
zonal statistic entry	2006	0.08	0.11	0.00	0.64
zonal statistic exit	1996	0.10	0.10	0.00	0.59
zonal statistic exit	2001	0.12	0.14	0.00	0.88
zonal statistic exit	2006	0.15	0.15	0.00	0.77
net zonal statistic level	1991	0.44	0.37	0.00	1.92
net zonal statistic level	1996	0.41	0.36	0.00	1.73
net zonal statistic level	2001	0.38	0.35	0.00	1.62
net zonal statistic level	2006	0.32	0.31	0.00	1.39
net zonal statistic flow	1996	-0.03	0.09	-0.42	0.42
net zonal statistic flow	2001	-0.03	0.13	-0.49	0.65
net zonal statistic flow	2006	-0.06	0.11	-0.45	0.24

* Hundreds of Euros in purchasing power parity

In a few specifications, we also add the spatially lagged version of the same variables, which is obtained by pre-multiplying the matrix of the explanatory variables by a row-standardized inverse distance matrix.⁴⁶

All the coefficients are allowed to change across the three periods (all the variables are included three times and interacted with period dummies), except the total population and the zonal statistic, which showed very little variability in the coefficient across time, and therefore are included only once for simplicity. In table 4.39 we report the summary statistics of the main variables.

⁴⁶ The distance matrix is limited to the maximum distance of two kilometres, and the distance is calculated at the centroids of the neighbourhoods.

Results

For every dependent variable, we estimate three different specifications: a "light" one including only average income, log of total population, and the period fixed effects (table 4.40); a complete one including all the variables (table 4.41); and one including all the variables and the spatial lags (table 4.42).

We first examine the regression of the net flow of zonal statistics. This variable indicates whether the bank presence in general, after taking into account entries and exits, became smaller or larger during a certain period of time (indicated by a negative or a positive sign of the indicator, respectively).

A general observation is that the regression yielded very different results for the three periods. Looking at the first column of table 4.40, we see that in the first period income is insignificant, while in the second and in the third period the effect is positive and highly significant. This means that the net inflow of banks to a neighbourhood is more strongly correlated with the level of income in later periods. Furthermore, in the second period the size of the coefficient is twice as large as it is in the third period, and the difference is significant. The size of the effect is remarkable: the standard deviation of the zonal statistic flow in 1996-2001 is equal to 0.13, thus, *ceteris paribus*, 1% less in the level of average income corresponds to more than one standard deviation outflow of banks. The standard deviation is itself very large, given that the mean of the same variable is equal to -0.03, and the mean of the level of the zonal statistic in 2001 is equal to 0.38 (as reported in 4.39). The inclusion of other controls, including the spatial lags (col. 1 of tables 4.41 and 4.42; in the latter table standard errors are omitted to ease readability, due to the large number of

variables included; the full table is available from the author upon request), affects the size of the coefficients and their significance, but leave basically unchanged the main conclusion we reached after the previous regressions, i.e., in the second period the net outflow of banks is much strongly correlated with the average income than in the two other periods.

The second column of tables 4.40, 4.41, and 4.42 report the regression of the zonal statistic of the entries. In this case, the effect of income is more similar across the three periods and, not surprisingly, is positive, although with variable levels of significance.

This necessarily implies that much of the variation of the income coefficient across time is due to the exits, as it is confirmed by the third columns of tables 4.40, 4.41, and 4.42. While in the first period the flow of exit is unexpectedly positively correlated with average income (which means that banks are exiting proportionately more from richer areas), in the second period the coefficient is clearly negative, to become then insignificant in the third one.

The size of the time dummies in table 4.40 is also interesting, as it summarizes the overall intensity of the flows in a given period. For the net flow, the largest dummy - in absolute value - is the second one, indicating that this has been the period with most of the outflow of banks. At the contrary, the entries have been roughly stable across time, while the exit are markedly more pronounced in the second and third period (the first dummy is significantly negative, while the two others are insignificant).

Of the demographics, only the percentage elderly yields significant and consistent results across the different specifications. Contrary to the effect of income, the effect of elderly is significant in the first and the third period. In these periods, we find a larger net

flow and more entries in neighbourhoods with a higher percentage of elderly. We suspect, however, that the reasons underlying the positive coefficient are different in the first and third periods. In the first period, it is probably due to the strategy of banks to expand their customer base; in the third period, the advent of internet banking is probably determinant, as older customers are less inclined to use the new technologies and are therefore relatively more targeted by branch location.

It also interesting to assess how these dynamics have affected the *level* of financial services availability across time. We therefore regress the level of the zonal statistic on the same set of variables of the previous models (excluding the lag of the zonal statistic). Results, reported in table 4.43, show two interesting aspects: first, the relationship between the levels of income and bank presence is not very strong, and it becomes significant only once other controls are included, especially the centre dummy; second, the trends in flow do affect the relationship, as coefficients on income become bigger and more significant over time (especially from the first to the second period), although only slightly. Therefore, we can conclude that at the beginning of the period under examination the level of "financial segregation" in the city of Antwerp was moderate, and that it has increased over that period in light of the inspected dynamics of bank flows. Overall, however, the difference between the first and last period are small, and the pairwise differences between the coefficients on income are never statistically significant.

Table 4.40. Regressions of zonal statistic, flows

VARIABLES	(1) Δ zonal statistic	(2) Entries	(3) Exits
Average income 1991	0.0284 (0.0395)	0.103** (0.0481)	0.0750* (0.0387)
Average income 1996	0.152*** (0.0363)	0.0662* (0.0393)	-0.0859** (0.0387)
Average income 2001	0.0716** (0.0345)	0.0546 (0.0414)	-0.0168 (0.0466)
Tot. pop. (log)	-0.0133*** (0.00459)	0.0296*** (0.00663)	0.0429*** (0.00746)
dummy first period	-0.0866 (0.225)	-0.711** (0.274)	-0.626*** (0.223)
dummy second period	-0.760*** (0.206)	-0.483** (0.226)	0.279 (0.228)
dummy third period	-0.356* (0.194)	-0.430* (0.234)	-0.0760 (0.263)
Observations	699	699	699
R^2	0.167	0.400	0.513

Heteroscedasticity robust standard errors in parentheses

*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Spatial diagnostics

Spatial dependency generally is detected through evidence of spatial autocorrelation in the residuals, and this may be due to three categories of spatial effects:

a) unobserved similarity in contiguous observations arising from factors uncorrelated with the included regressors, which could be either correlated or uncorrelated with the included regressors. In the first case there is an omitted variable bias, in the second case only the precision of the estimates are affected. For example, banks are entering in a specific area of the city because a new road has been built. We do not have data on roads therefore this factor is unobserved (and the construction of the road could, or could not, affect the socioeconomic characteristics of the neighbourhoods).

Table 4.41. Regressions of zonal statistic, flows, further controls

VARIABLES	(1) zonest_flow	(2) zon_entry	(3) zon_exit
Average income 1991	0.0968* (0.0552)	0.161** (0.0659)	0.0644** (0.0315)
Average income 1996	0.284*** (0.0526)	0.174*** (0.0429)	-0.110*** (0.0329)
Average income 2001	0.0814 (0.0598)	0.0931* (0.0516)	0.0119 (0.0378)
Zonal statistic	-0.115*** (0.0157)	0.196*** (0.0143)	0.311*** (0.0123)
Active/tot pop. 1991	0.152 (0.153)	0.330*** (0.110)	0.176 (0.112)
Active/tot pop. 1996	-0.511** (0.202)	-0.184 (0.160)	0.328*** (0.119)
Active/tot pop. 2001	0.198 (0.144)	0.136 (0.108)	-0.0617 (0.0931)
Tot. pop. (log)	0.00307 (0.00417)	-0.00308 (0.00426)	-0.00616* (0.00328)
Non Belgian/tot pop. 1991	0.254*** (0.0895)	0.159* (0.0886)	-0.0945 (0.0634)
Non Belgian/tot pop. 1996	0.162 (0.110)	0.130 (0.0905)	-0.0322 (0.0706)
Non Belgian/tot pop. 2001	0.0258 (0.117)	0.0609 (0.0922)	0.0345 (0.0844)
Elderly/tot pop. 1991	0.411*** (0.136)	0.352*** (0.113)	-0.0609 (0.0947)
Elderly/tot pop. 1996	0.000818 (0.162)	0.156 (0.162)	0.155 (0.110)
Elderly/tot pop. 2001	0.287** (0.122)	0.173* (0.0899)	-0.114 (0.0826)
Centre dummy	-0.00320 (0.0148)	0.0154 (0.0157)	0.0186** (0.00837)
Time f.e.	YES	YES	YES
District. f.e.	YES	YES	YES
Time-district f.e.	YES	YES	YES
Observations	699	699	699
R ²	0.342	0.685	0.849

Heteroscedasticity robust standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Table 4.42. Regressions of zonal statistic, flows, further controls and spatial lags

VARIABLES	(1) Δ zonal statistic	(2) Entries	(3) Exits
Average income 1991	0.0551	0.106*	0.0513
Average income 1996	0.229***	0.167***	-0.0629*
Average income 2001	0.0479	0.0763	0.0286
zonal statistic	-0.115***	0.197***	0.311***
Active/tot pop. 1991	0.227	0.492***	0.263**
Active/tot pop. 1996	-0.421*	-0.150	0.273*
Active/tot pop. 2001	0.291*	0.0994	-0.192*
Tot. pop. (log)	0.00151	-0.00438	-0.00590*
non Belgian/tot pop. 1991	0.228**	0.274***	0.0458
non Belgian/tot pop. 1996	-0.0127	-0.0112	0.00247
non Belgian/tot pop. 2001	0.148	0.0938	-0.0554
elderly/tot pop. 1991	0.397**	0.438***	0.0396
elderly/tot pop. 1996	-0.0495	0.128	0.179
elderly/tot pop. 2001	0.312**	0.137	-0.175*
W average income 1991	-0.0205	0.123	0.144
W average income 1996	0.587***	0.0968	-0.491***
W average income 2001	0.0685	-0.0313	-0.0999
W zonal statistic	0.0655	0.0850	0.0195
W active/tot pop. 1991	-0.369	-0.707*	-0.336
W active/tot pop. 1996	-0.569	-0.865**	-0.299
W active/tot pop. 2001	0.0305	0.208	0.179
W tot. pop. (log)	0.0392	-0.0136	-0.0526**
W elderly/tot pop. 1991	0.585	0.117	-0.468
W elderly/tot pop. 1996	-0.218	-0.329	-0.114
W elderly/tot pop. 2001	0.0365	0.0664	0.0327
W non Belgian/tot pop. 1991	-0.217	-0.541*	-0.324
W non Belgian/tot pop. 1996	0.938***	0.420	-0.520**
W non Belgian/tot pop. 2001	-0.385	-0.280	0.107
Centre dummy	-0.00805	0.0160	0.0241
Time f.e.	YES	YES	YES
District. f.e.	YES	YES	YES
Time-district f.e.	YES	YES	YES
Observations	699	699	699
R^2	0.374	0.700	0.858

*** p<0.01, ** p<0.05, * p<0.1

Time, district, and time-district dummies included in all the specifications

Table 4.43. Regressions of zonal statistic, levels

VARIABLES	(1) zonal statistic	(3) zonal statistic	(2) zonal statistic
Average income 1991	-0.0105 (0.121)	0.179 (0.128)	0.334** (0.139)
Average income 1996	0.0446 (0.111)	0.222* (0.119)	0.380*** (0.143)
Average income 2001	0.130 (0.103)	0.285** (0.110)	0.392** (0.158)
Active/tot pop. 1991			1.449*** (0.482)
Active/tot pop. 1996			0.467 (0.478)
Active/tot pop. 2001			0.710* (0.416)
Tot. pop. (log)	0.153*** (0.0225)	0.146*** (0.0222)	0.142*** (0.0204)
non Belgian/tot pop. 1991			0.764*** (0.276)
non Belgian/tot pop. 1996			0.536* (0.274)
non Belgian/tot pop. 2001			0.650* (0.351)
elderly/tot pop. 1991			2.431*** (0.428)
elderly/tot pop. 1996			1.870*** (0.408)
elderly/tot pop. 2001			1.793*** (0.353)
centre dummy		0.176*** (0.0501)	0.184*** (0.0627)
Time f.e.	YES	YES	YES
Time-district f.e.	NO	NO	YES
Observations	699	699	699
R^2	0.613	0.635	0.695

*** p<0.01, ** p<0.05, * p<0.1

Heteroscedasticity robust standard errors in parentheses

b) correlation of the spatial lag of the regressors with the dependent variable (i.e., WX affects Y): e.g., banks are entering in a neighbourhood because contiguous neighbourhoods are becoming richer (and are obtaining more banks as well).

c) causal effect of the contiguous dependent variable on the dependent variable (WY affects Y): e.g., banks are exiting from a neighbourhood because a lot of banks are entering in the contiguous neighbourhoods, thus raising the competition pressure.

The model reported in equation 4.24 fully accounts for the second effect by the inclusion of a set of spatially lagged variables, it partially deals with the first one by including the district dummies, and it does not consider at all the existence of the third kind of effects. However, to the extent that the third effect is actually in play, the bias in our estimates can be serious. Furthermore, this effect cannot be excluded on a theoretical ground as well, as flows of bank presence in a given neighbourhood may indeed have a true causal effects on flows in contiguous neighbourhoods. Models fitted to cope with that - called "spatial lag models" - cannot be estimated by OLS as the spatially lagged dependent variable is endogenous by construction (this is also known as "you are your neighbour's neighbour" problem), and are thus generally estimated by maximum likelihood (Anselin, 1988). However, in a longitudinal setting further complications arise and frontier econometric techniques need to be applied (for a survey of available methods see Elhorst, 2009). Avoiding these models, whenever they are not strictly necessary, is rewarding in terms of both efficiency and simplicity of estimates.

Appropriate statistical tests show that more complex models are indeed unnecessary in our context. More precisely, we estimate model 4.24 in a cross-section setting (thus

allowing all the coefficient to vary over time) including a spatial autoregressive parameter in the error with the following structure:

$$\epsilon_i = \lambda W \epsilon_i + u_i \quad (4.25)$$

where λ is the spatial autoregressive parameter, W is a spatial contiguity matrix, and u is a vector of homoskedastic and uncorrelated errors. We then run a Lagrange Multiplier test on the significance of the λ coefficient, in both the standard and robust version (Anselin and Hudak, 1992). Subsequently, we estimate a spatial lag version of model 4.24, by adding a spatially lagged dependent variable on the RHS of the equation (ρWY). Again, we then run a Lagrange Multiplier test (and its robust counterpart) of significance of the autoregressive parameter ρ . The results of the test are reported in table 4.44: none of the non robust versions of the tests are significant at 5% level, and only one statistic is significant at 10% (LM spatial lag for entry in 1991-96). Considering that the robust tests should not be considered when the non robust versions are not significant (Anselin et al., 1996; Anselin and Florax, 1995), we can therefore conclude that, overall, the model reported in equation 4.24 does not omit significant spatial effects.

Table 4.44. Spatial diagnostics, p-values

Period	Dep. var.	LM error	LM error robust	LM sp. lag	LM sp. lag robust
1991-1996	net flow	0.50	0.28	0.27	0.16
	entry	0.31	0.12	0.07	0.03
	exit	0.18	0.73	0.19	0.93
1996-2001	net flow	0.76	0.66	0.66	0.58
	entry	0.79	0.63	0.69	0.57
	exit	0.87	0.07	0.68	0.07
2001-2006	net flow	0.54	0.97	0.51	0.80
	entry	0.55	0.07	0.94	0.09
	exit	0.54	0.06	0.84	0.08

Further robustness: M&A-induced exits

A massive wave of M&As involved the Belgian banking sector during the period of analysis, and especially from the late '90s on. Branch closures may be related to M&As for two reasons. First, M&As are often followed by a general rationalization of the existing branch network; second, we may observe a number of closure of branches due to the fact that two contiguous banks belonging to different groups before the M&A became part of the same group after the M&A; as a consequence, one of the two closed. While the first factor is difficult to identify in the data (we do not know how many branches of the same groups would have closed in absence of the M&A), we can instead detect fairly precisely all the branches which closed for the second reason. We therefore decided to identify them, and to estimate the same regression of tables 4.40, 4.41, and 4.42 excluding them.

Specifically, for every bank exiting in a given period, we checked whether i) this bank has become of the same group of another bank located within the radius of 300 meters following a M&A, or ii) another bank within 300 meters has become part of the same group. If one of two conditions was satisfied, we identified this exit as due to a M&A. They account for around one third of all the exits in the second and third period (we do not observe any M&A in the first period).

Subsequently, we re-estimate model 4.24 subtracting from the dependent variables (net flows and exits) the exits due to mergers. Results, reported in table 4.45 and 4.46, show that, overall, excluding M&A-induced exits does not contradict the main results of the previous section;⁴⁷ nevertheless, estimates are less precise and coefficient are smaller,

⁴⁷ We omitted standard errors from table 4.46 to ease readability; the full table is available from the author

which suggests that this kind of exits is correlated with the general trend of all the exits, and that it is not randomly located in space.

Table 4.45. Regressions excluding exits due to mergers

VARIABLES	(1) Δ zonal statistic	(2) Exits
Average income 1991	0.0324 (0.0395)	0.0710* (0.0384)
Average income 1996	0.0893** (0.0362)	-0.0232 (0.0233)
Average income 2001	0.0404 (0.0284)	0.0144 (0.0365)
Tot. pop. (log)	-0.00412 (0.00451)	0.0337*** (0.00616)
dummy first period	-0.176 (0.226)	-0.536** (0.218)
dummy second period	-0.439** (0.201)	-0.0431 (0.143)
dummy third period	-0.210 (0.163)	-0.221 (0.209)
Observations	699	699
R^2	0.064	0.498

Heteroscedasticity robust standard errors in parentheses

*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

4.4.2 Bank desert and bank choice

We analyse the measures of "bank desert" and bank choice with a methodology similar to the one we adopted for bank location. Technically, the approach is exactly symmetrical, as we use the same model (reported in equation 4.24), and we include the same explanatory variables, thus the only difference is the dependent variable. However, these two measures have different spatial properties than the zonal statistics, as they are much more dependent on the value of the same variable in neighbouring observation - which, in passing, was

upon request.

Table 4.46. Regressions excluding exits due to mergers, further controls

VARIABLES	(1) Δ zonal statistic	(3) Exits	(2) Δ zonal statistic	(4) Exits
Average income 1991	0.0772	0.0840***	0.0300	0.0764**
Average income 1996	0.236***	-0.0617**	0.200***	-0.0338
Average income 2001	0.0466	0.0467	0.0237	0.0528*
zonal statistic	-0.0281*	0.225***	-0.0320*	0.229***
Active/tot pop. 1991	0.0273	0.301***	0.145	0.345***
Active/tot pop. 1996	-0.328	0.145	-0.389	0.241*
Active/tot pop. 2001	0.122	0.0145	0.111	-0.0124
Tot. pop. (log)	-0.00178	-0.00131	-0.00392	-0.000470
non Belgian/tot pop. 1991	0.115	0.0438	0.181*	0.0928
non Belgian/tot pop. 1996	0.169*	-0.0392	-0.0236	0.0134
non Belgian/tot pop. 2001	-0.0142	0.0744	0.0247	0.0681
elderly/tot pop. 1991	0.206	0.144*	0.240	0.196**
elderly/tot pop. 1996	0.142	0.0141	0.00344	0.126
elderly/tot pop. 2001	0.241**	-0.0683	0.206	-0.0695
W average income 1991			-0.0207	0.144
W average income 1996			0.387**	-0.291**
W average income 2001			-0.126	0.0951
W zonal statistic			0.163**	-0.0784*
W active/tot pop. 1991			-0.835	0.129
W active/tot pop. 1996			-0.352	-0.516
W active/tot pop. 2001			0.0386	0.171
W tot. pop. (log)			-0.0108	-0.00265
W elderly/tot pop. 1991			0.242	-0.125
W elderly/tot pop. 1996			0.0626	-0.394
W elderly/tot pop. 2001			0.151	-0.0820
W non Belgian/tot pop. 1991			-0.535*	-0.00554
W non Belgian/tot pop. 1996			0.832***	-0.414**
W non Belgian/tot pop. 2001			-0.361	0.0824
centre	0.0145	0.000901	0.00142	0.0146
Time f.e.	YES	YES	YES	YES
District. f.e.	YES	YES	YES	YES
Time-district f.e.	YES	YES	YES	YES
Observations	699	699	699	699
R ²	0.171	0.797	0.211	0.804

*** p<0.01, ** p<0.05, * p<0.1

exactly one of the aims of the statistic. This is due to how the measure is built, rather than to a specific spatial interaction process; however, the econometric consequence is that, in this case, the null hypothesis of insignificance of a spatial autoregressive component for the dependent variable is rejected. Nevertheless, a spatial lag model would mostly absorb this nuisance in the data, rather than a true spatial interaction process. We, therefore, prefer to keep the OLS estimates; however, alternative estimations - reported in Appendix B - based on spatial lag and spatial error models gave very similar results.

Although similar in construction, the two measures have two different aims: the bank desert is a measure of bank location, therefore similar to the zonal statistic, while the bank choice is aimed at capturing the change in the differentiation of financial services at the local level. Regressions using the bank desert index, therefore, can be seen as a robustness check of the previous results. The correlation of the change in bank desert with the change in zonal statistic at neighbourhood level is, however, in the region of 0.2, which suggests that the two measures are capturing different information.

Results - reported in the first column of table 4.47 and in the first and second column of table 4.48 - substantially confirm the main finding of the previous regressions. The change in the average distance to the first bank is negatively correlated to the average income in the neighbourhood only in the second period; the evidence is robust across the different specifications. In the first column of table 4.47 we can also compare the size of the different time dummies: the second is the only one significant, implying that, overall, in the second period the distance to the first bank has increased significantly more than in the two other periods.

The regressions of the distance to three different bank groups (col 2 of table 4.47 and col 3-4 of table 4.48) yield similar results. In the simpler specification of table 4.47, the coefficient on income is insignificant in the first period, negative in the second, and positive and barely significant in the third period. This latter result, however, is not robust to the inclusion of further controls (table 4.48). Again, the time dummy in table 4.47 is significant only in the second period.

Table 4.47. regression of bank "desert" and bank choice

VARIABLES	(1) Δ dist 1st bank	(2) Δ dist 3 groups
Average income 1991	1.966 (11.24)	22.53 (16.62)
Average income 1996	-15.26*** (4.802)	-37.69*** (13.90)
Average income 2001	24.08 (15.87)	41.54* (23.51)
Tot. pop. (log)	-4.240* (2.491)	-5.725** (2.761)
dummy first period	24.55 (67.84)	-76.71 (88.93)
dummy second period	115.8*** (35.92)	249.1*** (80.45)
dummy third period	-87.99 (81.58)	-154.8 (126.8)
Observations	699	699
R^2	0.070	0.133

Heteroscedasticity robust standard errors in parentheses

*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Table 4.48. regression of bank "desert" and bank choice, further controls

VARIABLES	(1) Δ dist 1st bank	(2) Δ dist 1st bank	(3) Δ dist 3 groups	(4) Δ dist 3 groups
Average income 1991	-5.476	4.463	14.99	35.41**
Average income 1996	-20.21***	-11.88**	-39.60***	-29.83**
Average income 2001	1.919	18.83	13.51	2.068
Zonal statistic	1.311	2.943	0.593	-0.611
Active/tot pop. 1991	2.605	0.948	7.319	-88.51**
Active/tot pop. 1996	-36.58	-31.09	68.03	90.10
Active/tot pop. 2001	-108.0*	-129.5*	-71.30*	-184.9**
Tot. pop. (log)	-4.523*	-3.544	-5.020*	-5.235*
non Belgian/tot pop. 1991	-22.08	0.855	-15.97	-5.167
non Belgian/tot pop. 1996	-26.86	-9.862	14.93	29.08
non Belgian/tot pop. 2001	-92.48	-23.07	85.76	2.853
Elderly/tot pop. 1991	-31.73	-19.07	-55.63	-102.1**
Elderly/tot pop. 1996	-44.63*	-34.87**	37.45	45.51
Elderly/tot pop. 2001	-197.4***	-183.6***	-65.48	-181.2**
W average income 1991		-12.06		9.497
W average income 1996		-71.36**		-27.25
W average income 2001		-140.6*		251.2**
W zonal statistic			-70.87***	-37.11
W active/tot pop. 1991		-75.63		374.7**
W active/tot pop. 1996		-109.2**		-301.6*
W active/tot pop. 2001		292.5*		229.4
W tot. pop. (log)		-20.28*		-33.67
W elderly/tot pop. 1991		-241.0*		-185.8
W elderly/tot pop. 1996		-73.54		-277.6*
W elderly/tot pop. 2001		-330.2*		130.2
W non Belgian/tot pop. 1991		-11.48		199.6**
W non Belgian/tot pop. 1996		-75.50		69.73
W non Belgian/tot pop. 2001		-463.6**		617.0***
centre	-0.768	-3.031	6.496	-12.33
Observations	699	699	699	699
R ²	0.201	0.265	0.394	0.430

Heteroscedasticity robust standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

4.5 Conclusion

The intense and fast process of deregulation and concentration which involved the Belgian banking sector in the late '90s makes the country an interesting quasi-natural experiment for assessing the impact of these processes on branch geographic penetration at the very local scale.

Therefore, using detailed neighbourhood level data for the metropolitan area of Antwerp, changes in the patterns of bank location are examined across three subsequent periods corresponding to distinct stages in the deregulation process (1991-1996, 1996-2001 and 2001-2006). Our results show that in the second period, in coincidence with the strongest wave of the deregulation and concentration process, banks are systematically exiting from low income neighbourhood. In the two other periods, there is no evidence of such association. The result is robust across a number of different measures and specifications. Nevertheless, these dynamics, although marked, increase only slightly the positive relationship between the level of bank presence and the level of the average income, which is significant, but not very strong.

The paper also presents new methods to deal with point-pattern data. This kind of data are generally analysed in two alternative ways: i) by adapting them to a discrete spatial classification, generally by means of a simple event count, or ii) through Point Pattern Analysis techniques. In the first case, a significant "discretisation bias" is often introduced. In the second case, the information is more precise but it is difficult to interact with other data available with a discrete (administrative) spatial classification. Therefore, we develop three new measures of bank location and diversity which allow to keep a discrete classi-

fication of space, minimizing, at the same time, the discretisation bias. The measures are easy to calculate and can be applied in other contexts.

4.A Regressions using bank counts

This appendix reports the results obtained by using the simple counts of banks instead of the zonal statistic. Although the correlation among the two measures is quite high (see table 4.49), the results obtained using the simple count are overall much less significant, as expected.

We reach this conclusion after estimating the model of equation 4.24 with two alternative specifications. The first is a standard OLS, to ease comparability with the previous estimates. The second one accounts for the fact that with count variables OLS are inconsistent, and - also considering the large number of zeros present in the entry and exit variables - the most appropriate econometric model is a Zero Inflated Negative Binomial. In the latter case, the number of exits and entries are regressed on all the variables of the regression reported in table 4.41, while the zeros are a function of the zonal statistic at the beginning of the period and the area of the neighbourhood. The difference in the bank count between the end and the beginning of the period, instead, contains only few zeros; therefore, it has been linearly transformed to a non-negative variable (by adding four to all the observations), and is regressed by means of a Poisson model.

Results are reported in tables 4.50 and 4.51. In both the cases, the results obtained with the simple counts of banks are clearly less significant, which suggests that the dependent variable is considerably less precise. This is particularly evident when the dependent variables are the number of entries and exits (col. 2 and 3 of table 4.50 and col. 2 and 4 of table 4.51).

4.B Spatial econometric models of desert and choice variables

In this appendix we report the results of the regressions based on spatial econometric models of the specifications we discussed in section 4.2. The models we estimate are the spatial error and the spatial lag model described in section 4.1.4; to ease computation, we split the sample by each period - this, however, affects only the structure of the residuals and leaves unchanged the point estimates, which are therefore fully comparable with the corresponding OLS results reported in table 4.47. As it is possible to see in table 4.52, 4.53, and 4.54, the results do not change the general picture obtained from the OLS estimates. In particular, in the second period (1996-2001), the only one for which the coefficients on income are significant, the difference in the point estimates is minimal, especially as compared to the spatial lag model.

Table 4.49. Zonal statistic and bank counts: pairwise correlations

Pairwise correlation	Zon. st. flow	Zon. st. entry	Zon. st. exit
Δ bank count	0.74		
Entry count		0.66	
Exit count			0.71

Table 4.50. Regression of bank counts, OLS

COEFFICIENT Model	(1)	(2)	(3)
	Δ Bank count OLS	entry count OLS	exit count OLS
Average income 1991	0.421 (0.61)	0.522 (0.57)	0.125 (0.25)
Average income 1996	1.231*** (0.45)	0.236 (0.34)	-0.210 (0.31)
Average income 2001	0.666 (0.43)	0.300 (0.38)	-0.336 (0.41)
Zonal statistic	-0.802*** (0.14)	0.707*** (0.11)	1.436*** (0.10)
Active/tot pop. 1991	1.059 (1.33)	1.961** (0.97)	1.012 (1.07)
Active/tot pop. 1996	-3.054* (1.58)	-1.203 (0.78)	0.306 (0.91)
Active/tot pop. 2001	0.999 (1.31)	1.098 (1.02)	0.117 (1.12)
Tot. pop. (log)	0.0460 (0.034)	0.0479** (0.023)	0.0234 (0.029)
non Belgian/tot pop. 1991	1.110 (0.89)	0.516 (0.78)	-0.650 (0.52)
non Belgian/tot pop. 1996	0.534 (0.86)	-0.185 (0.61)	-0.404 (0.66)
non Belgian/tot pop. 2001	-0.202 (0.98)	-0.329 (0.65)	-0.171 (1.03)
elderly/tot pop. 1991	2.345* (1.25)	2.129* (1.10)	-0.0578 (0.83)
elderly/tot pop. 1996	0.0402 (1.30)	-0.788 (0.83)	-0.434 (0.87)
elderly/tot pop. 2001	1.026 (0.99)	0.691 (0.85)	-0.191 (0.92)
Centre dummy	0.103 (0.12)	0.0830 (0.096)	0.0360 (0.093)
Constant	-3.692 (3.90)	-4.978 (3.76)	-1.666 (1.70)
Observations	699	699	699
R^2	0.14	0.21	0.42

Heteroscedasticity robust standard errors in parentheses

Time, district, and time-district dummies included in all the specifications

*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Table 4.51. Regression of bank counts, zero inflated negative binomial

VARIABLES	(1)	(2)	(3)	(4)	(5)
Model	Δ Bank count Poisson	entry count ZINB	inflate ZINB	exit count ZINB	inflate ZINB
Average income 1991	0.112 (0.148)	1.482 (1.243)		0.531 (0.654)	
Average income 1996	0.316*** (0.110)	1.048 (1.433)		-0.424 (0.936)	
Average income 2001	0.177 (0.112)	0.845 (0.777)		-0.727 (0.599)	
Zonal statistic	-0.219*** (0.0381)	0.902** (0.408)	-13.86** (6.711)	1.388*** (0.175)	-13.67*** (2.796)
active/tot pop. 1991	0.283 (0.324)	6.678** (3.161)		3.007 (2.341)	
Active/tot pop. 1996	-0.780* (0.400)	-4.285 (3.768)		3.343 (2.551)	
Active/tot pop. 2001	0.280 (0.343)	2.468 (3.407)		1.279 (1.779)	
Tot. pop. (log)	0.0128 (0.00836)	0.293** (0.135)		0.156* (0.0800)	
non Belgian/tot pop. 1991	0.294 (0.213)	1.167 (2.371)		-0.744 (1.508)	
non Belgian/tot pop. 1996	0.142 (0.206)	0.582 (2.900)		1.652 (2.005)	
non Belgian/tot pop. 2001	-0.0532 (0.258)	-2.946 (2.368)		-0.281 (2.082)	
elderly/tot pop. 1991	0.615** (0.297)	6.987** (3.507)		1.625 (2.268)	
elderly/tot pop. 1996	0.0152 (0.319)	-0.452 (3.584)		3.042 (2.648)	
elderly/tot pop. 2001	0.283 (0.260)	0.643 (2.649)		0.186 (1.678)	
centre dummy	0.0277 (0.0314)	0.390 (0.323)		0.118 (0.204)	
area			3.46e-07 (4.35e-07)		-3.39e-07 (2.40e-07)
Constant	0.403 (0.944)	-17.82** (8.953)	3.195*** (1.102)	-7.826** (3.922)	3.317*** (0.792)
Observations	699	699	699	699	699

Heteroscedasticity robust standard errors in parentheses

Time, district, and time-district dummies included in all the specifications

*** p<0.01, ** p<0.05, * p<0.1

Table 4.52. Spatial regression desert and choice variables, 1991-1996

COEFFICIENT Spatial model	(1)	(2)	(3)	(4)
	Δ dist 1st bank ERROR	Δ dist 3 groups ERROR	Δ dist 1st bank LAG	Δ dist 3 groups LAG
Average income	-3.165 (12.4)	16.67 (11.5)	-3.018 (10.5)	10.75 (11.1)
Active/tot pop.	26.54 (27.3)	-20.48 (23.0)	24.25 (27.4)	-1.045 (20.6)
Tot. pop. (log)	-0.935 (1.97)	-0.399 (2.94)	-1.125 (1.93)	-0.802 (2.77)
non Belgian/tot pop.	-2.038 (18.4)	5.154 (16.1)	-7.943 (19.2)	-11.10 (20.5)
elderly/tot pop. 1991	0.454 (27.0)	-38.38 (27.8)	-9.007 (30.2)	-48.77 (32.3)
centre	-3.618 (3.33)	-7.600 (6.77)	-1.384 (2.49)	-0.477 (4.08)
Constant	28.98 (84.2)	-48.61 (59.8)	23.15 (73.3)	-41.79 (73.1)
Observations	233	233	233	233

Heteroscedasticity robust standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Table 4.53. Spatial regression desert and choice variables, 1996-2001

COEFFICIENT Spatial model	(1)	(2)	(3)	(4)
	Δ dist 1st bank ERROR	Δ dist 3 groups ERROR	Δ dist 1st bank LAG	Δ dist 3 groups LAG
Average income	-10.66* (5.71)	-35.06*** (12.5)	-14.04** (5.48)	-34.34*** (10.5)
Active/tot pop.	-12.74 (17.8)	86.24 (56.3)	-19.17 (21.5)	37.88 (44.3)
Tot. pop. (log)	-1.120 (1.14)	-4.413** (2.14)	-1.555 (1.30)	-5.542** (2.27)
non Belgian/tot pop.	-9.437 (15.0)	9.996 (28.7)	-17.52 (17.6)	0.975 (27.0)
elderly/tot pop. 1991	-22.32 (14.7)	45.46 (50.8)	-31.04 (20.1)	15.64 (45.7)
centre	-1.279 (3.54)	1.292 (7.50)	-1.586 (2.85)	-0.0244 (4.45)
Constant	81.43** (35.2)	149.1** (74.3)	106.8*** (38.6)	200.0*** (73.2)
Observations	233	233	233	233

Heteroscedasticity robust standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Table 4.54. Spatial regression desert and choice variables, 2001-2006

COEFFICIENT	(1)	(2)	(3)	(4)
Spatial model	Δ dist 1st bank ERROR	Δ dist 3 groups ERROR	Δ dist 1st bank LAG	Δ dist 3 groups LAG
Average income	-3.165 (12.4)	16.67 (11.5)	-3.018 (10.5)	10.75 (11.1)
Active/tot pop.	26.54 (27.3)	-20.48 (23.0)	24.25 (27.4)	-1.045 (20.6)
Tot. pop. (log)	-0.935 (1.97)	-0.399 (2.94)	-1.125 (1.93)	-0.802 (2.77)
non Belgian/tot pop.	-2.038 (18.4)	5.154 (16.1)	-7.943 (19.2)	-11.10 (20.5)
elderly/tot pop. 1991	0.454 (27.0)	-38.38 (27.8)	-9.007 (30.2)	-48.77 (32.3)
centre	-3.618 (3.33)	-7.600 (6.77)	-1.384 (2.49)	-0.477 (4.08)
Constant	28.98 (84.2)	-48.61 (59.8)	23.15 (73.3)	-41.79 (73.1)
Observations	233	233	233	233

Heteroscedasticity robust standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Chapter 5

Conclusions

This thesis has explored a number of different issues in urban economics. In the following, we summarize the main findings of the individual papers, discussing some methodological issues we faced while working on them. We also highlight the policy implications of our results.

Main findings and policy implications

The first paper investigates the role played by labour market factors in shaping the spatial concentration pattern of manufacturing industries in the United States. Until very recently, the importance of labour market factors has been stressed only theoretically, but few contributions have been able to provide strong empirical evidence on their importance (one notable exception is the work by Ellison et al., forthcoming). In the first paper, we develop an original methodology to disentangle the effects of labour market determinants of agglomeration in the pattern of spatial concentration of manufacturing industries in the United States. We were able to rank industries according to the relative importance of labour market factors in determining their level of spatial concentration. Furthermore, our results suggest that labour market determinants explain around one quarter of the pattern of concentration.

For policy-making, industrial concentration and specialisation are extremely relevant phenomena; together with spatial differentials in productivity, they arguably are the principal factor generating spatial inequalities. Regions and cities face different economic fates

because they are endowed with industries which differ widely in labour productivity, technological content, innovation potential, etc. Although neglected in traditional Solovian growth models (Solow, 1956), sectoral specialisation assumes a crucial role in many other theories of economic growth, e.g. in Kaldor (1968), Romer (1986), and Lucas (1988). According to Kaldor, intra-industry knowledge spillovers and learning-by-doing processes generate increasing return to scale within industries, while in the models of endogenous growth developed by Romer and Lucas (among others) the presence of an R&D sector specialised in the production of knowledge represents the key factor promoting growth.

The evidence that the characteristics of the local labour market have a strong association with industrial location brings a number of interesting policy implications: for instance, training or attracting skilled workers may not only enhance productivity, but also affect, in the long run, the industrial structure of a city or region, by attracting the most technologically advanced sectors.

Another interesting outcome of the research is the evidence that the industries for which labour markets seem to play a very important role in explaining agglomeration are very different each other: e.g., both "Guided missile and space vehicle manufacturing" and "Women's, girls' cut and sew dress manufacturing" are at the top of the ranking. This implies that the effects of labour market factors on concentration are more complex than one may think. The availability of high skilled individuals is not the only factor attracting the location of firms; also large pools of low-wage workers play an important role.

Finally, the paper also shows that approaching industrial agglomeration using traditional regression methods may be limiting. First, we highlighted some of the limits of the

Ellison and Glaeser's index (Ellison and Glaeser, 1997), which is now used in almost all the studies of concentration. These limits are the following:

a) the index takes into account only a few elements (namely the numerosity, the variance, and the mean) of the size distribution of plants and regions

b) The underlying assumption that the probability of a plant to locate in a given region is equal to the region aggregate share of employment may be questioned, as the size of plants may be endogenously determined by the industry pattern of concentration

c) The index may be biased because of the MAUP, as it needs to be calculated using discrete spatial units.

In the paper, we calculated a measure of industrial agglomeration which copes with all these issues. Rather than adding to the index further elements which may control for the sources of bias listed, we took an empirical approach, based on calculating the industry-specific bias with random simulations. Our results show sizeable differences with the Ellison and Glaeser index. To our knowledge, this is one of the few attempts to explore the statistical properties of the Ellison-Glaeser index and to question its validity (notable exceptions are Duranton and Overman, 2005, and Mori et al., 2005).

Second, the aforementioned evidence of the heterogeneity of industries for which labour market seems to play a very important role in explaining agglomeration casts doubts on previous approaches to industrial agglomeration based on parametric methods, as our results suggest that the labour market effect is generally hard to parameterize. This may make a case for different methods of research on the issue, including non-parametric econometrics or qualitative (case study) analyses.

The second paper assess whether the growth of cities in India is alleviating rural poverty in nearby areas, using a rich dataset on a panel of Indian districts in the period 1981-1999. The research question is particularly relevant in light of the massive wave of urbanisation in most developing countries, also considering that most of the world poor reside in rural areas, where the incidence of poverty is higher than in urban areas. India is especially suited to this kind of analysis as, with over 316 million of \$1/day rural poor in 2002, the country is home to 36% of the world's rural poor (the largest number in the world).

In assessing the effects of urbanization on rural poverty, we devote particular attention to disentangling first and second round effects, both theoretically and empirically. First round effects entail only a statistical association between urbanization and changes in rural poverty due to the change in residency of some rural poor (who may or may not be lifted out of poverty in their move to the urban areas). On the other hand, second-round effects capture the impact of the urban population growth on the rural rate of poverty. Such a relationship is causal in nature and tells us how good or bad urbanisation is for rural poverty.

Results show that the effect is significant and strong: we find that an increase of urban population by one fifth causes a decrease of between 3 and 6 percentage points in the share of rural poverty. However, these poverty reducing effects appear to apply mostly to rural poor closer to the poverty line. Although the very poor do not seem to be negatively affected by urbanization, they are not able to reap the benefits of such a growth.

These findings suggest a number of interesting policy implications.

First, it is commonly argued that most of the investments aimed at reducing poverty in developing countries should be addressed to rural areas, as most of the poor are concentrated there (e.g. World Bank, 2008). However, these investments may not always be efficient, as rural areas are often sparsely populated and therefore huge resources are needed in order to benefit only a small number of people. To the extent that investing in cities may also be effective in reducing poverty in rural areas, they may constitute a valid alternative or complement to rural projects.

Second, our findings counterbalance the widespread argument that rural-urban migration is generally harmful for source areas as it impoverish them, by taking away the most productive components of their population. The relatively low rate of urbanisation of India itself may also be due to public policies which have not facilitated (and in certain instances even constrained) rural-urban migration (Deshingkar and Start, 2003). On the contrary, the results of our paper suggest that rural-urban migration may bring also some positive effects for source areas. Considering that individuals involved may also be positively affected by this kind of migration, its overall effect on the wellbeing of inhabitants of rural areas needs probably to be reconsidered. More generally, the evidence discussed in the paper can foster a wider discussion on the benefit of migration toward urban agglomeration in the light of new economic geography, as well as "old" urban economics, theories. This is a crucial issue in the discipline, as policy makers often have a "schizophrenic" view on the issue: on one side, migration is seen as an important free market mechanism optimizing the supply of labour; on the other, it is a huge social cost widening spatial disparities, and impoverishing lagging areas.

Third, as we found some evidence that the benefits from urbanisation do not reach the very poor in rural areas, specific actions may be needed to enable these rural dwellers to enjoy the benefits of urbanisation. Examples of this may include developing the types of skills useful for an expanding urban sector; or the provision of capital to cover the fixed costs of rural-urban migration.

Some interesting considerations regarding spatial inequality may also stem from the analysis. Deaton and Jean (2002) highlight that urban-rural inequality has been rising in India in the '90s, although this is partly explainable by a rise in public sector wages. According to our results, a growing urban population mainly led by rural-urban migration may partly counteract the phenomenon, as positive spillovers of urbanization will reach rural areas, and a lower demographic pressure in rural areas may increase agricultural productivity of labour (and thus agricultural wages).

The third paper examines the geographical implications of a peculiar feature of patenting activity, i.e., the skewness of the distribution of patents across inventors. After identifying two categories of patents - stars and comets - based on the average productivity of their inventors, we find evidence supporting a number of interesting conclusions. First, once controlling for the overall concentration of patenting activity, stars and comets are unevenly distributed in space: in a few cities around 90% of granted patents are stars, while in others the share of comets accounts for 60% of the total. Second, concentrations of stars and comets are associated with cities with different structural characteristics. In particular, comets are associated with diversified economic structure, concentration of small plants,

and establishment births; while stars are more likely to be found in metropolitan areas with a large patenting activity and a specialized economic structure. Third, comets seem to be much more connected to the local economic environment than stars. Fourth, the activity of star inventors is beneficial to the activity of comet inventors: in our preferred specifications, we find that the elasticity of comet patents to star patents is approximately equal to 0.3, which means that, on average, a 10% increase in the number of star patents in a city leads approximately to a 3% increase in the number of comets in the same city.

The policy recommendations which result from our findings are manifold. Given that the effect of stars on the local economy is rather weak, the case for designing policies aimed at attracting stars may not be strong. However, we found that stars have a sizeable positive effect on comets, and the latter seem to be much more connected to the local economy. Therefore, concentrations of stars may have a positive indirect impact on the local economy, mediated through an increase in the number of comet patents. This, however, does not take into account the fact that stars and comets tend to concentrate in different cities. It follows that attracting stars where there are comets can be an unsuccessful policy, as "comets' cities" may be a sub-optimal location for stars.

From a methodological point of view, the paper stands as an example of an empirical strategy able to overcome classical identification issues. We assume that star patents have positive effects on the production of comets, but the estimation of these effects may be biased due to reverse causality or unobserved variables. We cope with that by building an instrumental variable for the number of star patents developed in a given city, technological category, and time period. The methodology is inspired by previous contributions

in regional and migration economics based on the so-called "shift-share" approach:⁴⁸ once the main regression controls for a sufficient array of individual fixed effects, a source of exogenous variation in the number of star patents in a given city and time period is given by the interaction of two factors: an historical presence of companies (or inventors specialized in certain sectors), and the contemporaneous variation in the average productivity level in other cities of inventors working in the same company (or sector) in other cities. The IV strategy may be applied in various other contexts where the number of star patents - or, generally, of all patents - is the endogenous variable. For example, we may be interested in assessing whether the number of patents is positively affecting the number of plant births, or the generation of new employment.

Many directions for further research stem from our analysis of the composite "patent universe". For instance, we may investigate whether some cities, or some institutions, are more supportive of knowledge spillovers from stars to comets than others. Also, in the paper we present some preliminary analysis suggesting that comets and stars have a rather different linkage to the local economy: it may be interesting to further explore the topic.

The fourth paper looks at urban economic phenomena at a much more detailed level, i.e., small neighbourhoods of around 2,000 inhabitants each. Specifically, we are interested in whether the intense and fast process of deregulation and concentration of the Belgian banking sector in the late '90s affected the geography of branch penetration in the city of Antwerp.

⁴⁸ Well known examples are Bartik (1991) and Blanchard and Katz (1992) for regional growth; and Ottaviano and Peri (2005) for migration studies.

We therefore identify three periods corresponding to distinct stages in the deregulation process (1991-1996, 1996-2001 and 2001-2006), and we quantify changes in the patterns of bank location across 233 neighbourhoods of the metropolitan area of Antwerp.

Our results show that in the second period, coinciding with the strongest wave of the deregulation and concentration, banks systematically exit from low income neighbourhoods. In the two other periods, there is no evidence of such association. The result is robust across a number of different measures and specifications. Nevertheless, these dynamics, although marked, only slightly increase the positive relationship between the level of bank presence and the level of the average income, which is significant, but not very strong.

One implication of our results is that, if the deregulation process proceeds further, poor neighbourhoods may lose most of their branches, and therefore it may become more difficult and infrequent for poor people to access mainstream financial services. In turn, this may entail a few other related negative outcomes. First, many low income people may renounce to open a bank account, with potential negative consequences on personal saving, budgeting, regular payments. Second, areas abandoned by traditional retail banking may be "colonized" by so-called "alternative service providers", i.e., money lenders or money transfer providers, which generally charge much higher fees, and may also promote less virtuous financial practices.

Finally, the paper also explores some methodological issues, related to the analysis of "point events" in spatial economics, which we discuss in the next section.

Methods with spatial data

The thesis exploits a number of innovative or recently developed tools in the discipline to deal with spatial data. For instance, the first paper highlights the informative content of an original empirical approach based on comparing results obtained with different spatial classifications. So far, spatial economic analyses has been traditionally based on i) choosing a specific geography, often arbitrarily and based on administrative or statistical subdivisions, and ii) testing the nature and significance of the statistical association among a number of variables using regression analysis. With this paper, we experimented with a different way to proceed. We first hypothesized that a given spatial association may be stronger at a specific spatial level for substantive reasons in the light of specific theoretical hypotheses, and then we tested whether these hypotheses were true by building an appropriate random counterfactual, based on a mix of Montecarlo simulations and advanced data elaboration through GIS. In this way, we were able not only to test whether our hypothesis were true, but also to calculate the statistical significance of our results. Although preliminary, the methodology may inspire new and original empirical strategies in spatial economic analysis.

In the first paper we also developed a new method to obtain an unbiased measure of the spatial concentration of manufacturing industries based on a discrete spatial classification, which controls for both the employment concentration bias and the discretisation bias. To our knowledge, we are the first to develop a measure with such properties. Technically, the measure builds on previous ideas on random zoning algorithms developed by Openshaw and coauthors (e.g. Openshaw and Taylor, 1979), and on Montecarlo counterfactuals

of industry plant location, in the spirit of the paper by Duranton and Overman (2005). Although conceptually simple and computable with publicly available data, the measure seems to significantly improve on the widely used Ellison and Glaeser (1997) index, as the latter is prone to the discretisation bias arising from the arbitrary aggregation of point events in space.

Summing up, from the first paper we learnt that zoning algorithms may be an extremely profitable resource for statistical inference in spatial economics. Given that the tool is almost totally unknown in the discipline, we think that our preliminary experiments in the field may be a good basis for further work on similar methods.

The second and the third paper also exploit some modern tools for the analysis of spatial data. In particular, in the paper on urbanization in India we make large use of GIS techniques to ease the matching of cities with the correspondent district using the available information on spatial location. More precisely, we geolocated almost all the 5,000 cities listed in the 2001 Indian Census, and then matched them with various shape files of the Indian administrative geography. Given that the latter presents a number of boundaries change over times, and information reported in the Census tables are not always reliable, the support of GIS software has meant higher precision of the matching, and remarkable time saving. Similar considerations hold for the paper on patents: in that case, given that the original dataset provides only the information on the name of the city and the State of residence of the inventors, without any further linkage to the US administrative geography (notably counties and MSAs), we geolocated all the cities, and subsequently we spatially

matched them with a shape file of US counties. In this way, we were able to obtain a dataset where almost all the patents are assigned to an US county.

Finally, the fourth paper, on the location of banks in the city of Antwerp, is also based on GIS techniques. While most of the data in spatial economics are traditionally collected at the level of discrete spatial units, point events data are identified by a unique pair of x-y coordinates. In the cases in which the two different kinds of data need to be employed together, a number of technical complications may arise; in most situations, scholars apply a discretisation to point events in order to homogenize them with other available data. However, the operation is not innocuous and may bring serious biases and measurement errors. In the specific case of our paper, the socio-economic data used for the analysis are collected at neighbourhood level, while the location of the branches is given by their x-y coordinates (as derived by their precise address). A common approach would be to "discretise" bank location by transforming the precise point events into a neighbourhood statistic (most likely a count), and to run the analysis at neighbourhood level. An alternative methodology would be to follow a "Point Pattern Analysis" approach by calculating a summary statistic based on the continuous space of the location pattern of banks. While the former strategy brings the aforementioned bias, the latter is also critical to the extent that a point statistic cannot easily be matched with other data available at neighbourhood level.

In the same paper we also presented two new measures of branch location. The first one is a measure of "bank desert", and corresponds to the distance which an hypothetical customer needs to cover to reach the first bank. The second measure aimed at capturing the level of "bank choice", and it is based on the distance which an hypothetical customer

needs to cover to find branches belonging to three different bank groups. Both the measures were calculated combining programming language with GIS software, and provide with an intuitive and precise metric of the issues we want to analyze in the paper. Furthermore, the two measures are neighbourhood-specific, but at the same time are calculated on the exact location of the point events (the branches) under investigation.

Therefore, in the paper we adopt an half-way solution, based on building original neighbourhood statistics - inspired by Point Pattern Analysis techniques - which are robust to the discretisation bias. By comparing our results with more traditional discrete measures, we are able show that our approach significantly improve on the the precision of the estimates.

Concluding remarks

In the preceding sections, we summarized the main findings, methodological issues, and policy implications of the four papers included in the thesis. Three general conclusions can be drawn from the previous discussion. First, as compared to other field of economics, spatial economics is an extremely rich discipline, which embraces a variety of different issues using a coherent and distinct set of analytical tools. To the extent that the variety of horizons feeds the debate and foster cross-fertilization and generation of new ideas, this can be a strong asset. However, there is also the concrete risk for the discipline of losing its identity - that is why some reasoning on the *distinctive* contributions of the subject is needed. Second, rigorous quantitative investigations can be extremely effective in informing the action of policy makers. This does not only relate to causal effects, but also to the analysis

of the spatial extent of economic phenomena, and to the degree of spatial association of socio-economic variables. Third, the discipline may be a fertile field for original contributions, experiments and innovations, especially related to empirical methods and data. The fascinating new data sources which are now available, especially remote sensing data and micro-geographic databases, offer to scholars in the field the opportunities for outstanding contributions.

References

- Acs ZJ, DB Audretsch MP Feldman, 1992. Real Effects of Academic Research: Comment. *American Economic Review*, 82:1
- Acs ZJ, DB Audretsch, 1990. *Innovation and Small Firms*, the MIT Press
- Aghion P., R. Burgess, S. Redding F. Zilibotti 2008. The Unequal Effects of Liberalization: Evidence from Dismantling the License Raj in India. *American Economic Review*, 98:4, pp. 1397–1412.
- Almeida, P., B. Kogut. 1999. Localization of knowledge and the Mobility of Engineers in Regional Networks. *Management Science* 45
- Amrhein C G, 1995. Searching for the Elusive Aggregation Effect: Evidence from Statistical Simulations. *Environment and Planning A* 27:1, pp 105–119
- Anselin L., 1988. *Spatial Econometrics: Methods and Models*. Dordrecht, Kluwer Academic Publishers.
- Anselin, L., R. J. Florax, 1995. Small Sample Properties of Tests for Spatial Dependence in Regression Models: Some further results. In Anselin, L. and Florax, R. J., editors, *New Directions in Spatial Econometrics*, p 21–74. Springer-Verlag, Berlin.
- Anselin, L., A. Bera, R. J. Florax, M. Yoon, 1996. Simple Diagnostic Tests for Spatial Dependence. *Regional Science and Urban Economics*, 26:77–104.
- Arbia G., 1989. *Spatial Data Configuration in Statistical Analysis of Regional Economic and Related Problems*. Springer.

- Arbia G., 2001. The role of Spatial Effects in the Empirical Analysis of Regional Concentration. *Journal of Geographical Systems*, Vol. 3 3, 271-281
- Audretsch DB, 2002. The Dynamic Role of Small Firms: Evidence from the US. *Small Business Economics*, 18:1-3
- Avery, R.B., R.W. Bostic, P.S. Calem, and G.B. Canner, 1999. Consolidation and Bank Branching Patterns. *Journal of Banking & Finance*, vol. 23, no. 2-4, pp. 497-532.
- Azoulay, P., J. S. Graff Zivin, and J. Wang, 2008. Superstar Extinction, NBER Working Paper No. 14577
- Balasubramanian N., J. Sivadasan, 2008. What Happens when Firms Patent? New Evidence from US Economic Census Data, working paper
- Bartik T.J., 1991. Who Benefits from State and Local Economic Development Policies? W.E. Upjohn Institute
- Baum, C., M. Schaffer, S. Stillman, 2008. IVREG2: Stata module for extended instrumental variables/2SLS and GMM estimation. Statistical Software Components from Boston College Department of Economics.
- Becker, G.S., K.M. Murphy, 1992. The Division of Labor, Coordination costs, and Knowledge. *Quarterly Journal of Economics*, 1074, pp. 1137-60.
- Berdegue, J.A., E. Ramirez, T. Reardon, and G. Escobar, 2001. Rural Nonfarm Employment and Incomes in Chile. *World Development*, 293, pp. 411-425.
- Bessen J, 2008. The value of U.S. Patents by Owner and Patent Characteristics. *Research Policy* 37:5

- Birthal, P. 2007. Linking Smallholder Livestock Producers to Markets: Issues and Approaches. Mimeo, Indian Society of Agricultural Economics.
- Breschi S., F. Lissoni, 2009. Mobility of skilled workers and co-invention networks: an anatomy of localized knowledge flows. *Journal of Economic Geography* 9 pp. 439–468
- Breschi, S. Lissoni, F., 2001. Knowledge Spillovers and Local Innovation Systems: a Critical Survey. *Industrial and Corporate Change*, 10: 975–1005.
- Briant A., P.P. Combes, M. Lafourcade, 2008. Does the Size and Shape of Geographical Units Jeopardize Economic Geography Estimations? CEPR Discussion Paper No. DP6928
- Cairncross F., 1997. *The Death of Distance: How the Communications Revolution Will Change our Lives*. Harvard Business School Press, Boston.
- Carlino G.A., S. Chatterjee, R.M. Hunt, 2007. Urban density and the Rate of Invention, *Journal of Urban Economics*, 61: 389–419
- Cerasi, V. Chizzolini, B., and M. Ivaldi, 2002. Branching and Competition in the European Banking Industry. *Applied Economics*, vol. 34, no. 17, pp. 2213-2225.
- Chakravarty, S.P., 2006. Regional Variation in Banking Services and Social Exclusion. *Regional Studies*, Vol. 40.4, pp. 415–428
- Chandra Mohan Reddy, Y.V.R. 2000. A Study of Livestock Markets and Marketing of Livestock in Rangareddy District of Andhra Pradesh, MSc Dissertation, Department of Agricultural Economics, Acharya N.G. Ranga Agricultural University, Hyderabad.
- Cheshire P.C., 1979. Inner Areas as Spatial Labour Markets: A Critique of the Inner Area Studies. *Urban Studies*, 16, 29-43

Cheshire P.C., D. Hay, 1989. *Urban Problems in Western Europe: An Economic Analysis*. Unwin Hyman

Combes P., H. G. Overman, 2004. *The Spatial Distribution of Economic Activities in the European Union*. In V. Henderson and J-F. Thisse (Eds.) *Handbook of Regional and Urban Economics*, vol. 4, Ch 64, pp 2845-2909. Elsevier

Coyle, D., 1999. *The Weightless World: Strategies for Managing the Digital Economy*, MIT Press

Damar, H.E. 2007. Does post-crisis restructuring decrease the availability of banking services? The case of Turkey. *Journal of Banking and Finance*, vol. 31, no. 9, pp. 2886-2905.

Datt, G., M. Ravallion, 1998. *Farm Productivity and Rural Poverty in India*. FCND Discussion Paper No. 42, International Food Policy Research Institute, Washington D.C.

Datt, G., M. Ravallion, 2002. Is India's Economic Growth Leaving the Poor Behind? *Journal of Economic Perspective* 16, 89–108.

Deichmann, U. F. Shilpi, R. Vakis, 2008. *Spatial Specialization and Farm-non-farm Linkages*. Policy Research Working Paper 4611: The World Bank.

Deshingkar, P., 2005. *Seasonal Migration: How Rural is Rural?*, ODI Opinions no. 52.

Deshingkar, P. and D. Start, 2003. *Seasonal Migration for Livelihoods in India: Coping, Accumulation and Exclusion*. ODI Working Paper 220.

Dore, R., 1987. *Taking Japan Seriously: a Confucian Perspective on Leading Economic Issues*. Stanford University Press: Stanford.

- Duranton G. and D. Puga, 2004. Microfoundation of Urban Agglomeration Economies. In V. Henderson and J-F. Thisse (Eds.) Handbook of Regional and Urban Economics, vol. 4, Helsevier
- Duranton G. and H. G. Overman, 2005. Testing for Localization Using Micro-Geographic Data. *Review of Economic Studies*. 72 4, 1077-1106
- Duranton G., D. Puga, 2005. From Sectoral to Functional Urban Specialisation. *Journal of Urban Economics*, 57:2, 343-370
- Duranton G., L. Gobillon, and H.G. Overman, 2006. Assessing the Effects of Local Taxation Using Microgeographic Data. CEP D.P. N. 748
- Ellis, F., 1998. Household Strategies and Rural Livelihood Diversification, *Journal of Development Studies*. Vol. 351: 1-38.
- Ellison G. and E. L. Glaeser, 1997. Geographic Concentration in U.S. Manufacturing Industries: a Dartboard Approach. *Journal of Political Economy*, 1997, 105, 5, 889-927
- Ellison, G.D., E.L. Glaeser, and W.R. Kerr, forthcoming. What Causes Industry Agglomeration? Evidence from Coagglomeration Patterns. *The American Economic Review*
- Ernault, J., Hemetsberger, W., Schoppmann, H., and C. Wengler, 2008. *European Banking and Financial Services Law*, Larcier, Gent.
- Eswaran, M. A. Kotwal, B. Ramaswami, and W. Wadhwa, 2008. How Does Poverty Decline? Suggestive evidence from India, 1983-1999, mimeo.
- European Central Bank, 2004. Report on EU Banking Structure. November. Available for download from <http://www.ecb.int/pub/>.

European Central Bank, 2008. EU Banking Structures. October. Available for download from <http://www.ecb.int/pub/>

European Commission, 2003. The internal market. Ten years without Frontiers. Memo 03/2, available from <http://ec.europa.eu/10years>.

Fafchamps, M. and F. Shilpi, 2003. The spatial Division of Labor in Nepal. *Journal of Development Studies* 39(6): 23-66.

Fafchamps, M. and F. Shilpi, 2005. Cities and Specialization: Evidence from South Asia. *Economic Journal*, Vol. 115, April: 477-504.

Foster, A. and M. Rosenzweig, 2004. Agricultural Productivity Growth, Rural Economic Diversity and Economic Reforms: India 1970-2000. *Economic Development and Cultural Change*, 52(3), p.509-542.

Fotheringham, A.S., and D.W.S. Wong, 1991. The Modifiable Areal Unit Problem in Multivariate Statistical Analysis. *Environment and Planning A* 23,1025-44

Friedman, T., 2005. *The World is Flat*, New York: Farrar, Strauss and Giroux

Fujita, M., P. Krugman, P., and A.J. Venables, 1999. *The Spatial Economy: Cities, Regions, and International Trade*, MIT Press: Cambridge and London.

Gehlke C. E., K. Biehl, 1934. Certain Effects of Grouping Upon the Size of the Correlation Coefficient in Census Tract Material. *Journal of the American Statistical Association*, 29, No. 185, Supplement: Proceedings of the American Statistical Journal, pp. 169-170

Glaeser, E.L., Kallal, H.D., J.A. Scheinkman, and A. Shleifer, 1992. Growth in cities, *Journal of Political Economy* 100, 1126-1152.

- Goddard, J., Molyneux, P, Wilson, J.O.S. and M. Tavakoli, 2007. European Banking: an Overview. *Journal of Banking and Finance*, vol. 31, no. 7, pp. 1911-1935.
- Government of India, various years. *Census of India*.
- Greenstone M, R Hornbeck, E Moretti, 2008. Identifying Agglomeration Spillovers: Evidence from Million Dollar Plants. NBER Working Paper n. W13833
- Griffith R., L. Sokbae, J. Van Reenen, 2007. Is Distance Dying at Last? Falling Home Bias in Fixed Effects Models of Patent Citations. NBER Working Paper 13338
- Griliches Z, 1990. Patent Statistics as Economic Indicators: A Survey. *Journal of Economic Literature*, 28: 4
- Gual, J. 1999. Deregulation, Integration and Market Structure in European Banking. *Journal of the Japanese and International Economies*, vol. 13, pp. 372-396.
- Hall, B.H., A. B. Jaffe, M. Trajtenberg, 2001. The NBER Patent Citations Data File: Lessons, Insights and Methodological Tools. NBER Working Paper 8498
- Hasan, R., D. Mitra and B.P. Ural, 2007. Trade Liberalization, Labor-Market Institutions and Poverty Reduction: Evidence from Indian States. *India Policy Forum* 3, pp. 71-122.
- Holmes and J.J. Stevens, 2004. The Spatial Distribution of Economic Activities in the North America. In V. Henderson and J-F. Thisse (Eds.) *Handbook of Regional and Urban Economics*, vol. 4, Helsevier,
- Holmes T., 1998. The Effect of State Policies on the Location of Manufacturing: Evidence from State Borders. *Journal of Political Economy*, 106, 4, 667-705

Holmes T., and J.J. Stevens, 2002. Geographic Concentration and Establishment Scale. *The Review of Economics and Statistics*. 84 4, 682-690

Ioannides Y., H.G. Overman, E. Rossi-Hansberg, and K. Schmidheiny, 2008, The Effect of Information and Communication Technologies on Urban Structure. *Economic Policy*, 23, pp 201-242

Isserman A.M., Westervelt J., 2006. 1.5 Million Missing Numbers: Overcoming Employment Suppression in County Business Patterns Data. *International Regional Science Review*

J.P. Elhorst, 2009. Spatial Panel Data Models. In Fischer M.M., Getis A. (Eds.) *Handbook of Applied Spatial Analysis*. Springer-Verlag, Berlin, Forthcoming.

Jacobs, J., 1969. *The Economy of Cities*, Random House: New York.

Jaffe A.B., 1989. Real Effects of Academic Research. *The American Economic Review*, 79:5, pp. 957-970

Jaffe A.B., M. Trajtenberg, 2005. *Patents, Citations, and Innovations: A Window on the Knowledge Economy*, The MIT Press

Jaffe A.B., M. Trajtenberg, and R. Henderson, 1993. Geographic localization of knowledge spillovers as evidenced by patent citations. *Quarterly Journal of Economics* 10, 577-598

Jha, V., 2008. Trickle down effects of Inter State Migration in a Period of High Growth in the Indian economy. Mimeo.

Kesteloot, C., Slegers, K, Vanden Broucke, L., Ippersiel, B., de Bethune, S. R Naiken, 2006. Dynamische Analyse van de buurten in moeilijkheden in de Belgische stadsgewesten, report for the Programmed Government Service PGS for Social Integration, November.

Kijima, Y. and P. Lanjouw, 2003. Poverty in India During the 1990s: a Regional Perspective. World Bank Policy Research Working Paper 3141.

Kochar, A. 2004. Urban Influences on Rural Schooling in India. *Journal of Development Economics* 74, pp. 113– 136

Kuznets, S., 1955. Economic Growth and Income Inequality. *American Economic Review*, 1, Vol. XLV.

Lafourcade M., G Mion, 2007. Concentration, Agglomeration and the Size of Plants. *Regional Science and Urban Economics* 37:1, pp. 46-68

Lanjouw, P. and A. Shariff 2002. Rural Nonfarm Employment in India: Access, income, and poverty impact. Working Paper Series no 81. New Delhi: National Council of Applied Economic Research.

Lewis, W. A., 1954. Economic Development with Unlimited Supplies of Labour. *The Manchester School*, 22: 139-191.

Leyshon A. and N. Thrift, 1996. Financial Exclusion and the Shifting Boundaries of the Financial System. *Environment and Planning A* 28 1150–6

Leyshon A., S. French, P. Signoretta, 2008. Financial Exclusion and the Geography of Bank and Building Society Branch Closure in Britain. *Transactions of the Institute of British Geographers*, 33 447-465

- Magrini S., 1999. The evolution of Income Disparities among the Regions of the European Union. *Regional Science and Urban Economics*, 29 2, 257-281
- Magrini S., 2004. Regional (Di)Convergence. In V. Henderson and J-F. Thisse (Eds.) *Handbook of Regional and Urban Economics*, vol. 4, Ch. 62, pp 2741-2796, Helsevier
- Manski, C.F., 1999. *Identification Problems in the Social Sciences*. Harvard University Press
- Marcon E. and F. Puech, 2003. Evaluating the Geographic Concentration of Industries using Distance-Based Methods. *Journal of Economic Geography*, 3 4, 409-428
- Marshall A., 1920. *Principles of Economics*. London: Macmillan and Co.
- Matsuyama, K., 1992. Agricultural Productivity, Comparative Advantage, and Economic Growth," *Journal of Economic Theory* 58, pp. 317-334.
- Maurel F. and B. Sedillot, 1999. A Measure Of the Geographic Concentration in French Manufacturing Industries. *Regional Science and Urban Economics*, 1999, 29 5, 575-604
- Nurske, R., 1953. *Problems of Capital Formation in Underdeveloped Countries*, Oxford University Press: New York.
- OMB Office of Management and Budget, 2000. Standard for Defining Metropolitan Statistical Areas; Notice. *Federal Register*, December 27th
- Openshaw, S. and P. Taylor, 1979. A Million or So Correlation Coefficients: Three Experiments on the Modifiable Areal Unit Problem. In N. Wrigley. Ed. *Statistical Applications in the Spatial Sciences*, 127-144, Pion, London.

Openshaw, S., 1977. A Geographical Solution to Scale and Aggregation Problems in Region-Building, Partitioning and Spatial Modelling. Transactions of the Institute of British Geographers, vol 2, pp. 459-72.

Parthasarathy Rao,P., P.S. Birthal, P.K. Joshi and D. Kar, 2004. Agricultural Diversification in India and Role of Urbanization, MTID Discussion Paper No. 77, IFPRI.

Peri G, 2005. Determinants of Knowledge Flows and Their Effect on Innovation, The Review of Economics and Statistics, vol. 87, issue 2, pages 308-322

Plantinga, A.J., Lubowski, R.N. and R.N. Stavins, 2002. The Effects of Potential Land Development on Agricultural Land Prices, Journal of Urban Economics 52, 561-81.

Quah, D., 1999. The Weightless Economy in Economic Development. CEP Discussion Paper No. 417

Ravallion, M., 2002. On the Urbanization of Poverty. Journal of Development Economics 68, 435-442.

Ravallion, M., S. Chen and P. Sangraula, 2007. New Evidence on the Urbanization of Global Poverty. World Bank Policy Research Working Paper 4199.

Ripley B. D., 1976. The Second-Order Analysis of Stationary Point Processes. Journal of Applied Probability, Vol. 13, No. 2 Jun., 1976, pp. 255-266

Rosenthal, S.S. and W.C. Strange, 2001. The Determinants of Agglomeration. Journal of Urban Economics, vol. 50:2, pp. 191-229

Rosenthal, S.S. and W.C. Strange, 2004. Evidence on the Nature and Sources of Agglomeration Economies. In V. Henderson and J-F. Thisse (Eds.) Handbook of Regional and Urban Economics, 2004, vol. 4, ch. 49, pp 2119-2171, Elsevier

Schaffer, M., 2007. XTIVREG2: Stata module to perform extended IV/2SLS, GMM and AC/HAC, LIML and k-class regression for panel data models, Statistical Software Components from Boston College Department of Economics

Silverberg G, B Verspagen, 2007. The Size Distribution of Innovations Revisited: an Application of Extreme Value Statistics to Citation and Value Measures of Patent Significance. *Journal of Econometrics*, 139:2

Silverman BW, 1986. *Density Estimation for Statistics and Data Analysis*. Chapman and Hall

Smith T.E., M.M. Smith, and J. Wackes, 2008. Alternative Financial Service Providers and the Spatial Void Hypothesis. *Regional Science and Urban Economics* 38:205–227

Stark, O. 1980. On the Role of Urban-Rural Remittances in Rural Development, *Journal of Development Studies*, 163: 369–74.

Stark, O., and R.E.B. Lucas, 1988. Migration, Remittances and the Family. *Economic Development and Cultural Change*, Vol. 36 3:465-482.

Steel, D.G., and D. Holt, 1996. Rules for Random Aggregation. *Environment and Planning A* 28, 957-78

Stock, J.H. and M. Yogo, 2005. Testing for Weak Instruments in Linear IV Regression, in D.W.K. Andrews and J.H. Stock (Eds.) *Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg*, Cambridge: Cambridge University Press, 2005.

Thanh, H.X., D.T., Thu Phuong, N. Thu Huong, and C. Tacoli, 2008. Urbanization and Rural Development in Vietnam's Mekong Delta, IIED Working Paper 14.

Topalova, P., 2005. Trade Liberalization, Poverty, and Inequality: Evidence from Indian Districts, NBER Working Paper No. W11614.

Trajtenberg M., G. Shiff, R. Melamed, 2006. The "Names Game": Harnessing Inventors' Patent Data for Economic Research, NBER Working Paper No. 12479

Udell, G.F., 1999. Comment on Avery, Bostic, Calem and Canner. *Journal of Banking and Finance*, vol. 23, no. 2-4, pp. 533-536.

United Nations, 2008. World Urbanization Prospects: The 2007 Revision Population Database, <http://esa.un.org/unup/>, accessed 23 May 2008.

von Thünen, J. H., 1966. *Isolated State: an English Edition of Der Isolierte Staat*. Pergamon Press, Oxford.

Waldinger F., 2009. Peer Effects in Science - Evidence from the Dismissal of Scientists in Nazi Germany. CEP Discussion Paper No 910

Wong D.W.S., 1997. Spatial Dependence of Segregation Indices. *Canadian Geographer / Le Géographe canadien*, 41:2, pp 128 - 136

World Bank, 2008. World Development Report 2008: Agriculture for Development. The World Bank: Washington DC.

Zenou Y., N. Boccoard, 2000. Racial Discrimination and Redlining in Cities. *Journal of Urban Economics* 48, 260–285

Zucker, L G. and M.R. Darby, 2007. Star Scientists, Innovation and Regional and National Immigration NBER Working Paper Series no. 13547

Zucker, L. G., M. R. Darby, , J. Armstrong, 1998. Geographically localized knowledge: spillovers or markets? *Economic Inquiry*, 36: 65–86.

List of tables

1.1	CBSA and PSA (average across 1000 its.) distributions: moments and percentiles	61
1.2	G Employment Concentration Index in the noise counterfactual, top 20 industries	70
1.3	CBSA minus NOISE, top 20 industries	71
1.4	CBSA - PSA, top 50 industries, 6-digit	72
1.5	EG index in the noise counterfactual	73
1.6	CBSA - PSA, 4-digit*, positive values	76
1.7	Regression of total agglomeration on labour market effect	77
1.8	Regression output	86
2.9	Urban population growth	100
2.10	Descriptive statistics of the main variables, 1981-99	117
2.11	The effects of urbanization on rural poverty across Indian districts, 1981-1999	121
2.12	The effects of urbanization on rural poverty across Indian districts, 1981-1993, OLS	124
2.13	The effects of urbanization on rural poverty across Indian districts, Further robustness	128
2.14	The effects of urbanization on rural poverty across Indian districts, Further robustness	130

2.15	The effects of urbanization on rural poverty across Indian districts, 1983-1999, IV Estimation, first stage	132
2.16	The effects of urbanization on rural poverty across Indian districts, 1983-1999, IV Estimation	133
2.17	The effects of urbanization on rural poverty across Indian districts, 1981-1993, IV Estimation	136
3.18	Period classification	149
3.19	Definition requirements	151
3.20	citations' shares, comets and stars	152
3.21	Regression of citations received	153
3.22	Patents by MSAs over total employment, rank correlation	154
3.23	Regression of comets/stars shares at MSA level	157
3.24	Regression of establishment births at MSA level	158
3.25	Summary statistics of stars and comets	169
3.26	regression of comet patents	170
3.27	Negative Binomial count regressions	173
3.28	Regressions with spatially lagged variables	176
3.29	Inventors' surname initial and patent authors' sequence	182
3.30	regression of comet patents, multi-author	183
3.31	regression of comet patents, alternative definitions, OLS	185
3.32	regression of comet patents, alternative definitions, IV	186

3.33	First stage regression	189
3.34	IV, overidentified regressions, 2SLS and LIML	190
3.35	citations' shares, comets and stars, within tech. category	191
3.36	Regression of citations received with tech. subcat. fixed effects	191
3.37	Regression of citations received excluding the top 5 per cent cited patents	192
3.38	Regression of comets/stars shares at MSA level, SUR	192
4.39	Summary statistics	218
4.40	Regressions of zonal statistic, flows	222
4.41	Regressions of zonal statistic, flows, further controls	223
4.42	Regressions of zonal statistic, flows, further controls and spatial lags	224
4.43	Regressions of zonal statistic, levels	225
4.44	Spatial diagnostics, p-values	227
4.45	Regressions excluding exits due to mergers	229
4.46	Regressions excluding exits due to mergers, further controls	230
4.47	regression of bank "desert" and bank choice	232
4.48	regression of bank "desert" and bank choice, further controls	233
4.49	Zonal statistic and bank counts: pairwise correlations	237
4.50	Regression of bank counts, OLS	238
4.51	Regression of bank counts, zero inflated negative binomial	239

List of tables

272

4.52	Spatial regression desert and choice variables, 1991-1996	240
4.53	Spatial regression desert and choice variables, 1996-2001	240
4.54	Spatial regression desert and choice variables, 2001-2006	241

List of figures

1.1	CBSAs and populated places	49
1.2	A stylized example	51
1.3	CBSA classification, Indianapolis area	55
1.4	PSA classification (single iteration), Indianapolis area	56
2.5	Indian towns (2001 Census)	93
2.6	Urban population growth and poverty reduction, by district 1981-99	95
3.7	Star and comet patents over employment, MSAs	193
3.8	Share of comets by MSAs	194
3.9	Share of stars by MSAs	194
4.10	Herfindahl index	200
4.11	The city of Antwerp: total polulation	203
4.12	The city of Antwerp: total polulation	204
4.13	The zonal statistic	210
4.14	Measure of "bank desert"	212