THE EMPIRICAL RELEVANCE OF GEOGRAPHICAL ECONOMICS

MAARTEN BOSKER

ISBN 978 90 393 4821 5

The empirical relevance of geographical economics

De empirische relevantie van geografische economie

(met een samenvatting in het Nederlands)

Proefschrift

ter verkrijging van de graad van doctor aan de Universiteit Utrecht op gezag van de rector magnificus, prof.dr. J.C. Stoof, ingevolge het besluit van het college voor promoties in het openbaar te verdedigen op vrijdag 30 mei 2008 des middags te 12.45 uur

door

Erik Maarten Bosker

geboren op 6 augustus 1980
te Leiderdorp

Promotor:	Prof.dr. J.H. Garretsen
Co-promotor:	Dr. M.C. Schramm

Voor mijn ouders

Dankwoord - Acknowledgements

At the end of my period as a PhD student, there are several people that I would like to thank explicitly. Without them this book may well have not been written.

First and foremost, I want to thank my promotor Harry Garretsen. Harry, you have been the best thesis supervisor I could have wished for, giving me all the freedom in following my own research interests and being always available and patient in hearing and commenting on my research ideas, doubts and plans. Without your help, criticism, encouragement and friendship, this thesis would certainly not have been what it is now. I am proud to have had you as my supervisor and hope to continue our collaboration for many years to come.

Second, I am very thankful to my co-promotor, Marc Schramm. Thank you for the many comments and ideas that have significantly improved this thesis. In some way or the other you always manage to point exactly at the weaknesses in my stories. Also, I will never forget pinpointing over five hundred Italian cities on a large-scale road map, nor your ability to forget to bring a jacket when visiting Toronto in mid-November.

I thank Waldo Krugell for inviting me over to South Africa and introducing me to the interesting and important issues that his country is facing. Chapter 4 of this thesis is the result, and hopefully only the start, of our collaboration.

Finally, I thank the members of the reading committee, Rob Alessie, Steven Brakman, Bernard Fingleton, Steve Redding and Keith Head, for reading and commenting on my thesis. Your feedback on my work has been a great stimulus to continue working in academia.

I also feel very lucky to have met a few people at work that made the last four years as a PhD student a lot more enjoyable.

Gerard Marlet and Clemens van Woerkens showed me that academic research does indeed have its practical use, be it slightly differently packaged and presented than in academia itself. Thank you for introducing me to the policy side of economic research, for the mind-clearing experience of always cycling against the wind along the Amsterdam-Rijnkanaal, and the free lunches at Broodje Martin.

An especially big thank you goes to Jordan Otten and Joppe de Ree. From the start of my work at USE, the many hours spent together drinking bakkies, listening to Barry White or Schnappi, watching the Tour and the World Cup on our pc's, playing indoor soccer, and discussing everything but economics, helped a great deal in putting academic research into perspective and made my time at USE a lot of fun.

Special thanks here go to Joppe. I have learned a lot from your persistent drive to get at the very bottom of every econometric issue you are facing, never to write anything down on the basis of empirical methods that you do not completely understand. Our long discussions have greatly improved the quality of my thesis, but have also led to many interesting, but academically questionable discoveries, such as a reinvention of several long-existing estimation strategies, trying to integrate out all kinds of error processes, and addressing the spatial nickell bias.

Ook wil ik deze gelegenheid gebruiken om enkele andere mensen te bedanken, niet zozeer voor hun directe bijdrage aan dit proefschrift, maar des te meer omdat ik hun vriendschap enorm waardeer en ze voor de nodige afleiding en vooral ook relativering van mijn onderzoek zorgen.

Dave, unfortunately Washington is not around the corner, the annual detour on the way back home makes my US conferences a whole lot more worthwhile. Bedankt Arjan, Joost, Frans, Bart en Victor, voor de vele avondjes bieren, de weekendjes Texel en Ardennen, en alle wielrenavonturen. David-Jan, voor de avondjes voetbal en onze discussies over het wel en wee van een bestaan binnen de wetenschap. Meinte, voor onze langdurige vriendschap, ik heb veel respect voor de manier waarop je in het leven staat. Ysbrand, voor de ‘twee handen op een buik’ en je onovertrefbaar goede slechte grappen. Jan-Willem, bedankt voor onze vriendschap en natuurlijk je fanatieke (maar vaak toch vergeefse) tennistegenstand. Niets beter dan een goede buur en beste vriend.

Bedankt oma, voor de soms bijna wekelijkse telefoontjes en de koffie en proviand tijdens mijn trainingsrondjes. Annieke, Hannebeth en Hans Rutger, voor de lol en gezelligheid in Oosterbeek en steeds vaker ook daarbuiten.

Papa en mama, aan jullie draag ik dit boekje op. Altijd staan jullie voor mij klaar met jullie steun, advies en liefde. Jullie zijn een groot voorbeeld voor mij, betere ouders kan ik me niet voorstellen.

Lieve Maud, je vroeg je zelf af waarvoor ik jou nu moest bedanken. Bedankt voor de vrolijkheid, en het gevoel van geluk en rust dat je me geeft. Op een of andere manier accepteer je me onvoorwaardelijk zoals ik ben. Bij jou voel ik me altijd thuis.

Maarten

THE EMPIRICAL RELEVANCE OF GEOGRAPHICAL ECONOMICS

MAARTEN BOSKER

Contents

Dankwoord – Acknowledgements i

Contents v

Introduction 1

Part I

Chapter 1 Adding geography to the New Economic Geography

1.1 Introduction 9

1.2 The Puga (1999) model 11

1.2.1 Setup of the model 11

1.2.2 Long run equilibrium and the degree of interregional labor mobility 14

1.2.3 Economic integration 16

1.3 Beyond an equidistant setup 17

1.3.1 The simulation setup 18

1.3.2 Introducing more realistic geography structures 20

1.3.3 Same overall degree of agglomeration – but different spatial distribution 23

1.3.4 The importance of the geography structure imposed 24

1.4 European integration when (relative) geography matters 26

1.4.1 Estimating the structural parameters 26

1.4.2 Simulating the impact of ongoing EU integration 27

1.4.2.1 Decrease in transport costs 28

1.4.2.2 Decrease in border impediments 31

1.5 Conclusions 32

Appendix 1.A Breakpoints of the Puga (1999) model 34

Appendix 1.B Combining distance and border effect 34

Chapter 2 Trade costs, market access and economic geography: Why the empirical specification of trade costs matters

2.1 Introduction 36

2.2 Trade costs and the wage equation in NEG 38

2.2.1 The basic NEG model 38

2.2.2 Estimating the wage equation 40

2.2.2.1 Direct non-linear estimation of the wage equation 40

2.2.2.2 Two-step linear estimation of the wage equation making use of trade data 41

2.3 The trade cost function 42

2.4 A third estimation strategy: implied trade costs 48

2.5 Data 51

2.6 Estimation results: trade costs and market access 51

2.6.1 Inferring trade costs from trade flows 52

2.6.1.1 Inferring the trade cost function's parameters from trade flows 52

2.6.1.2 Implied trade costs and trade cost functions 54

2.6.2 Varying trade costs and the impact on market access 56

2.6.3 Trade costs and the spatial reach of an income shock in Belgium 58

2.7 Conclusions 62

Appendix 2.A 63

Chapter 3 Economic geography and economic development in Sub-Saharan Africa

3.1 Introduction 64

3.2 The NEG model: income, trade costs and market access 66

3.3 Data set 69

3.4 Step 1: Trade estimation results 70

3.5 Step 2: Constructing market access and baseline results 75

3.5.1 Constructing market access from the trade estimations 75

3.5.2 The relevance of market access for GDP per capita in SSA: baseline results 78

3.6 Additional results: robustness checks, human capital and policy shocks 82

3.6.1 Robustness of the results 82

3.6.2 Human capital and market access 83

3.6.3 Policy experiments 86

3.7 Conclusions 88

Appendix 3.A Data definitions and sources 89

Part II

Chapter 4 Regional income evolution in South Africa after Apartheid

4.1 Introduction 92

4.2 A brief history of the South African spatial economy 93

4.2.1 The spatial economy up to 1994 93

4.2.2 The spatial economy after Apartheid 94

4.3 The evolution of the regional income distribution 95

4.3.1 The regional per capita income distribution 96

4.3.2 Markov chain analysis 98

4.4 The spatial evolution of the regional income distribution 102

4.4.1 Regionally conditioned Markov chain analysis 106

4.4.2 Spatial Markov chain analysis 109

4.5 Explaining the observed regional income distribution 111

4.6 Conclusions 112

Appendix 4.A 114

Chapter 5 A century of shocks: the evolution of the German city size distribution 1925-1999

5.1 Introduction 117

5.2 Data 119

5.3 The evolution of the city size distribution 120

5.3.1 Distribution characteristics 121

5.3.2 Distributional dynamics 123

5.4 Zipf's Law and Gibrat's Law of proportional effect 128

5.4.1 Individual city size evolution and Gibrat's Law of proportional effect 130

5.4.2 Nonparametric evidence on Gibrat's Law 130

5.5 Unit root testing and parametric evidence on Gibrat's Law 133

5.6 Unit root testing and WWII's impact on relative city size 137

5.7 Conclusions 141

Appendix 5.A Data 143

Appendix 5.B City characteristics and break stationarity 144

Conclusions 146

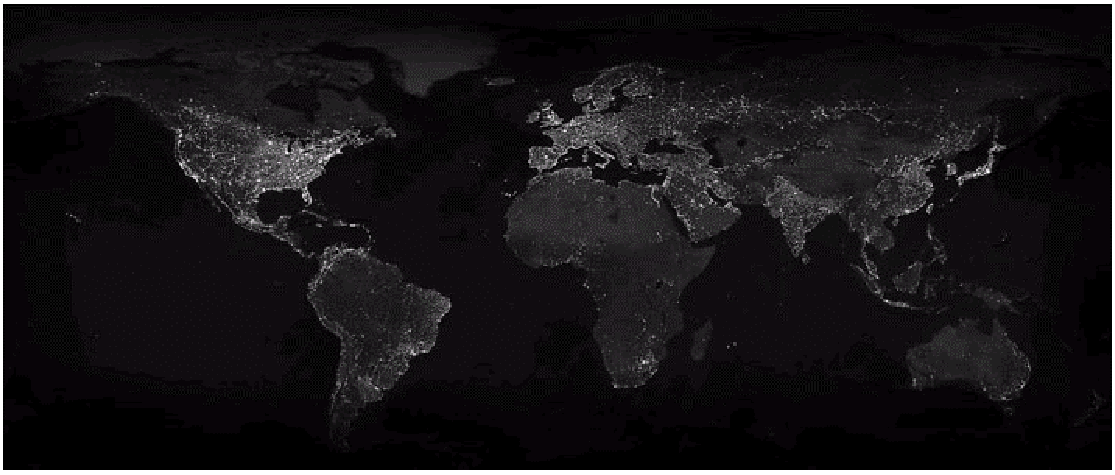
Nederlandse samenvatting 152

References 157

Introduction

The distribution of lights around the globe that is revealed by a satellite picture of the world at night, immediately suggests that economic activity is not evenly distributed across space.

Figure I.1 The world at night



Source: NASA

This is not only the case at the global scale, where in 2006 the US, Japan, the EU and Canada alone produced about 70% of total world output[1]. Also when 'zooming in' on the world's continents (compare North-West Europe to Southern Europe, or South Africa to the rest of Sub-Saharan Africa), countries (compare coastal to inland China or Paris to peripheral France), or even cities (New York's financial district is clearly different from Harlem as is Utrecht's Kanaleneiland from Wittevrouwen), the clustering of (specific types of) economic activity is hard to ignore.

Traditionally economic theory explains these spatial differences in economic development by location-specific factors. First nature geography, for example the presence of natural resources, climate, soil quality or the presence of natural harbors and navigable rivers, is one of the main candidates but also location-specific differences in technology or institutions have received considerable attention in the literature. Such factors can for example explain why deserts, dense jungles and the polar regions of the world are largely void of economic activity, why Singapore, Cape Town, or Rotterdam developed as large hubs of international trade, or why some countries specialize in exporting bananas (Ecuador or Costa Rica), while others in exporting opium (Afghanistan and Myanmar), gold (South Africa) or coffee (Brazil or Vietnam).

In effect, traditional economic theory explains the differences in economic development and the resulting trade patterns, by (exogenous) differences in endowments or in

[1] Source: World Economic Outlook database 2007, International Monetary Fund.

the ability of a specific location to produce more efficiently than others. As a result, it has a much harder time in explaining why two a priori similar locations can develop in totally different ways. Why did South-East Asia manage to dramatically raise its income level in the last twenty years, whereas most Sub-Saharan African countries (many of which are no more disadvantaged than their Asian counterparts when it comes to first nature geography) experienced decades of economic stagnation? Why did London become Europe's financial capital and not Amsterdam, Paris or Stockholm? Also it does not provide a satisfactory answer to the phenomenon that some places continue to thrive after having long lost the initial advantage that led to their economic rise (e.g. Johannesburg was founded on the site of one of the world's largest gold reserves, but nowadays it is far from dependent on natural resource extraction alone) whereas others fall into despair (e.g. the Mediterranean economy after the diversion of the major trade routes following the discovery of the direct route to the East by the Portuguese). Economic activity can be surprisingly sticky in some cases and surprisingly footloose in others. It may therefore seem quite surprising that until recently mainstream economists only paid attention to the first nature aspect of geography to explain the location of (different types of) economic activity and the resulting trade patterns.

The regional science and economic geography literature already did provide other explanations for the observed spatial inequality of economic activity than first nature geography, stressing the importance of the spatial interactions between regions, countries or cities[2] (also known as second nature geography), for their respective economic development. The main insight from these two strands of literature is that the observed spatial distribution of economic activity is the result of agglomeration and dispersion forces. Agglomeration forces are for example proximity to large markets, the presence of a large variety of goods that can be consumed, and the ease of finding a job or someone who does a job for you. Working against these are dispersion forces such as congestion, pollution, more competitors that sell the same good, and higher prices of immobile factors (such as space to live and produce on, but also immobile services are generally more expensive). It is the interplay of, and the relative changes in, these dispersion and agglomeration forces that determine whether people, and firms alike, tend to favor living in large agglomerations or instead prefer to live in a less crowded environment. Despite the important contributions in the regional science and economic geography literature (for an excellent review, see Ottaviano and Thisse, 2004), the role of second nature geography in the mainstream economic literature was for a long time ignored, mainly due to the inability of the standard economic models to deal with space in a coherent general equilibrium framework based on perfect competition (as exemplified by Starrett's 1978 Spatial Impossibility Theorem[3]).

[2] If not otherwise stated, I will use region to refer to any geographical unit of analysis (country, city, province, etc).

[3] This theorem states that if space is homogenous (i.e. each region is the same in terms of endowments, consumer preferences and firms' production possibilities) and transportation is costly, there does not exist a competitive equilibrium involving goods being traded between regions. Perfect competition combined with transport costs and homogenous space would result in so-called backyard capitalism: each region producing for itself (hereby avoiding to pay transport costs).

It was the influential paper by Krugman (1991) that finally put space back on the agenda of many mainstream economists. It gave rise to a renewed interest in second nature geography and quickly became known as New Economic Geography (NEG) or geographical economics[4]. The latter arguably being the more appropriate term given the aim of this strand of literature to put geography or space back on the agenda of economists instead of economics on the agenda of geographers (see Martin, 1999 for a critical view on the 'new' in new economic geography). Its appeal is that it combines insights from the earlier regional science and economic geography literature, most notably increasing returns to scale (so that firms have an incentive to produce in one place) and transport costs (so that it matters where you produce) into a coherent general equilibrium framework based on imperfect competition. The "trick" of using a market structure of imperfect competition allowed Krugman (1991) to "combine old ingredients through a new recipe" (Ottaviano and Thisse, 2004), modeling the distribution of economic activity, after controlling for first nature geography, as a trade-off of exactly those agglomeration and dispersion forces put forward in the earlier economic geography and regional science literature.

Since Krugman's seminal paper, NEG theory "has come of age" (Neary, 2001, p.536). Many models that build on Krugman (1991) using different and/or extra agglomeration and dispersion forces and different assumptions have been developed (e.g. Puga, 1999; Ottaviano, Tabuchi and Thisse, 2002; Baldwin, Martin and Ottaviano, 2001; Pflüger, 2004). Empirical work on the other hand has been lagging behind with only a few papers that explicitly base their empirical work on NEG theory, being published in the decade following NEG's conception in 1991 (see Head and Mayer, 2004 and Overman, Redding and Venables, 2003). As a result, both Neary (2001) and Krugman (1998) mention the lack of empirical evidence regarding the relevance of geographical economics as one of the major shortcomings of the NEG literature. Following the influential studies of Redding and Venables (2004), Hanson (2005), and Davis and Weinstein, (2002), this gap is however closing quickly with many valuable contributions being published since I started working on this thesis in 2004 (see e.g. Breinlich, 2007; Amiti and Javorcik, 2008; Knaap, 2006; Brakman, et al. 2006; Amiti and Cameron, 2007; Bosker, Brakman, Garretsen and Schramm, 2007a; Brakman, Garretsen and Schramm, 2004a,b; Crozet, 2004; Combes, Duranton and Gobillon, 2008).

Against this background, the aim of this thesis is to contribute to this growing body of empirical evidence and assess the importance of second nature geography in explaining the spatial distribution of economic activity. Each of the five chapters does so from a different point of view.

The first three chapters of this thesis (Part I) stay very close to NEG theory; the empirical backbone of each of these chapters is the estimation of a structural equation that follows directly from the most commonly used type of NEG models (i.e. the wage equation, focusing on the NEG proposition that market potential raises local factor prices, see Head and Mayer, 2004). The first two chapters discuss some of the problems one encounters when taking NEG to the data and the obtained estimation results back to NEG theory. Chapter 3 in

[4] From now on I will use geographical economics and NEG interchangeably, although I will mostly use NEG as a convenient abbreviation.

turn is fully concerned with an empirical application that tries to establish the importance of second nature geography for the countries in Sub-Saharan Africa.

The focus of the last two chapters (Part II) is somewhat different. They each take a closer look at the consistency of the observed spatial distribution of economic activity (and its evolution over time) with predictions made by NEG theory. Both chapters look at the effect of (large) shocks on the distribution of economic activity. Each chapter focuses on different predictions made by NEG theory, employing data at a different geographical level (regions and cities), and differing in the particular empirical method(s) used. Chapter 4 considers the effect of a shock to trade costs and labor mobility on the regional distribution of economic activity in South Africa and chapter 5 looks in more detail at the effect of a shock to economic activity itself by considering the effects (both in the short and in the long run) of the destruction suffered by West German cities during World War II.

Chapter 1 starts of Part I of this thesis, introducing the basic ideas underlying the theoretical NEG literature. It discusses a general NEG model (Puga, 1999) that encompasses two of the benchmark NEG models (Krugman, 1991 and Krugman and Venables, 1995). It describes the model's ability to explain the spatial distribution of economic activity endogenously (NEG's virtue compared to earlier economic geography models), focusing on, and partly extending, the model's main predictions regarding the relationship between the degree of spatial interdependency (the cost of trading goods between regions) and the location of economic activity. In particular, I discuss two conditions that the typical NEG model requires in order to remain analytically solvable. These two conditions, namely considering only two regions and/or a very simple (unidimensional) depiction of regions' spatial interdependency, can be argued to be problematic when doing empirical work based on NEG models. Virtually all empirical work done in new economic geography is based on multi-region datasets and a depiction of their interdependency that is not unidimensional but region-pair specific instead (based on e.g. bilateral distance, sharing a common border/language, etc). The common practice of comparing the resulting multi-region based estimates to the theoretical results/predictions from a two-region model can be highly misleading (see also Brakman, Garretsen and Schramm 2006). Next, I suggest a way to bridge this gap and argue that by conducting extensive simulations, one can get a very good idea of the behavior of the underlying NEG model even when considering many regions and a realistic depiction of their spatial interdependency. The results obtained from these simulation exercises allow the empirical researcher to much more adequately compare his/her estimation results with the underlying theory. They also show that many, but not all, of the insights obtained from the simple, but analytically solvable, two-region NEG models carry over to the multi-region case with a more realistic depiction of regions' interdependencies. The chapter ends with an illustration of the usefulness of combining structural empirical estimates with extensive simulation of the underlying NEG model, using a panel dataset of European regions and a depiction of their spatial interdependencies based on actual geographical information.

Chapter 2 builds on chapter 1 by focusing in detail on the way empirical work in NEG has dealt with the modeling of arguably the most important element of these models, namely regions' spatial interdependencies (i.e. trade costs between regions), in the absence of which

NEG would immediately lose its relevance. Ideally one would like to have information on the level of trade costs, that is on all costs incurred when shipping goods from one region to another, between all pairs of regions that are considered in a particular empirical exercise. Unfortunately these data are usually not readily available (see Hummels, 2007). Even data on transport costs and tariffs, seemingly easily measurable variables, are notoriously hard to come by, let alone information on other (more intangible) components of trade costs such as non-tariff barriers, administrative costs involved with trade, costs of delays that increase the time in transit, cultural and/or language barriers, and in the more obscure cases bribes to government officials, road blocks, etc. To still be able to model regions' spatial interdependencies in the absence of directly observable data, one usually assumes a so-called trade cost function (see e.g. Breinlich, 2007; Knaap, 2006; Redding and Venables, 2004; Hanson, 2005). That is, it is assumed that the level of trade costs between two regions depends on several (bilateral) characteristics that are observed such as the distance between the two regions, the quality of the two regions' infrastructure, membership of the same free trade area, language similarity, etc. The functional form of the assumed trade cost function subsequently specifies the way that trade costs depend on each of the observable components. Chapter 2 discusses how empirical studies in NEG have usually dealt with the specification of the trade cost function when estimating the 'workhorse equation' of the empirical NEG literature, the wage equation. It discusses what (implicit) assumptions one has to make when using such a trade cost function and introduces an alternative way to deal with the unavailability of trade cost data, based on the method proposed in Head and Ries (2001) that solely relies on the information revealed in bilateral trade flows to calculate so-called implied trade costs. A subsequent empirical example shows that the choice of trade cost approximation used nontrivially affects the conclusion(s) drawn about the relative importance of regions' spatial interdependencies when using two of the most common ways to estimate the NEG wage equation.

Chapter 3 leaves the discussion of how to estimate and interpret NEG models in chapters 1 and 2 behind and fully turns to an empirical application. More specifically, I (structurally) estimate an NEG model to be able to say something about the importance of second nature geography in determining the spatial distribution of economic development in Sub-Saharan Africa (SSA). The role of geography in explaining Sub-Saharan Africa's poor economic performance is usually confined to its physical geography that is widely viewed as being one of the important causes of the sub-continent's current dismal state of economic development (Gallup, Sachs and Mellinger, 1999; Collier and Gunning, 2001). Sub-Saharan Africa's climate makes many countries vulnerable to large shocks in agricultural production as a result of rainfall volatility (in the worst case resulting in widespread famine). It also makes it hospitable to many diseases such as malaria and dengue fever that heavily affect the infectant's ability to work productively. Another unfortunate aspect of Sub-Saharan Africa's physical geography is that it is void of navigable rivers, making the large-scale use of (relatively cheap) transportation over water impossible. Even the apparent blessing of the continent with the abundance of large deposits of natural resources (e.g. diamond, copper, gold and platinum) has turned itself into a curse (Sachs and Warner, 2001), resulting in corruption and even outright civil war. In chapter 3 I argue that, when considering Sub-Saharan Africa's economic development, the second nature aspect of geography is typically overlooked, whereas ample evidence exists that the sub-continent's ability to trade with the rest of the world is heavily impeded by the high level of trade costs involved in getting goods into and out of SSA countries. Using the estimation strategy introduced by Redding and Venables (2004), I first use trade data to reveal the relative importance of several trade cost measures in determining the strength of SSA countries' economic interdependency with other SSA countries and with the rest of the world respectively. Second, using the thus revealed degree of interdependency, I construct theoretically grounded measures of each Sub-Saharan African country's market access, that measures the ease with which a country can export its products to other countries (i.e. each SSA country's (dis)advantage in terms of second nature geography). It is these market access measures that are subsequently used to assess the impact of second nature geography on income per capita levels in the sub-continent. Explicitly grounding the construction of these market access terms on NEG theory furthermore allows the assessment of the effect of trade cost reducing policies such as infrastructure improvements or the establishment of regional or free trade agreements on economic development in Sub-Saharan Africa.

As briefly mentioned before, chapters 4 and 5 in Part II of this thesis constitute a break with the focus of the previous three chapters in Part I. Instead of estimating structural relationships that can be directly derived from NEG theory, these two chapters focus on the consistency of the observed (spatial) evolution of economic activity across regions with predictions made by NEG theory. On the one hand I lose the ability to directly link the resulting estimates back to an NEG model (indeed, results may sometimes also be consistent with alternative theories), on the other hand I gain 'more empirical freedom' in choosing appropriate (spatial) econometric methods.

More in detail, chapter 4 looks at the evolution of regional income levels in South Africa in the decade after the end of Apartheid in 1994. Besides ending years of racial injustice that heavily disadvantaged the black population in terms of almost all social and economic activities, the end of the Apartheid era marked two important changes from an economic geography point of view. First, the sanctions that were imposed on South Africa by the international community in response to South Africa's Apartheid-regime were lifted, opening up the country to international trade. Second, the end of Apartheid gave (economic) freedom to the largest part of South Africa's population that was suddenly free to move wherever it wanted instead of being confined to their so-called Homelands under the Apartheid-regime. NEG theory makes clear predictions about the effects of these two changes, the one being a decrease in trade costs and the other an increase in interregional labor mobility, on the location of economic activity (see Krugman and Livas-Elizondo (1996), Puga (1999) or chapter 1). On the basis of a large data set containing information on per capita income levels of 351 magisterial districts over the period 1994 – 2002, I provide a detailed characterization of the (spatial) evolution of regional income disparities in South Africa and verify the consistency of their evolution with the predictions from NEG theory. Instead of using a structural approach that is firmly based in NEG theory, as employed in chapters 1, 2, and 3, this is done by using (spatial) Markov chain techniques that characterize the evolution of the entire regional income distribution in terms of its intra-distributional dynamics. The use of these techniques provides a very clear picture of the (spatial) evolution of regional income inequalities in South Africa. The spatial Markov chain techniques in particular reveal interesting insights into the consistency of the observed evolution with some of NEG's predictions.

As chapter 4, chapter 5 also addresses the relevance of new economic geography models by carefully looking at the (in)consistency of the observed changes in the spatial distribution of economic activity in response to sudden changes in or large shocks to the economic system. Instead of focusing on the effect of a sudden change in the level of trade costs or the degree of interregional labor mobility as in chapter 5, this chapter builds on work by Davis and Weinstein (2002, 2005), Brakman, Garretsen and Schramm (2004a), Redding and Sturm, 2005 and Bosker, Brakman, Garretsen and Schramm (2007a) and looks at the response of an economic system to large shocks to the spatial distribution of economic activity itself. In particular I compare the evolution of the West German urban system over the period 1925-1999 with the theoretical predictions made by the NEG models regarding the stability of an established urban system when affected by large (exogenous) shocks. The West German case is of particular interest for this purpose as its urban system has been subject to two of history's largest urban shocks. First, the West German urban population suffered tremendously during World War II (WWII) as a result of the systematic Allied bombing raids aimed specifically at destroying Germany's urban centers and the subsequent heavy fighting when the Allied forces invaded Germany in 1944. Second, after World War II, the country was divided into East and West Germany and split by the Iron Curtain for more than 40 years until the fall of the Berlin Wall in 1989 and the subsequent German reunification in 1990. The use of several parametric as well as non-parametric empirical methodologies provides an in-depth analysis of the effect(s) and (relative) importance of these two shocks on the West German urban system. This, and in particular the evidence on the persistency of shocks in (relative) city size, allows me to compare the relevance of NEG-theories (or more generally theories based on increasing returns to scale) to two other dominant strands of urban economic theory that stress random city growth or the importance of locational fundamentals respectively.

Finally, the concluding chapter summarizes the main findings of my thesis. It focuses in particular on what can (and cannot) be said about the relevance of geographical economics on the basis of the empirical evidence provided in each of the chapters. Also, it reflects on potential fruitful areas of future empirical research, in my view mainly in the micro-, institutional and historical dimension, that will help in furthering our understanding of why, how and when second nature, or relative location, matters for economic development.

Part I

Chapter 1

Adding geography to the New Economic Geography[5]

1.1 INTRODUCTION

The seminal paper by Krugman (1991) gave rise to what became known as the new economic geography (NEG) literature, where the 'new' refers to the fact that the spatial distribution of population, production and consumption emerges endogenously from full general equilibrium models (Fujita, Krugman and Venables, 1999a; Helpman, 1998; Krugman and Venables, 1995; Venables, 1996; or Puga, 1999). Any particular spatial distribution of economic agents goes along with spreading (e.g. congestion, housing prices) and agglomeration (e.g. better market access, good access to intermediate suppliers) forces that dis- or encourage agglomeration respectively. The interplay of these two forces determines whether or not agents move to another region, hereby possibly changing the spatial economic landscape. As a result, the theoretical models are able to give predictions about the effect of a change in these spreading or agglomeration forces, for example the effect of lowered trade costs, on the distribution of economic activity across space (see also Fujita and Thisse, 2002).

These predictions are, however, typically based on models that treat geography in a very simple way (see also Neary, 2001, p.551). Attention is largely confined to simple 2-region models or multi-region models exhibiting a unidimensional spatial structure, with regions lying on a circle, i.e. a racetrack economy (Fujita et al., 1999a ch.6), with the distance between each pair of regions the same, i.e. an equidistant economy (Puga, 1999, Tabuchi et al., 2005) or regions lying on a line, i.e. a line economy (Fujita, et al., 1999b). A few papers, notably Krugman and Livas Elizondo (1996) and Monfort and Nicolini (2000), introduce 3-region models that allow for differences in the cost of intranational and international trade, but in order to keep these models tractable the authors have to assume the economic mass of the third region to be exogenous. A notable exception is the paper by Behrens, et al. (2005), which presents analytical results in a multi-region trade model with a somewhat more complex characterization of geography, i.e. a transportation network locally described by a tree, showing that in that case changes in transport costs have spatially limited effects.

The reason for making these simplifying assumptions is analytical tractability. Adding a more realistic, asymmetric, geography structure to an NEG model would render the model analytically insolvable (see Behrens et al., 2005, p.16 and Fujita and Mori, 2005, p.396). It is the assumption of a simple geography structure and/or the focus on a 2-region model that allows for the establishment of all the well-known analytical results in the NEG literature (e.g. multiple equilibria, catastrophic (de)agglomeration, etc).

However, when doing empirical or policy work, these simplifying assumptions become problematic since it is unclear whether the conclusions from these simple models

[5] This chapter is an adapted version of Bosker, Brakman, Garretsen and Schramm (2007b).

carry over to the more heterogeneous asymmetric geographical setting faced by the empirical researcher or policy maker (see also Behrens and Thisse, 2007). For empirical work, this makes it difficult to relate the underlying theory to empirical estimates of the structural model parameters obtained using multi-region or multi-country data, as e.g. presented in Redding and Venables (2004), Hanson (2005) and Brakman et al. (2006). When doing policy work, it becomes ambiguous to provide policy recommendations for the clearly asymmetric multi-region setting in the 'real world' on the basis of an equidistant (often 2-region) model. To study the effects of increased economic integration in such a 'real-world'-setting, attempts by e.g. Forslid et al. (2002a), Forslid et al. (2002b) and Bröcker (1998) all resort to the simulation of a computable general equilibrium (CGE) model of an asymmetric multi-region and/or multi sector world. These simulations give some interesting predictions about the effect of the ongoing economic integration in the EU. But the results obtained are difficult to link back to the theoretical model because the properties of the CGE-models that are used for the simulations are generally not known, not even for the simple 2-region or equidistant multi-region case.

This chapter takes a more theoretically grounded approach by adding more geographical realism to a well-known NEG-model (Puga, 1999) that encompasses several benchmark NEG-models. A particularly nice feature of that model is that it presents analytical results for both the 2-region and the equidistant multi-region setting, which serve as the theoretical benchmark to which we[6] can compare our findings. In doing so, we follow the recommendations made by Fujita and Krugman (2004), p.158; Behrens and Thisse (2007), section 3; Krugman (1998), p.15 or Fujita and Mori (2005), and it is useful to quote the latter paper at some length:

> *"While it will continue to be important to pursue building analytically solvable models regarding the basic mechanism of agglomeration and dispersion, it will become even more important to build numerically computable models. After all, there is great need to finally go beyond the basic two-region-two-industry models and go to asymmetric many-region-many-industry models of trade and geography in order to attain practically useful policy implications. Most models with emphasis on the analytical solvability are solvable only in a very limited low dimensional setup, but they are often not computable numerically (at least not in a reasonable amount of time) once more spatial and industrial structure is incorporated. A most desirable model would be one that has solvability at the low dimensional setup and computability even at the fairly high dimensional setup." (Fujita and Mori, 2005, p.396)*

The only other paper that we know of that simulates a NEG model when adding a more realistic depiction of geography is Stelder (2005). Using the Krugman (1991) model, Stelder (2005) tries to replicate the actual spatial distribution of cities across Europe by simulating the

[6] In contrast to the introduction and the conclusion, I will use the first person plural in chapter 1 – 5. Given that I have worked on each of the chapters with one or more co-authors, I find using the 'we form' in these chapters no more than appropriate.

Krugman (1991) *cum* geography model. He however does not relate any of his model simulations back to theory, focusing instead on simulating the current spatial distribution of economic agglomerations as closely as possible. Our aim is quite different: we systematically show the impact of introducing several asymmetries (asymmetric 2nd nature geography structure(s), asymmetric initial endowments) to a multi-region NEG model that is analytically solvable when considering its equidistant version.

This chapter is organized as follows. In the next section we introduce the Puga (1999) model. In section 1.3 we introduce our depiction of geography. To restrict our attention to the introduction of more realistic geography structures, we deliberately assume that all regions are initially of equal size or mass. The introduction of non-equidistant regions does by and large not change the qualitative results from the benchmark 2-region (or equidistant multi-region) model with respect to the impact of a change in trade costs on the equilibrium degree of agglomeration. With interregional labor mobility, a continued fall in trade costs will ultimately, and 'catastrophically', lead to complete agglomeration. Without interregional labor mobility, moving from very high to very low trade costs will also initially result in sudden (or partial) agglomeration but when trade costs become very low the degree of agglomeration will decrease resulting in a sudden (or partial) return to more spreading. A notable difference between the long run equilibria in our non-equidistant, multi-region model and the equidistant Puga (1999) model is that the same long run equilibrium level of agglomeration may go along with a *different* spatial distribution of economic activity across the individual regions. In section 1.4 we add the role of economic mass to our model by not only allowing for an asymmetric geography structure but by also taking differences in initial size (employment, arable land area) into account. Again, in a qualitative sense the main results from the equidistant model, with respect to the impact of changes in trade costs on the long run equilibrium level of agglomeration, carry through. In addition, and by using estimation results of the key NEG model parameters for a sample of 194 European NUTSII regions, we are in a position to answer questions about the impact of increased EU-integration. We find that lowering trade costs will most likely imply more and not less agglomeration for the EU regions. Moreover, both the extent and the spatial pattern of agglomeration depend crucially on the assumption about interregional labor mobility. Section 1.5 concludes.

1.2 THE PUGA (1999) MODEL

1.2.1 Setup of the model

This section provides a brief description of the model introduced by Puga (1999). As mentioned before we use this model as it captures two important benchmark NEG-models, i.e. Krugman (1991) and Krugman and Venables (1995) as special cases. Also Puga (1999) derives analytical results in the 2-region case as well as in the equidistant multi-region case, which allows for a ready comparison to our simulation results. The model set up is as follows[7]. Consider a world consisting of M regions, each populated by L_i workers and

[7] We use the same notation as Puga (1999) for ease of exposition.

endowed with K_i units of arable land ($i = 1,...,M$). Each region's economy consists of two sectors, agriculture and industry. Labor is used by both sectors and is perfectly mobile between sectors within a region and is either perfectly mobile or immobile between regions. Land on the other hand is used only by the agricultural sector; all land is always in use for agricultural production[8], and is immobile between regions.[9]

Production

The agricultural good is produced under perfect competition and free entry and exit using Cobb-Douglas technology[10] and is freely tradable between regions. The industrial sector produces heterogeneous varieties of a single good under monopolistic competition and free entry and exit, incurring so-called 'iceberg' trade costs when shipped between regions: $\tau_{ij} \geq 1$ goods have to be shipped from region *i* to let one good arrive in region *j* (the rest 'melts away' in transit, hence the name iceberg transport costs)[11]. Industrial production technology is characterized by increasing returns to scale, i.e. production of a quantity *x(h)* of any variety *h* requires fixed costs α and variable costs $\beta x(h)$ that are assumed to be the same in each region. This, together with free entry and exit and profit maximization, ensures that in equilibrium each variety is produced by a single firm in a single region. The production input is a Cobb-Douglas composite of labor and intermediates in the form of a composite manufacturing good, with $0 \leq \mu \leq 1$ the Cobb-Douglas share of intermediates. The composite manufacturing good is specified as a CES-aggregate (with $\sigma > 1$ the elasticity of substitution across varieties) of all manufacturing varieties produced. The resulting minimum-cost function associated with the production of a quantity *x(h)* of variety *h* in region *i* can be written as:

$$C(h) = q_i^{\mu} w_i^{M\,1-\mu}(\alpha + \beta x(h)) \tag{1.1}$$

where q_i is the price index of the composite manufacturing good, and w_i^M the manufacturing wage in region *i*.

Preferences

All consumers have Cobb-Douglas preferences over the agricultural good and a CES-composite (also with $\sigma > 1$ the elasticity of substitution across varieties) of manufacturing varieties, with $0 \leq \gamma \leq 1$ the Cobb-Douglas share of the composite manufacturing good. Specifying the composite manufacturing good in this way ensures demand from each region

[8] There is no other use for land. As a result there exists no active land market with landowners deciding to supply their land for agricultural production or put it to some other use; they are basically price takers that are always willing to put all their land to agricultural use.

[9] Defining the two sectors as being agriculture and industry is arbitrary. The main point is that one sector employs an immobile (both between sectors and regions) factor of production, producing a homogeneous good that is freely tradable between regions under perfect competition, and that the other sector employs a mobile (be it between sectors and/or regions) factor of production, producing heterogeneous varieties of the same good that are costly to trade between regions under monopolistic competition.

[10] Puga (1999) defines the agricultural sector somewhat more general. However, when deriving analytical results, he also resorts to the use of a Cobb-Douglas production function in agriculture, see p.318 of his paper.

[11] This is of course a very simplistic depiction of transport costs. It is however very useful from a modelling perspective, as it avoids having to fully specify a transport or logistics sector.

for each manufacturing variety, which, together with the fact that each variety is produced by a single firm in a single region, implies that trade takes place between regions.

Equilibrium

Having specified preferences over, and the production technologies of, the manufacturing and agricultural good, the equilibrium conditions of the model can be calculated. In the agricultural sector, profit maximization and free entry and exit determine the share of labor employed, L_i^A, the wage level w_i^A in agriculture, which equals the marginal product of labor, and the rent earned per unit of land $r(w_i^A)$[12]. The former two in turn pin down the share of workers in manufacturing, ς_i. Given the assumed Cobb-Douglas production function in agriculture, with labor share *θ*, we have that:

$$\varsigma_i = \frac{L_i^M}{L_i} = 1 - \frac{L_i^A}{L_i} = 1 - \frac{K_i}{L_i}\left(\frac{\theta}{w_i^A}\right)^{\frac{1}{1-\theta}} \tag{1.2}$$

where $0 \leq \theta \leq 1$ denotes the Cobb-Douglas share of labor in agriculture, and L_i^M and L_i^A the number of workers in manufacturing and agriculture respectively. Equation (1.2) shows that, in contrast to Krugman (1991), where agriculture uses only land[13] ($\theta = 0$), or to Krugman and Venables (1995), where agriculture employs only labor ($\theta = 1$), the share of a region's labor employed in manufacturing is endogenously determined in this model. It increases with a region's labor endowment and agricultural wage level and decreases with a region's land endowment and with the Cobb-Douglas share of labor in agricultural production. Consumer preferences in turn determine total demand for agricultural products in region *i* as:

$$x_i^A = (1-\gamma)Y_i \tag{1.3}$$

where Y_i is total consumer income (see also (1.9) below).

In the industrial sector, utility maximization on behalf of the consumers, combined with profit maximization and free entry and exit, gives the familiar result that all firms in region *i* set the same price for their produced manufacturing variety as being a constant markup over marginal costs:

$$p_i = \frac{\sigma\beta}{\sigma-1} q_i^{\mu} w_i^{M\,(1-\mu)} \tag{1.4}$$

where q_i is the price index of the composite manufacturing good in region *i* defined by:

$$q_i = \left(\int_j \tau_{ij}^{1-\sigma} n_j p_j^{(1-\sigma)}\right)^{\frac{1}{1-\sigma}} \tag{1.5}$$

where n_i denotes the number of firms in region *i* and

$$w_i^M = \left[(1-\mu)n_i p_i\left(\frac{(\sigma-1)}{\sigma\beta}(\alpha+\beta x_i)\right)\right](\varsigma_i L_i)^{-1} \tag{1.6}$$

is the manufacturing wage level in region *i*.

[12] Given the absence of an alternative use of land, the land rent implicitly only depends on the wage rate offered to workers, which determines how many people are willing to work in agriculture and thus the demand for land.

[13] Krugman (1991) does not call this immobile production factor land; he refers to it as being immobile labor, i.e. farmers.

It also gives total demand for each manufacturing variety produced (coming from both the home region *i* as well as foreign regions *j*) which is the same for each variety in the same region due to the way consumer preferences are specified:

$$x_i = \int_j p_i^{-\sigma} e_j q_j^{(\sigma-1)} \tau_{ij}^{1-\sigma} \tag{1.7}$$

In (1.7) demand from each foreign region *j* is multiplied by τ_{ij} because (τ_{ij} –1) of the amount of the product ordered from region *i* melts away in transit (the iceberg assumption), and

$$e_i = \gamma Y_i + \mu n_i p_i \left(\frac{(\sigma-1)}{\sigma\beta}(\alpha+\beta x_i) \right) \tag{1.8}$$

is total expenditure on manufacturing varieties in region *i* (the first term representing consumer expenditure and the second term producer expenditure on intermediates), where

$$Y_i = w_i^A (1-\varsigma_i) L_i + w_i^M \varsigma_i L_i + r(w_i^A) K_i + n_i \pi_i \tag{1.9}$$

is total consumer income consisting of workers' wage income, landowners' rents and entrepreneurs' profits respectively. Due to free entry and exit these profits are driven to zero ($\pi_i = 0$), thereby uniquely defining a firm's equilibrium output at:

$$x_i = \alpha(\sigma-1)/\beta \tag{1.10}$$

Finally, to close the model, the labor markets are assumed to clear:

$$L_i = L_i^M + L_i^A = \left[(1-\mu) n_i p_i \left(\frac{(\sigma-1)}{\sigma\beta}(\alpha+\beta x_i) \right) \right] \left(w_i^M \right)^{-1} + K_i \left(\frac{\theta}{w_i^A} \right)^{\frac{1}{1-\theta}} \tag{1.11}$$

where the demand for labor in agriculture, L_i^A, follows from the assumption of Cobb-Douglas technology in agriculture and the term between square brackets represents the total manufacturing wage bill. Moreover equating labor supply to labor demand in the industrial sector gives an immediate relationship between the number of firms and the number of workers in industry:

$$n_i = \frac{\varsigma_i L_i}{\alpha\sigma(1-\mu) q_i^{\mu} w_i^{M-\mu}} \tag{1.12}$$

1.2.2 Long run equilibrium and the degree of interregional labor mobility

Next, to solve for the long run equilibrium (LRE)[14], Puga (1999) distinguishes between the case where labor is both interregionally and intersectorally mobile and the case when it is only intersectorally mobile. Without interregional labor mobility, long run equilibrium is reached when the distribution of labor between the agricultural and the industrial sector in each region is such that wages are equal in both sectors. This is ensured by labor being perfectly mobile between sectors driving intersectoral wage differences to zero. When instead labor is also interregionally mobile, not only intersectoral wage differences are driven to zero in all regions in equilibrium. Workers now also respond to real wage (utility) differences

[14] The use of long run and short run is quite common in the NEG literature. Note however that all NEG models are essentially static and equilibrium is reached instantaneously so that the terminology long run and short run can be somewhat confusing to the reader not familiar with the literature. The distinction between the short and the long run is only introduced to help in getting the intuition behind the model.

between regions by moving to regions with the higher real wages (utility) until real wages are equalized between all regions, hereby defining the long run equilibrium.[15] In effect, the model (and its two variants) can be summarized by the following scheme or decision tree[16]:

Table 1.1 Model outline to find long run equilibrium (LRE)

a. Initial distribution of labor over regions and over sectors within each region

b. Labor moves between sectors within each region until sectoral wages are equal.

c. Interregional labor mobility?

> **C1.** NO: long run equilibrium
>
> **C2.** YES: short run equilibrium → **d.**

d. Interregional real wage equality?

> **D1.** NO: labor moves between regions in response to differences in real wages, with workers moving to those regions with higher real wages, hereby changing the distribution of labor over the regions → process restarts at **a.** with this new distribution of labor over regions and sectors.
>
> **D2.** YES: long run equilibrium

Interregional labor immobility

The long run equilibrium in case of interregional labor immobility can now be shown to be a solution *{w*$\{w_i, q_i\}$ of three equations that have to hold in each region. In our case (when using wage-worker space) these are, using the fact that in equilibrium $w_i^M = w_i^A = w_i$:

$$q_i = \frac{\sigma\beta}{\sigma-1}\left(\frac{1}{\alpha\sigma(1-\mu)}\sum_j\left(\varsigma_j L_j q_j^{-\mu\sigma} w_j^{1-\sigma(1-\mu)} \tau_{ij}^{1-\sigma}\right)\right)^{1/(1-\sigma)} \qquad (1.13)$$

$$w_i = \left(\frac{\sigma\beta}{\sigma-1}\right)^{\mu-1} q_i^{\mu/(\mu-1)}\left(\frac{\beta}{\alpha(\sigma-1)}\sum_j e_j q_j^{\sigma-1}\tau_{ij}^{1-\sigma}\right)^{1/(\sigma(1-\mu))} \qquad (1.14)$$

$$e_i = \gamma(w_i L_i + K_i r(w_i)) + \mu/(1-\mu) w_i \varsigma_i L_i \qquad (1.15)$$

, where (1.13) is obtained by substituting (1.4) and (1.12) into (1.5), (1.14) by substituting (1.4) and (1.10) into (1.7), and (1.15) by substituting (1.4), (1.10) and (1.12) into (1.8).

Interregional labor mobility

In case of interregional labor mobility, a solution to (1.13)-(1.15) merely constitutes a short run equilibrium (SRE). With interregional labor mobility, workers will move between regions in response to real wage differences until the interregional real wage differences, that are possible to persist when workers are unable (or unwilling) to move between regions, are no

[15] Note that in case of interregional labor *im*mobility real wages can possibly differ between regions.

[16] Note that the model is actually static which implies that the economy immediately adjusts to the LRE. The model outline in Table 1.1 merely serves to get some intuition behind the model. Also, it is the algorithm we use to find the LRE when simulating the model.

longer present. More formally, the LRE solution *{w_i,q_i}* for each region *i* has to adhere to the additional condition that real wages, ω_i, are equal across all regions:

$$\omega_i = q_i^{-\gamma} w_i = \omega \qquad \forall i \tag{1.16}$$

Having specified the equilibrium equations, the next point of interest is to determine the equilibrium distribution of firms and people over the *M* regions in the model and how this distribution depends on the level of economic integration modeled here by the level of trade costs, τ_{ij}.

1.2.3 Economic integration

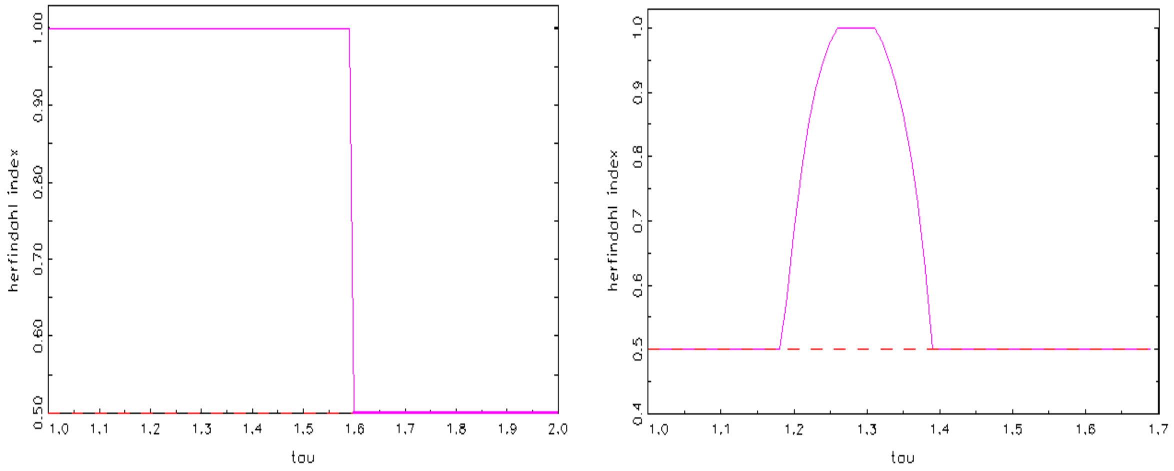
Puga (1999) makes the following simplifying assumption in order to derive analytical results with respect to the effect of ongoing economic integration on the distribution of economic activity across regions: trade costs between each pair of regions are the same and there are no costs of transporting goods within one's own region, i.e.:

$$\tau_{ij} = \tau,\ if\ i \neq j \qquad and \qquad \tau_{ij} = 1,\ if\ i = j \tag{1.17}$$

Making these assumptions, Puga (1999) is able to show analytically that the effect of ongoing integration on the degree of agglomeration depends crucially on whether one assumes perfect interregional labor mobility or interregional labor immobility. This difference is best summarized by Figures 1.1a and 1.1b respectively[17]. Figures 1.1a and 1.1b are obtained from a simulation of the symmetric 2-region model. These two figures replicate Figure 2 and 6 in Puga (1999) and are also known as the tomahawk and the bell shaped curve, respectively.

Figure 1.1 Trade costs and the long run equilibrium in the 2 region model

Figure 1.1a Figure 1.1b



Notes: Simulation parameters as in Puga (1999), p.333. In Figure 1.1a, μ = 0.2, γ = 0.1, θ = 0.55, σ = 4 and in Figure 1.1b: μ = 0.3, γ = 0.4, θ = 0.94, σ = 4. The breakpoints, see Appendix 1.A, are in Figure 1.1a, τ_S = *1.6002*. Figure 1.1b, $\tau_{S,1}$ = *1.1839 and* $\tau_{S,2}$ =*1.3887*. Note, that in the case of only two regions the Herfindahl index on the y-axis is similar to depicting one region's share. See section 1.3.1 for more details on the use of the Herfindahl index.

[17] See Appendix 1.A for the analytics behind these Figures.

Figures 1.1a and 1.1b show that the assumption about interregional labor mobility crucially affects the sensitivity of the spatial distribution of economic activity to increased levels of economic integration. Starting from a relatively high level of trade costs (e.g. $\tau = 1.7$), increased integration (moving from right to left along the x-axis) will in the case of interregional labor mobility result in a sudden (catastrophic) change in the economic landscape characterized by a shift from perfect spreading to complete agglomeration. In case of interregional labor immobility, increased integration will also first result (but less catastrophically) in agglomeration, but as integration continues, the economy ultimately moves back to perfect spreading[18].

This return to symmetry in case of interregional labor immobility is caused by the fact that the spreading force imposed by the increased difficulty with which firms have to attract their workers from the agricultural sector is not weakened (as in case of an interregional labor mobility) by the possibility to attract workers from the other region. As with ongoing economic integration trade or transport costs become relatively small, this means that wage differences become more important as a cost factor in production. Eventually the spreading forces (i.e. the lower wage level in the periphery) 'take over' and industrial firms spread out over both regions again. This does not happen with interregional labor mobility as the higher real wage levels in agglomerations keep attracting workers from the periphery (see also e.g. Helpman (1998), as to how not only 'non-traded production inputs' (here the interregionally immobile labor force), but also non-traded consumption goods can give rise to such a return to symmetry at low levels of trade costs).

1.3 BEYOND AN EQUIDISTANT SETUP

The results regarding the impact of increased levels of integration on the long run equilibrium, as summarized by Figures 1.1a and 1.1b, crucially depend on the assumption of an equidistant regional structure, i.e. equation (1.17). It is difficult to envisage such a regional structure with more than three regions on a flat plain; more important it is at odds with the real world. There, regions are related to each other by a more complicated geography structure:

$$\tau_{ij} \neq \tau \quad for\ all\ \ i,j \tag{1.18}$$

All empirical work within the new economic geography literature, be it multi-country (e.g. Redding and Venables, 2004) or multi-region (e.g. Hanson, 2005; Brakman et al., 2006; Crozet, 2004, Breinlich, 2006 and Knaap, 2006) studies, imposes such a *multi-dimensional* geography structure on the data. The geography structure depends usually on bilateral distances between regions and sometimes also incorporates the idea that ex- or importing to a region in a different country involves extra trade costs (e.g. tariffs or language barriers. See

[18] Here and in the rest of the chapter, the main focus is essentially only on the sustainability of the symmetric equilibrium. The sustainability of a perfectly agglomerated equilibrium has also received a lot of interest in the NEG literature. However, whereas in a two-region model perfect agglomeration is a well-defined concept, i.e. all economic activity is located in one region, this is not so straightforward in the multi-region case. Also, in case of an interregionally immobile labor force, a formal analysis of the sustainability of a perfectly agglomerated equilibrium is already analytically very cumbersome and as a result quite uninformative in a simple two-region setup (see p.325 in Puga, 1999).

chapter 2 for a more detailed discussion on the issue of modeling regions' spatial interdependencies). The practice of relating the estimated structural parameters back to the analytical results obtained from the theoretical models (mostly the simplest 2-region version of the underlying model) to answer questions like "where on the bell (tomahawk) are we?" can thus be considered largely tentative or even misleading (see also Behrens and Thisse, 2007).

The most elegant solution to this problem would of course be to develop an analytically solvable version of an NEG-model with a multi-dimensional geography structure. However, given the mathematical difficulties that are far from straightforward (probably even impossible) to overcome[19], we propose a different strategy in this chapter: *simulation*. Instead of trying to explicitly solve for equilibrium using equations (1.14)-(1.16), and also (1.17) in case of interregional labor mobility, making some necessary simplifying assumptions in the process, one can also use these equations to simulate model outcomes. A major advantage of this is that it does not require any simplifying assumptions about the geographical dependencies between regions. A drawback, however, of performing merely simulations is that one is never 100% certain whether or not the results found are due to the particular parameter setting used in the simulation and whether or not the equilibrium solution found is unique or not[20]. However by extensive simulation (starting at different initial distributions of labor over the regions and/or sectors, and using many different, though cleverly chosen on the basis of the analytical results in the unidimensional case, model parameters) one can get a good grasp of the model's behavior in the multi-region, multi-dimensional geography case. Even more so from an empiricist's point of view, where the number of parameters to use is restricted to merely one set of parameters, i.e. those estimated (see also section 1.4), hereby substantially limiting the number of simulations needed when performing robustness checks.

1.3.1 The simulation setup

The version of the model that we simulate consists of 194 regions, the number of NUTSII[21] regions that make up the 15 countries of the European Union before its eastward expansion in 2004. Our simulation model solves for the long run equilibrium (LRE) in case of an interregionally immobile labor force using a sequentially iterative search algorithm that follows the schematic outline of the model as presented in Table 1.1, where the algorithm stops whenever the nominal wages in each region change less than 0.00000001% between iterations. In case of interregional labor mobility, we also have to specify the way workers

[19] The model not only becomes analytically intractable, also the concept of agglomeration breaking becomes more problematic when considering more than two regions (what is agglomeration and when is it breaking is not as clear-cut as in the 2-region case).

[20] Given the fact that the symmetric 2-region version of the model with interregional labor mobility is characterized by multiple equilibria (see Robert-Nicoud (2004) for a formal proof) it is not unthinkable to also be a characteristic of a multi-region model with a multi-dimensional geography structure.

[21] Nomenclature of Territorial Units for Statistics, a division of the EU15 in regions for which statistical information is collected by Eurostat. Excluding Luxembourg and the overseas territories of Portugal, Spain and France, the following 14 countries (nr. regions) are included in the sample: Belgium (11), Denmark (3), Germany (30), Greece (13), Spain (16), France (22), Ireland (2), Italy (20), The Netherlands (12), Austria (9), Portugal (5), Finland (6), Sweden (8) and The United Kingdom (37).

move in response to real wage differences between regions (and subsequently solve for the equilibrium distribution of labor between manufacturing and agriculture in order to have identical wages within a region). Following Fujita et al. (1999), we assume that workers move according to the following simple dynamics, which can be reconciled with for example evolutionary game theory (Weibull, 1995; see also the discussion in Baldwin et al., 2003):

$$d\lambda_i / \lambda_i = \psi(\omega_i - \overline{\omega}), \qquad with\ \overline{\omega} = \sum_j \lambda_j \omega_j \tag{1.19}$$

where $\lambda_i = L_i / \sum_j L_j$, $\overline{\omega}$ the average real wage per capita and ψ is a parameter governing the speed at which people react to real wage differences. Again we define equilibrium to be reached whenever the number of people in each region changes with less than 0.00000001% between iterations. We explicitly mention the stopping criterion used in our algorithm as we found the equilibrium solution quite sensitive (especially in case of interregional labor mobility) to its specification. With a less stringent stopping criterion (e.g. 0.000001%) than the one we use in our baseline simulations, the search algorithm may stop 'too early', presenting a short run equilibrium characterized by partial agglomeration as the long run equilibrium (see also page 23). In general we stress that the type of search algorithm used to find the LRE, i.e. the way 'dynamics' are artificially but necessarily introduced in an essentially static model, is of paramount importance and can potentially give misleading results regarding the agglomeration pattern in the LRE (as in e.g. Fowler, 2007 and Brakman et al., 2001, ch.10; for more discussion on this see also section 1.3.2).

Next, we have to choose the parameter values for which to show the simulation outcomes. Figures 1.1a and 1.1b already showed that our simulation model replicates the findings in Puga (1999) using the same parameter values as in that paper. For our multi-region simulations, however, we use different parameter values, namely μ = 0.6, γ = 0.2, θ = 0.55, σ = 5. This choice is made for the following important reason. Using this set of model parameters, we can isolate the impact of the assumption made about the interregional mobility of the labor force on the conclusions drawn regarding the effect of increased integration on the spatial distribution of economic activity. It precludes a situation where the choice of parameters is such that it results in (uninteresting) LRE characterised by complete agglomeration or symmetry for all levels of trade costs in either of the two interregional mobility scenarios (note: the latter is the case when using the same parameter values as in Puga, 1999[22]).

To provide a benchmark for the simulation results in the rest of the chapter, Figures 1.2a and 1.2b show the effect of increased integration, using the above mentioned parameter values, in case of the simplest, analytically solvable, equidistant 194-region version of the model. Because we are dealing with more than two regions, the vertical axis depicts the

[22] We found that being able to obtain the effect of increased integration in case of an interregionally immobile labor force similar to Figure 1.1b, is quite sensitive to two of the structural parameters., namely θ and μ. Either θ or μ needs to be set 'large enough'. Instead of, as in Puga (1999), picking a high value of θ, we decided for the latter option, where our choice is mainly driven by the fact that such a high share of labor in agriculture seems to be more at odds with reality than assuming a high share of intermediates in final production (see e.g. Hummels, 2001, who documents a large increase in trade in intermediates over the last decades).

Herfindahl index, $HI = \sum_i \lambda_i^2$, as our agglomeration measure (where in case of interregional labor (im)mobility λ_i denotes the share of a region's (manufacturing) workers in the total number of (manufacturing) workers in the economy). The advantage of using the HI-index is that it allows us to distinguish between different levels of agglomeration in a multi-region setting.[23]

Figure 1.2 shows that the effect of integration on the spatial distribution of industrial activity is qualitatively similar to the effect shown in Figure 1.1 and depends crucially on the assumption of whether or not labor is mobile between regions. In Figure 1.2a labor is mobile between regions and, as in Figure 1.1a, ongoing integration results in a sudden move from symmetry to agglomeration.

Figure 1.2 Trade costs and the long run equilibrium for 194 equidistant regions

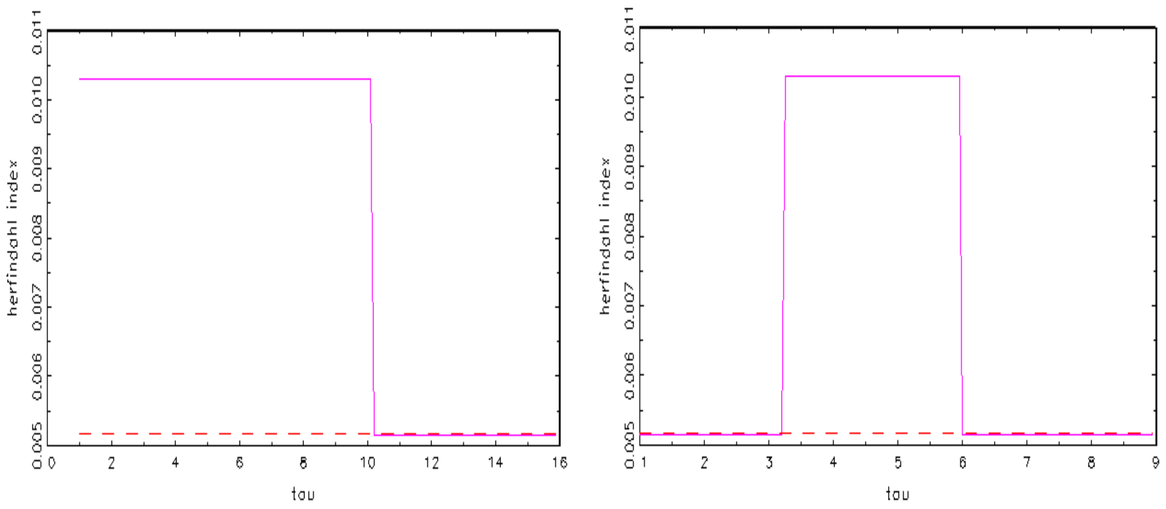


Figure 1.2a

Figure 1.2b

Notes: Simulation parameters: $\mu = 0.6$, $\gamma = 0.2$, $\theta = 0.55$, $\sigma = 5$. In Figure 1.2a, $\tau_S = 10.107$. Figure 1.2b, $\tau_{S,1} = 3.2024$ and $\tau_{S,2} = 5.9710$. M = 194. Stability is checked by equally shocking half of the regions in terms of number of workers in Figure 1.2a or in terms of number of workers in manufacturing in Figure 1.2b. Given that we equally shock half the regions, agglomeration means an equal division of labor/firms over these 97 regions, i.e. HI =0.0103. If we instead shock a different number of regions equally, these shocked regions will attract all footloose activity equally. For example, when shocking only a single region, this region will attract all activity (HI = 1). The dashed line shows the value of the HI associated with a perfect spreading equilibrium, HI = 1/194.

Figure 1.2b shows the same move from symmetry to agglomeration and back to symmetry, as in Figure 1.1b, although here we find that the shift from symmetry to agglomeration is not gradual as in Figure 1.1b (see Puga (1999), footnote 18 for a discussion of this result).

1.3.2 Introducing more realistic geography structures

Using Figure 1.2 as a benchmark, we now turn to introducing an asymmetric geography structure to the model. Instead of assuming all regions equidistant to each other we define the level of trade costs between region *i* and *j* as being pair specific, i.e.

[23] Although there are other arguably preferable measures of agglomeration (see e.g. Bickenbach and Bode, 2006), we deem the HI suitable when looking at the change of agglomeration level in response to changes in trade costs. Using other, more sophisticated measures does not change our results qualitatively.

$$\tau_{ij} = \tau_{ji} = \tau D_{ij}^{\delta}(1 + bB_{ij}),\ \ if\ i \neq j \quad and \quad \tau_{ij} = 1,\ \ if\ i = j \tag{1.20}$$

where D_{ij} is the great-circle distance between region *i*'s and region *j*'s capital city, τ the transport cost parameter, B_{ij} an indicator function taking the value zero if two regions belong to the same country and one if not, δ is the so-called distance decay parameter and $b \geq 0$ a parameter measuring the strength of the border impediments. Specifying trade costs this way is common in empirical studies (see e.g. Anderson and van Wincoop (2004) and Redding and Venables (2004) and also chapter 2 of this thesis). It captures the notion that trade costs increase with distance and it also allows international trade to differ from intranational trade (due to either tangible costs in the form of e.g. tariffs, but also due to intangible costs in the form of differences in language, culture, etc). Note, that each simulation starts with an *equal* distribution of labor and land over the regions (and in case of labor also over the sectors within a region) in the sample. In this way we are able to isolate the effect of non-equidistant regions from the influence of economic mass. In the next section, we also allow regions to differ in their initial economic mass, measured by interregional differences in initial employment and arable land area, as in reality the resulting economic geography is very much the result of the interplay between relative (distance) and absolute (economic mass or size) geography.

We simulate the effect of ongoing integration on the spatial distribution for the following two cases (for a combination of these two, see Appendix 1.B):

a. Assuming no border effect, $b = 0$, and looking at the effect of lowering the transport cost parameter, τ given a fixed distance decay parameter δ. → *see Figure 1.3*[24].

b. Assuming no transport costs, i.e. $\tau = 1$ *and* $\delta = 0$, and looking at the effect of lowering the border effect, *b*. → *see Figure 1.4.*

Figure 1.3 Transport costs and the LRE for 194 non-equidistant regions

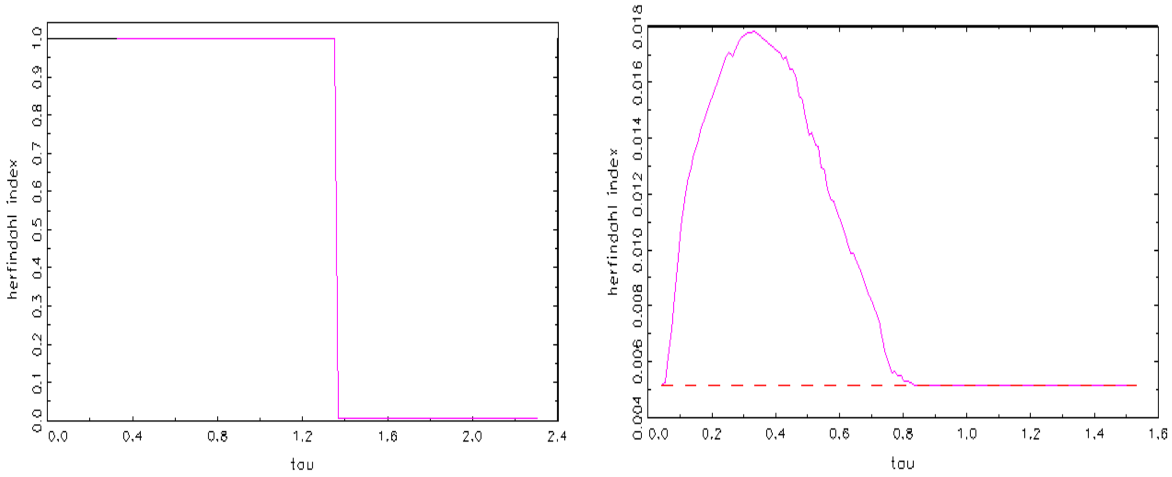


Figure 1.3a

Figure 1.3b

Notes: μ = 0.6, γ = 0.2, θ = 0.55, σ = 5 and b = 0, δ = 0.38 see estimation results in Brakman et al. (2006). The dashed line shows the value of the HI associated with a perfect spreading equilibrium, HI = 1/194.

[24] One could also use the distance decay parameter δ to simulate the effect of decreasing transport costs. Results are very similar to those presented in Figure 1.3 and are available upon request.

The ‘a’ panels of Figures 1.3 and 1.4 show the results of increased integration[25] in case of an interregionally mobile labor force and the ‘b’ panels when labor is immobile between regions. Comparing Figures 1.3 and 1.4 to the benchmark equidistant case presented in Figure 1.2, we observe that the effect of ongoing integration still crucially depends on the assumption whether or not the labor force is interregionally mobile. Without interregional labor mobility (see Figures 1.3b and 1.4b), ongoing integration will, as in the equidistant case, first result in an increased agglomeration followed by a return to symmetry with further integration. The shift from symmetry to agglomeration and back to symmetry is however not as sudden as in the equidistant case (resembling much more the bell-shaped curve as found when using Puga’s parameter settings, recall Figure 1.1b).

Figure 1.4 The border effect and the LRE for 194 non-equidistant regions

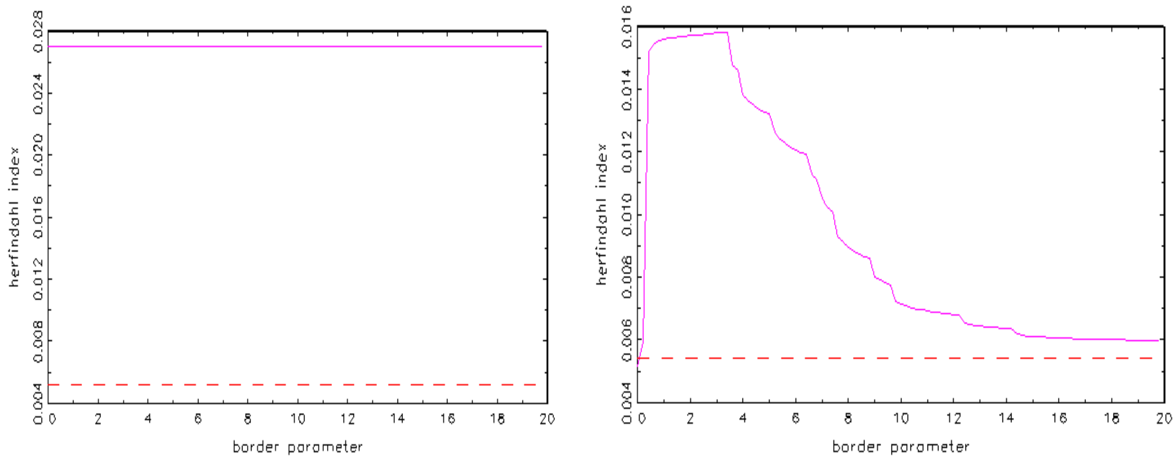


Figure 1.4a Figure 1.4b

Notes: $\mu = 0.6$, $\gamma = 0.2$, $\theta = 0.55$, $\sigma = 5$ and $\delta = 0$, $\tau = 1$. The dashed line shows the value of the HI associated with a perfect spreading equilibrium, HI = 1/194.

Moreover, complete agglomeration is never reached; manufacturing activity is still present in several regions. With interregional labor mobility, ongoing integration in the form of decreasing trade costs, as depicted in Figure 1.3a, also has a similar effect as in the equidistant case. It results in a sudden (catastrophic) change in the economic landscape from symmetry to complete agglomeration. With a positive border effect, as in Figure 1.4a, full agglomeration is always the long run equilibrium outcome for any level of the border effect shown here (see the next section for more details on this finding).

The above results in case of interregional labor mobility are different from the findings in Stelder (2005) and Brakman et al. (2006) (the latter is also based on Stelder’s

[25] Note that when modeling increased integration as a lowering of transport costs τ, eventually one reaches a point at which trade costs between two regions are exactly equal to one, i.e. $\tau_{ij} = \tau_{ji} = 1$. From this point on a further decrease of τ would result in these trade costs becoming smaller than one, which would violate the rationale behind the iceberg-assumption (this would imply that in order to deliver one unit, less than one unit would have to be shipped). We restrict the minimum transportation costs to $\tau_{ij} = 1$: if a further reduction of τ results in a value of trade costs between two regions less than 1, we fix it at 1. As a result the minimum value that τ takes is such that trade costs between all pairs of regions equal 1.

model but starting the simulations from the actual instead of an equal distribution of economic activity across regions). In these two papers, multi-region simulations of the Krugman (1991) model (where labor is mobile between regions) with an asymmetric geography structure give rise to LRE characterized by incomplete agglomeration, with the level of agglomeration increasing and the number of agglomerated regions decreasing, the lower trade costs. Here we find that agglomeration forces are so strong in a model with interregional labor mobility, that when the spreading equilibrium becomes unstable each introduced asymmetry (be it initial endowment, or (as here) geographical location) in the limit results in the one region that is the most favorable in terms of net asymmetries attracting all industrial activity. That the above-mentioned papers instead find partial agglomeration when labor is interregionally mobile could possibly be explained by the particular geography structure used in those papers. For example the exponential distance decay function, resulting in highly localized areas of relatively cheap trade, or the particular distance grid used in these papers (in Brakman et al. (2006) also the initial labor distribution, see also section 1.4). However, the only way we are able to find partial agglomeration patterns similar to those presented in the above-mentioned papers (even when using similar model parameters and the same distance decay function) is when using a higher stop criterion in our search algorithm. In that case partial agglomeration would wrongly be interpreted as being the LRE (see p.19 for a discussion of the sensitivity of the simulated LRE to the stop criterion).

1.3.3 Same overall degree of agglomeration – but different spatial distribution

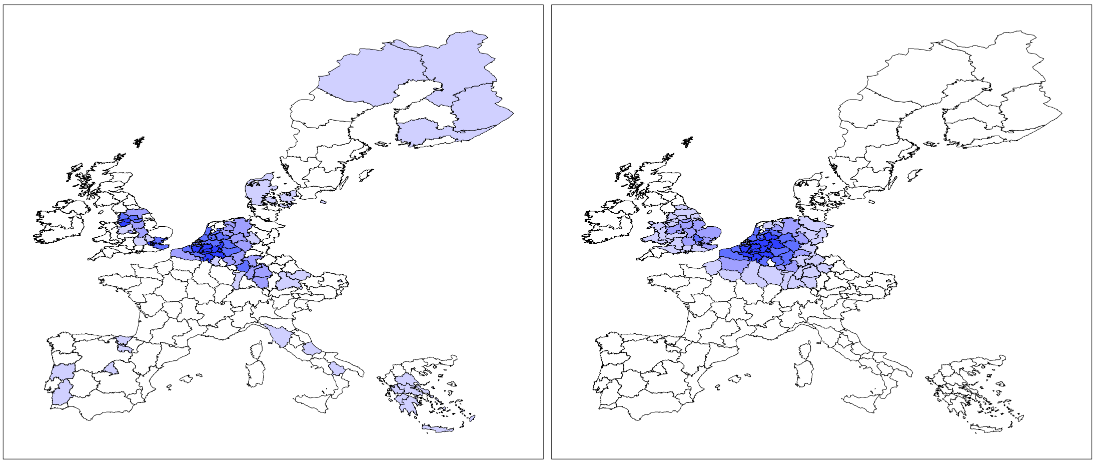
A major difference with the equidistant case (even more so with the simple 2-region version of the model) is that the same level of agglomeration as measured by the Herfindahl index does not necessarily mean the same spatial distribution[26]. This is especially so when there is interregional labor *im*mobility as illustrated by Figure 1.5, but also in case of interregional labor mobility, the same level of agglomeration does not necessarily mean the same spatial distribution (see Appendix 1.B). In Figure 1.5, the left and the right panel show the spatial distribution of the manufacturing sector obtained using the same parameters as in Figure 1.3 but for two different values of τ, chosen such that the distribution in both panels gives rise to the *same* value of the Herfindahl index. That is, the left (right) panel shows the distribution 'on the right (left) side of the bell' in Figure 1.3b, corresponding to a lower (higher) level of economic integration respectively.

In the simple equidistant models these two distributions would be exactly the same. As can be seen from Figure 1.5 below, this no longer holds when allowing for a more realistic geography structure: the left panel shows a distribution with a group of centrally located core regions (in Belgium, The Netherlands and the western part of Germany) but still some industrial activity in the peripheral regions (Sweden, Greece, Portugal). The right panel shows

[26] Note that this can also happen in the 2-region (or equidistant multiple region) version of the model. The difference is that this can only happen due to an exogenous shift of economic activity from one region to the other (i.e. agglomeration can occur in either of the two regions for a given level of trade costs at which perfect spreading is an unstable equilibrium). A change in trade costs does not shift the agglomeration from one region to another, whereas here this can happen (see Appendix 1.B).

a larger group of centrally located core regions (now also extending into the southern UK and northern France) and no more industrial activity in the peripheral regions.

Figure 1.5 Similar agglomeration but different regional distribution



Notes: Simulation parameters as in Figure 1.3b. Left panel: $\tau = 0.5$. Right panel: $\tau = 0.16$. HI = 0.0146.

More generally we find, on the basis of more extensive simulations (not shown here), that starting from a symmetric distribution of industrial activity, increased economic integration has the following effect on the spatial distribution of economic activity[27]: At a certain level of integration, agglomeration starts, with a number of core regions attracting activity from nearby regions, still leaving some level of industrial activity in the peripheral regions. As integration proceeds, this process continues until the peripheral regions are completely specialized in agriculture, and industrial activity only takes place in the centrally located core regions. A further falling of trade costs eventually reverses this process, with industrial activity gradually spreading from the core, at first to nearby regions (not to the peripheral ones!) and eventually reaching the peripheral regions again.

1.3.4 The importance of the geography structure imposed

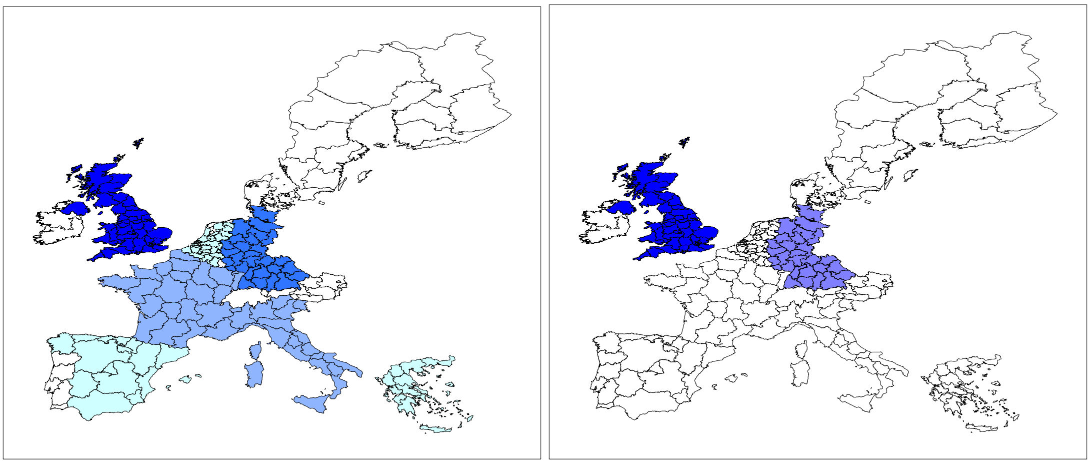
Essentially, the way regional interactions in the regional economy are modeled, i.e. the imposed geographical structure connecting the regions, crucially and predictably influences the way integration affects the distribution of economic activity (see also Behrens and Thisse, 2007). That is, in our case, the distance matrix, D_{ij}, and the borderdummy matrix, B_{ij}, together determine the equilibrium outcomes whereas the parameters (δ, τ and b) in the transport cost function determine the relative strength of the D_{ij} and B_{ij} effects.[28] When only D_{ij} is allowed to have an effect by setting the border parameter b equal to zero, agglomeration will always be in or around the most centrally located regions in case of interregional labor immobility (see Figure 1.5) and *in* the most centrally located region in case of interregional labor mobility (i.e. Brabant-Wallon in our case).

[27] Modeling a decrease in transport costs by using the distance decay parameter δ gives exactly the same pattern.

[28] If more asymmetries are introduced (as e.g. in the next section an asymmetric initial distribution of labor) these will also play a crucial role.

When instead only B_{ij} is allowed to have an effect by setting δ and τ equal to zero and one respectively, Figure 1.6 shows what happens when border impediments are decreasing in the case of an interregionally immobile labor force. Now agglomeration, if it occurs, will be in countries with many regions relative to other countries, with the regions within these countries all having the same share of footloose industrial activity.

Figure 1.6 Changing the border impediments B_{ij} (Fig. 1.4b in more detail)



Notes: Simulation parameters as in Figure 1.4. Left panel: b = 8. Right panel: b = 3.

As can be seen when comparing the left and right panel of Figure 1.6, when the border effect becomes less important, ever fewer countries retain footloose activity. In case of interregional labor mobility (not shown here), the largest country in terms of number of regions, i.e. the UK in our case will eventually attract all industrial activity (again equally spread over the regions within the UK)[29]. Appendix 1.B shows the results when considering both geography structures, B_{ij} and D_{ij}, simultaneously.

To sum up, many of the qualitative conclusions obtained from the simple symmetric NEG models carry over when introducing a more realistic asymmetric geography structure like the one depicted by our equation (1.20). Catastrophic agglomeration as a result of increased integration remains a characteristic of the model with interregional labor mobility. Also in case of interregional labor immobility, the impact of increased integration shows a similar pattern in terms of the long run equilibrium agglomeration levels (first increasing and finally decreasing) as in the simple symmetric models. However, as shown in this section, a big difference with the symmetric versions of the model is that the same level of agglomeration (in terms of some agglomeration index) does not necessarily mean the same spatial distribution of economic activity once a more realistic geography structure is added to the model. Finally, the simulated effects of increased integration depend crucially (and predictably) on the type(s) of asymmetric geography structure imposed.

Now that we have established what the effects are of introducing non-equidistant regions in the Puga (1999) model, we next turn to the question whether and how the long run

[29] Note that this shows the importance of the definition of a region. Using a different subdivision of countries into regions will have an impact on the simulation results when considering the importance of border effects.

equilibria are affected by the introduction of regional differences in economic size alongside the asymmetries introduced in the present section.

1.4 EUROPEAN INTEGRATION WHEN (RELATIVE) GEOGRAPHY MATTERS

The asymmetric geography structure between regions is certainly not the only "real world" asymmetry faced by the empirical researcher or policy maker. Instead, the current (unequal) distribution of economic activity is also of paramount importance, either in being used as input in the estimations or as the basis of implementing (new) policies. This section takes note of this and tries to offer some guidance in how to use estimated parameters from a structural NEG model to be able to provide policy makers with predictions regarding the effect of ongoing European integration on the current distribution of economic activity. Instead of relating these estimated parameters back to the simple symmetric version of the NEG model (see e.g. Crozet, 2004), we argue that predictions regarding the impact of ongoing integration should instead (ideally) be obtained by simulating the same NEG model as estimated, i.e. with the same asymmetries present and using the estimated model parameters (see Brakman et al. (2006) for a first pass at this).

1.4.1 Estimating the structural parameters

To illustrate the usefulness of this strategy, we need to obtain estimates of the structural model parameters in the Puga (1999) model. Using data from Cambridge Econometrics on compensation per employee and gross value added (GVA) for our sample of 194 EU15 NUTS-II regions over the period 1992-2000[30], we obtain the estimates of σ, δ and b by estimating, using NLS panel data techniques, the wage equation (in logs) shown in (1.14) while substituting (1.20) for τ_{ij}.[31]

Table 1.2 Structural parameter settings

σ	7.122
δ	0.102
b	*285.65 [set to 0 in simulation exercises]*
γ	0.335
μ	0.284
θ	0.234
Labor interregionally mobile	
τ_s	3.199
Labor interregionally immobile	
$\tau_{s,1}$ and $\tau_{s,2}$	symmetry always stable

Notes: In the estimation of the wage equation σ and δ are significant (p-value: 0.000). b is insignificant (p-value: 1.000).

[30] Due to wage data availability we use data at the NUTS I-level for Germany and London, which leaves us with 183 regions.

[31] For more detail, see Brakman et al. (2006). Like in Brakman et al. (2006) we set $\mu = 0$ in the estimation of the wage equation. First-nature geography variables are omitted as explanatory variables, we do include country dummies.

Parameter values of μ and γ are calibrated using data from Input-Output Tables provided by the OECD (edition 2002) and θ is calibrated by using Eurostat data on the compensation of employees and gross value added in the agricultural sector in the EU-15 for the year 1995. Table 1.2 above shows the resulting parameter estimates, together with the breakpoint(s) that would apply at these parameter settings for our 194-region model if we would stick to an equidistant geography structure (shown in case of both interregional labor mobility and immobility, respectively).

Next, we use these parameters in a simulation exercise with $b = 0$[32]. Also, instead of introducing only the geography structure by means of a non-equidistant relative geography structure between regions based on (1.20) as in section 1.3, we now also introduce, staying as close as possible to the Puga (1999) model, the true initial distribution of labor (total employment share) and land (arable land share), as shown in Figures 1.7a and 1.7b respectively, as additional asymmetries to the simulation exercises.

Figure 1.7 Adding economic mass: actual labor and land distributions

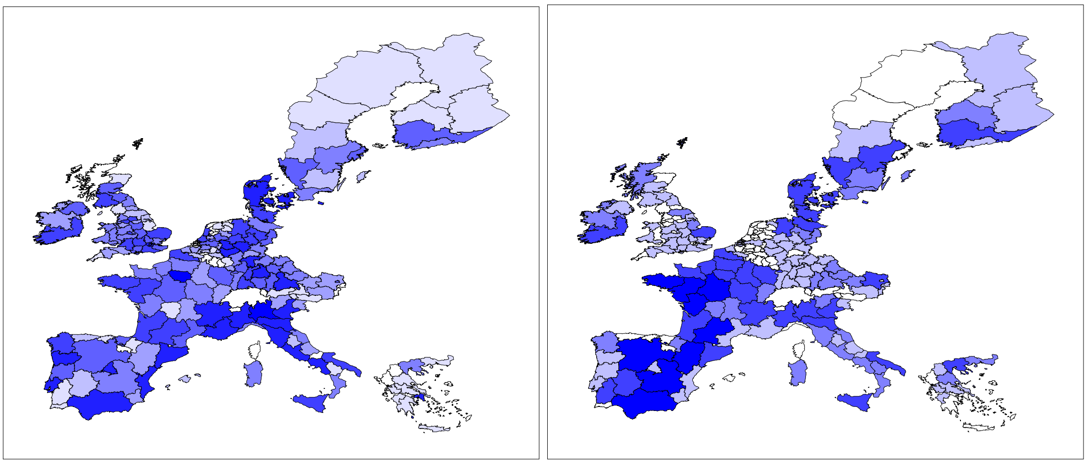


Figure 1.7a: Total employment

Figure 1.7b: Arable land

1.4.2 Simulating the impact of ongoing EU integration

Having specified the simulation settings in the previous section, we now turn to simulating the effect of ongoing integration. As in the previous section, we do this in two different ways:

1. *a decrease in interregional transport costs* τ, e.g. the EU supports the construction and upgrading of transportation links (roads, railways, etc).
2. *a decrease in border impediments b*, e.g. the EU stimulates the formation of an internal market by removing trade barriers (streamlining national regulations, removing border controls, etc).

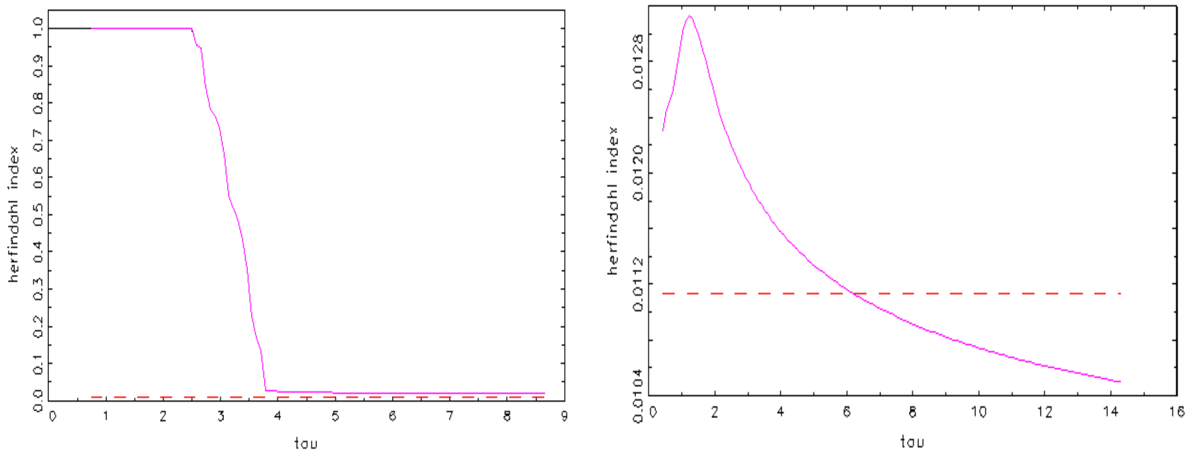
[32] One can choose any parameter value for the border effect as one can be 99% sure that it lies within the range [-1.16 x 10^{14}, 1.16 x 10^{14}]. A possible reason for the insignificance of this parameter may be that the extent of the border effect differs substantially among different pairs of EU15 countries (see Breinlich (2006), who provides evidence on this).

1.4.2.1 Decrease in transport costs

Here we look at the effects of lowering the transport cost parameter τ on the spatial distribution of economic activity[33]. Figure 1.8 below shows the resulting long run equilibrium when labor is either a) interregionally mobile (left panel) or b) interregionally immobile (right panel) for each value of transport cost.

Figure 1.8 Transport costs and the long run equilibrium when geography matters

Figure 1.8a Figure 1.8b



Notes: Simulation parameters as in Table 1.2 and the simulations are started using the actual distributions of arable land and total employment (see Figure 1.7). Left panel: interregional labor mobility. Right panel: interregional labor immobility.

In both Figure 1.8a and Figure 1.8b, the dashed line shows the value of the Herfindahl index associated with the actual, initial spatial distribution of economic activity across the 194 regions. With interregional labor mobility (see left panel Figure 1.8), agglomeration increases with decreasing transport costs. Note, however, that the process is gradual rather than catastrophic. This is in contrast to the findings in section 1.3. A closer look at the map of Europe with the equilibrium distribution for different levels of transport costs provides the reason for this finding (see Figure 1.9 below).

At very low levels of transport costs we find (as before) complete agglomeration in the region Île-de-France with Paris as the main city (not shown in Figure 1.9). When transport costs increase, spreading forces take over and the economy moves towards a more equal distribution across the 194 regions. As Figure 1.9 shows, the most important spreading force turns out to be the (unchanging) distribution of arable land! Once agglomeration forces become less important due to the increasing transport costs, labor starts moving towards the regions that by virtue of their large supply of arable land offer higher wages (see Figure 1.9a). With very high transport costs this finally results in the distribution of manufacturing activity being (almost) the same as the distribution of arable land (compare Figure 1.9b to Figure 1.7b, the corresponding correlation coefficient is 0.962).

[33] Again, the results using the distance decay parameter δ instead are similar and available upon request.

Figure 1.9 Interregional labor mobility and arable land as spreading force

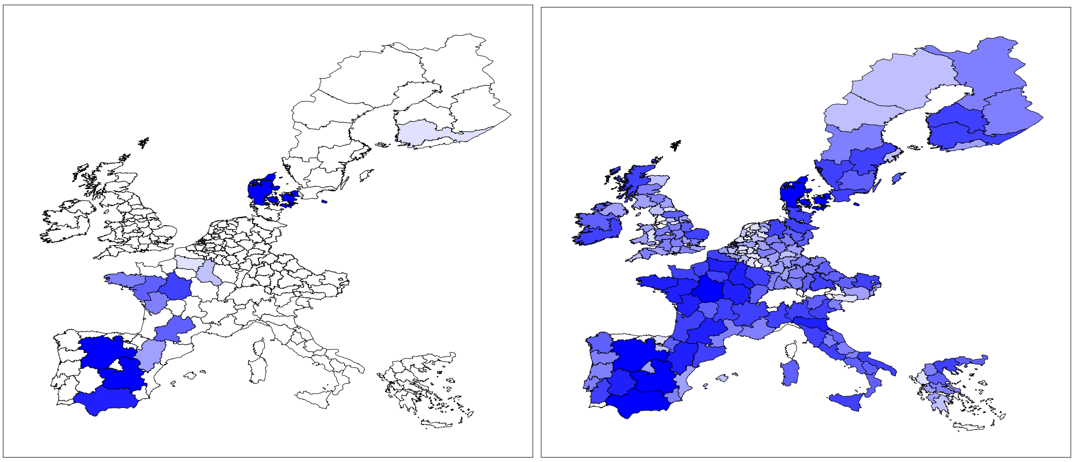


Figure 1.9a: τ = 3.15 Figure 1.9b: τ = 10

Notes: The simulation parameters are set as in Table 1.2 and starting the simulation using the actual distribution of arable land and total employment.

The catastrophic agglomeration patterns found in section 1.3 in the case of interregional labor mobility are thus (partly) due to the fact that in the simulations presented there, land is equally distributed over the regions, imposing no additional asymmetries between regions. We are more likely to find catastrophic agglomeration when initial differences in economic size or mass are not taken into account[34].

Without interregional labor mobility (see Figure 1.8b), we again find a 'bell-shaped' agglomeration pattern.

Figure 1.10 Interregional labor immobility: actual and simulated long run equilibrium

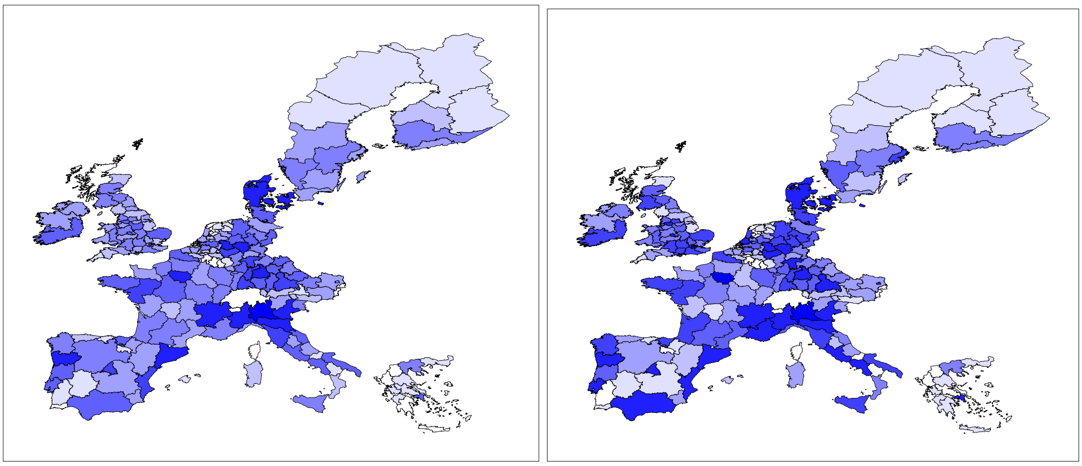


Figure 1.10a: Actual manufacturing labor Figure 1.10b: LRE at τ = 6.181

Notes: In the right panel the simulation parameters are set as in Table 1.2 and τ = 6.181 and starting the simulation using the actual distributions of arable land and total employment.

What is particularly interesting about the long run equilibrium depicted by Figure 1.8b, however, is that for τ = 6.183, *the Herfindahl index of the simulated LRE distribution is*

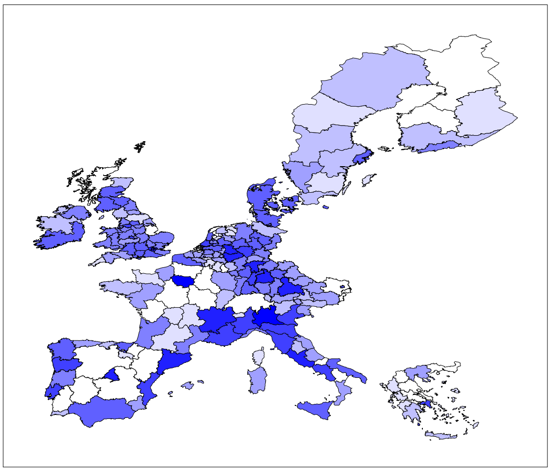
[34] When we endow each region with the same amount of land while using the actual distribution of employment, we again find catastrophic agglomeration with decreasing transport costs. Results available upon request.

identical to the Herfindahl index for the actual distribution of manufacturing employment. Even more striking (as the same Herfindahl index does not necessarily mean the same distribution across regions), when we compare this long run equilibrium distribution for this level of transport costs (Figure 1.10b) with the actual distribution of manufacturing labor for our 194 regions (Figure 1.10a), they are remarkably similar (the correlation coefficient is 0.809). This suggests that the model is able to reproduce the actual spatial distribution of economic activity in the EU quite accurately. This finding immediately suggests a way to provide an answer to the question "*where on the bell are we?*" that has up to now been unsatisfactory answered in case of more than two regions[35].

Given that we can pinpoint the actual location on the bell curve, at τ =6.181, our simulation example suggests (see Figure 1.8b) that, in case of our sample of 194 EU NUTS II regions, increased integration will most likely result in more agglomeration in the future, with only a return to a more equal regional distribution of economic activity once integration reaches very high levels (note however that this return will not mean going back to a completely symmetric distribution of footloose activity).

Of particular interest is Figure 1.11, which shows the distribution at the 'top of the bell' (top of the curve shown in Figure 1.8b).

Figure 1.11 A simulated blue banana



Notes: The simulation parameters are as in Table 1.2 with τ = 1.3 and the simulation is based on the actual distributions of arable land and total employment.

Remarkably, this resembles the so-called Blue Banana (the name given to the pattern of industrial agglomeration in Europe ranging from the southern UK to the Netherlands, through Germany to the north of Italy) quite well. This finding suggests that increased European integration will lead to a more pronounced 'Banana' pattern in the EU (note also the stark core-periphery patterns in France and Spain).

[35] Head and Mayer (2004) show that in case of two regions one can cleverly pinpoint the location on the bell-curve (or tomahawk) using bilateral trade data and estimates of the model parameters needed to calculate the break- and sustain-point of the bell-curve (or tomahawk). See Appendix 1.A, equation 1.A3 for the analytical expressions, and chapter 3 on how to construct implied trade costs from bilateral trade data following Head and Ries (2001).

1.4.2.2 Decrease in border impediments

Again as in section 1.3, we model a decrease in border impediments by a decrease of the border parameter *b*. Figure 1.12 shows the resulting long run equilibria in case of a) an interregionally mobile and b) an interregionally immobile labor force (see Figures 1.12a and 1.12b respectively).

Figure 1.12 A decrease of border impediments and the long run equilibrium

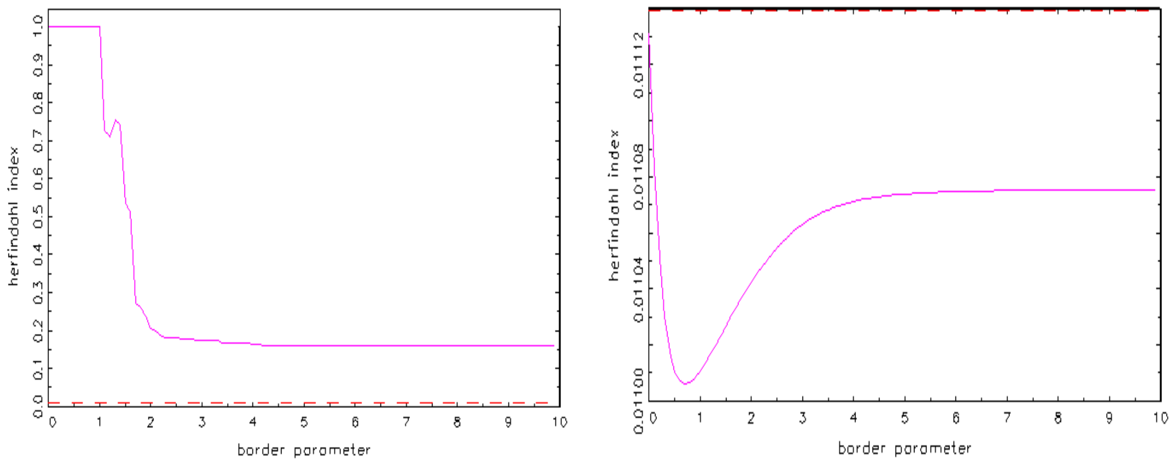


Figure 1.12a

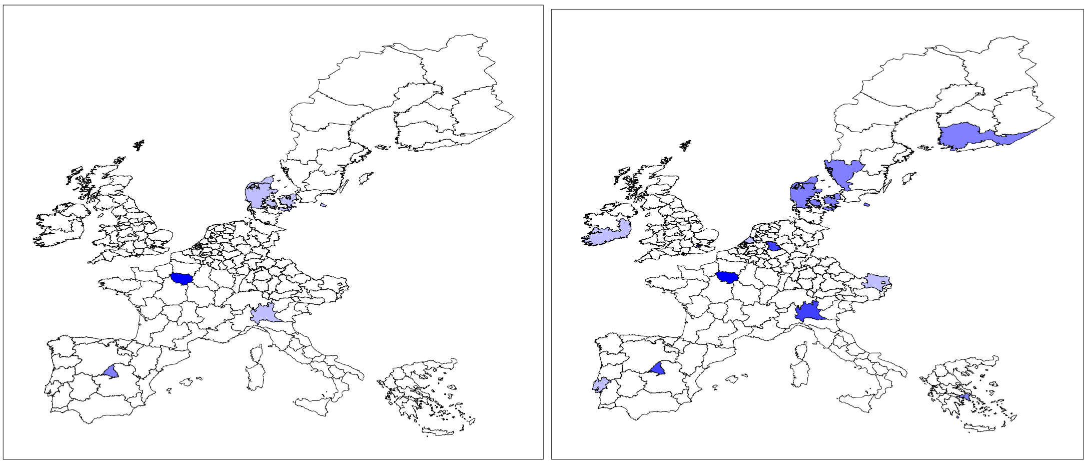
Figure 12b

Notes: Simulation parameters as in Table 1.2 and the simulations are started using the actual distributions of arable land and total employment (see Figure 1.7). Left panel: interregional labor mobility, τ = 1. Right panel: interregional labor immobility, τ = 6.183. The dashed line shows the HI associated with the initial distribution.

With interregional labor mobility, the effect of less border impediments is qualitatively similar to that of a decrease of transport costs *τ*: it is characterized by a move from a more equal distribution over the regions towards full agglomeration (again in Île-de-France) once the border effect disappears. This process is not catastrophic, but the agglomeration level increases gradually towards full agglomeration as the border impediments become weaker.

Figure 1.13 The effect of increasing the border impediments

Figure 1.13a: border effect = 1.6 Figure 1.13b: border effect = 10



Notes: The simulation parameters are set as in Table 1.2 with τ = 1 and starting the simulation using the actual distributions of arable land and total employment.

Different from the case of very high transport costs, however, the distribution of manufacturing activity does not become the same as the distribution of arable land over the regions when border impediments become extremely high. Instead, the distribution at these very high levels of border impediments is characterized by one region in each country attracting all of its country's manufacturing activity (compare Figures 1.13a and 1.13b, where border impediments are respectively low and very high). As the border impediments decrease (*b* falls), the number of countries with a region containing some industrial activity decreases (the countries containing regions with a relatively high supply of arable land are the ones retaining some manufacturing activity the longest) until eventually all industrial activity takes place in Île-de-France.

In case of interregional labor *im*mobility, we model the impact of a decrease of border impediments using $\tau = 6.181$, the value that resulted in a very similar simulated distribution as the actual distribution of manufacturing labor (recall Figure 1.8b and Figure 1.10). Although a first look at Figure 1.12b shows the interesting pattern of an 'inverted bell curve', a closer look at the scale of the y-axis shows that the difference in terms of agglomeration level is extremely small (also when mapping the distribution), so that one can say that, in case of an immobile labor force, an increase of the border impediments has almost no apparent effect on the spatial distribution of footloose activity (the correlation coefficient between the distribution at $b = 0$ and at $b = 10$ is 0.987!). Note that the insensitivity of the simulated long run equilibrium distribution to changes in the border parameter is very much consistent with the results found in the estimation of the wage equation where the estimated border parameter was highly insignificant (see Table 1.2 and footnote 32).

1.5 CONCLUSIONS

Most new economic geography models treat geography in a very simple way: attention is either confined to a simple 2-region or to an equidistant multi-region world. As a result, the main predictions regarding the impact of e.g. diminishing trade costs are based on these simple models. In empirical work these simplifying assumptions become problematic, as conclusions from these simple models may not carry over to the heterogeneous geographical setting faced by the empirical researcher or policy maker. This chapter partly fills this gap by establishing, through careful simulation, the effect of adding more realistic geography structures to the NEG model of Puga (1999), one of the main NEG models that encompasses several other core NEG models.

We first show that many, although not all, conclusions from the simple models do carry over to a multi-region setting with more realistic geography structures. The effect of increased levels of integration on the level of agglomeration is very similar to that found in the simple equidistant (2-region) models. With interregional labor mobility agglomeration levels increase with the level of integration, where, as in the 2-region models, this increase is mostly catastrophic. Without interregional labor mobility, increased integration is accompanied by a steady (not a catastrophic) increase in the level of agglomeration. And when integration proceeds even further this process is reversed, resulting in a return to an equal distribution of economic activity over all the regions, hereby confirming the bell shaped pattern in the 2-region models. Although the qualitative results are similar to the simple equidistant (2-region) models, a major difference is that the same level of agglomeration – as measured by, e.g. the HI-index –corresponds to very different spatial distributions, especially when labor is interregionally immobile.

Having established the effect of introducing more realistic geography structures to a multi-region NEG-model, we next introduce more asymmetries to the simulations, i.e. the actual distributions of employment and arable land. Moreover, using estimates of the structural model parameters, we simulate the impact of increased EU integration on the spatial distribution of regional economic activity for a sample of 194 NUTSII regions. Again the results depend crucially on the assumption about the mobility of the labor force between regions. When labor is interregionally mobile, the model's predictions about the level of agglomeration are probably too extreme (only few (or even one) regions will attract all footloose activity). When labor is interregionally mobile the model's prediction becomes less extreme. More importantly, we are in that case able to answer questions like 'where on the bell are we?'. In our example of the former EU15, we find that more integration will most likely lead to more agglomeration. For EU policy makers an important challenge, as they will have to deal with increased spatial inequality.

APPENDIX 1.A BREAKPOINTS OF THE PUGA (1999) MODEL

Focusing on the 2-region version of the model only (Appendix A1 and A2 in Puga (1999) show the results for the M-region equidistant version of the model), we briefly summarize the analytics behind Figures 1.1a and 1.b.

In case of an interregionally mobile labor force, see Figure 1.1a, there exists a minimum level of transport costs at which a symmetric distribution of firms and workers is a stable equilibrium, i.e. symmetry is stable for levels of transport costs[36]:

$$\tau_{ij} = \tau \geq \tau_S = \left(1 + \frac{2(2\sigma-1)[\gamma+\mu(1-\gamma)]}{(1-\mu)\{(1-\gamma)[\sigma(1-\gamma)(1-\mu)-1]-\gamma^2\eta\}}\right)^{1/(\sigma-1)} \qquad (1.A1)$$

where η is the wage elasticity of labor supply from a region's agricultural to its manufacturing sector[37]. Also a maximum level of transport costs at which agglomeration (i.e. with industrial production and the labor force located in only one region) is a stable equilibrium can be derived, i.e. agglomeration is stable for levels of transport costs smaller or equal to τ_B, being a solution to

$$\tau^{[\sigma(1-\mu)(1-\gamma)-1]}\left(\tau^{2(1-\sigma)} + \frac{(1-\mu)(1-\gamma)}{1+\tau^{\theta\gamma/(1-\theta)}(\tau^{2(1-\sigma)}-1)}\right) = 1 \qquad (1.A2)$$

In case of an interregionally immobile labor force, see Figure 1.1b, the results are quite different. In that case Puga (1999) shows that a symmetric distribution of industrial production is an unstable equilibrium for the following range of transport costs: $0 < \tau_{S,1} < \tau < \tau_{S,2} < \infty$, with $\tau_{S,1}$ and $\tau_{S,2}$ the solutions of the following quadratic expression[38]:

$$\begin{aligned} &[\sigma(1-\mu)-1][(1+\mu)(1+\eta)+(1-\mu)\gamma][\tau^{1-\sigma}]^2 \\ &\quad -2\{[\sigma(1+\mu^2)-1](1+\eta)-\sigma(1-\mu)[2(\sigma-1)-\gamma\mu]\}\tau^{1-\sigma} \\ &\quad +(1-\mu)[\sigma(1-\mu)-1](\eta+1-\gamma)=0 \end{aligned} \qquad (1.A3)$$

and stable for levels of transport costs smaller than $\tau_{S,1}$ and larger than $\tau_{S,2}$.

APPENDIX 1.B COMBINING DISTANCE AND BORDER EFFECT

To illustrate what happens when both geography structures (distance and border effects) are allowed to play a role consider the different effect of increased transport costs on the one hand and increased border impediments on the other hand when starting from the same initial distribution. We choose this initial distribution as the one where $\delta = 0.38$, $\tau = 0.5$ and $b = 0$, which in case of interregional labor mobility amounts to full agglomeration in the region Brabant-Wallon and in case of labor being interregionally immobile in the situation shown in Figure 1.6a. Starting from this distribution, increasing the level of transport costs would eventually result in an equal spreading of industrial activity over all 194 regions (see Figure 1.3b) whether or not labor is interregionally immobile. This is not the case when increasing the magnitude of the border effect.

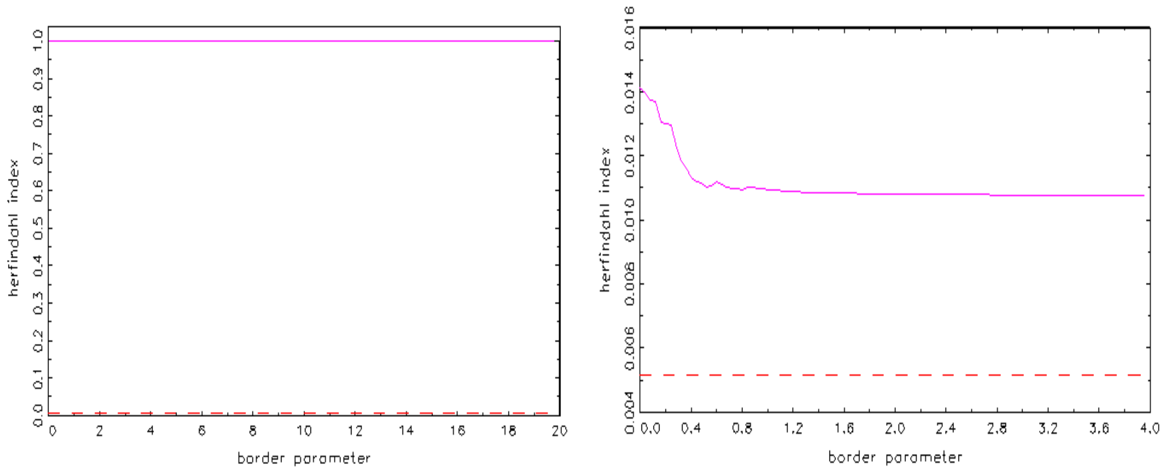
[36] This is the case provided that the denominator is larger than zero. If this does not hold, agglomeration is the only stable equilibrium for all possible levels of transport costs.

[37] Here $\eta = \theta(1-\gamma)/\gamma(1-\theta)$ given the assumed Cobb-Douglas production function in agriculture.

[38] Note that in order for equation (1.21) to have 2 solutions in the required range, the model parameters have to adhere to some additional requirements (see Puga (1999), or Bosker (2007c) for more details).

Figure 1.A1 Changing the border impediments B_{ij} when distance also plays a role

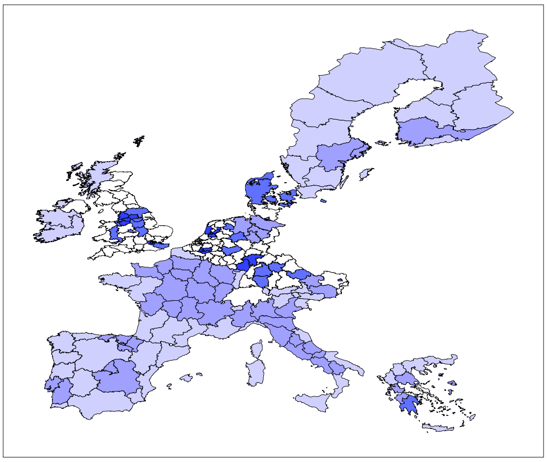
Figure 1.A1a Figure 1.A1b



Notes: Simulation parameters: $\mu = 0.6$, $\gamma = 0.2$, $\theta = 0.55$, $\sigma = 5$ and $\tau = 0.5$, $\delta = 0.38$.

In that case, see Figure 1.A1a, agglomeration remains the only equilibrium outcome as the border effect increases in strength when labor is mobile across regions and we observe no return to spreading. However the region in which all activity agglomerates switches from the most centrally located region, i.e. Brabant-Wallon, to the most centrally located region in a large country with many regions, i.e. Rheinhessen-Pfalz. This is the result of the increased cost of international trade (i.e. B_{ij} gaining in importance in determining the equilibrium outcome). With labor immobile between regions, see Figure 1.A1b, symmetry is also not reached with an increase in the border effect; instead the economy settles at the situation depicted in Figure 1.A2 once international trade becomes more than four times as costly as intranational trade.

Figure 1.A2 Figure 1.A1b in more detail



Notes: Simulation parameters as in Figure 1.A1b, $b = 4$.

Chapter 2

Trade costs, market access and economic geography: Why the empirical specification of trade costs matters[39]

2.1 INTRODUCTION

Trade costs are a key element of new economic geography models in determining the spatial distribution of economic activity (see e.g. Krugman, 1991; Venables, 1996; Puga, 1999 and chapter 1 of this thesis): without trade costs there is no role for geography in NEG models. It is therefore not surprising that trade costs are also an important ingredient of empirical studies in NEG (see e.g. Redding and Venables, 2004; Hanson, 2005 or Head and Mayer, 2004). They are a vital ingredient of a region's or country's (real) market potential, that measures the ease of access to other markets (Redding and Venables, 2004; Head and Mayer, 2006). In the empirical trade literature at large, trade costs are also a main determinant of the amount of trade between countries (see e.g. Limao and Venables, 2001; Anderson and van Wincoop, 2004).

The empirical specification of trade costs is, however, far from straightforward[40]. Problems with the measurement of trade costs arise because trade costs between any pair of countries are very hard to quantify. Trade costs most likely consist of various subcomponents that potentially interact, overlap and/or supplement each other. Obvious candidates are transport costs, tariffs and non-tariff barriers (NTBs), but also less tangible costs arising from cross-border trade due to e.g. institutional and language differences have been incorporated in previous studies (Limao and Venables, 2001). An additional difficulty arises with what is arguably the most obvious measure of trade costs, transport costs. Accurate transport cost data between country pairs are very difficult to obtain and even completely unavailable when considering transport costs between regions[41]. In principle, between-country transport costs can be inferred from cif/fob ratios. The IMF provides for instance extensive trade data on the basis of which these cif/fob ratios can be calculated; see e.g. Limao and Venables (2001) and Baier and Bergstrand (2001). However as put forward by Hummels (1999, p.26) these data "suffer from severe quality problems and broad inferences on these numbers may be unwarranted"[42].

[39] This chapter is an adapted version of Bosker and Garretsen (2007b).

[40] The specification of trade costs may also be not that straightforward from a theoretical point of view, see McCann (2005) and Fingleton and McCann (2007). Also, see chapter 1 for the effect of introducing more realistic second nature geography structures to an NEG model.

[41] Where actual transport cost data are used in the empirical literature the coverage in terms of the number of country pairs is very limited (e.g. costs of shipping a standard 40-foot container from Baltimore (USA) to 64 different countries in Limao and Venables, 2001) or only the evolution of average (by world-region) transport costs over time is available (e.g. Norwegian and German shipping indices and air cargo rates in Hummels, 1999).

[42] Hummels infers transport costs by making use of more accurate data, but these data are only available for very few countries.

The problems of measuring trade costs that beset the empirical trade literature also apply to empirical studies into the relevance of NEG since any attempt to shed light on the empirical relevance of NEG calls for the availability of bilateral trade costs between a sufficiently large number of countries or regions (e.g. Redding and Venables, 2004; Hanson, 2005; Brakman et al., 2006; Knaap, 2006). Given the unavailability of a direct measurement of bilateral trade costs, all NEG studies turn to the indirect measurement of trade costs. In doing so, they closely follow the empirical trade literature (see Anderson and van Wincoop, 2004 for a very good survey of the latter) and assume a so-called trade cost function. This trade cost function aims to proxy the unobservable trade costs by combining information on observable trade cost proxies such as distance, common language, tariffs, adjacency, etc with assumptions about the unobservable trade cost component. The assumptions made about this trade cost function, e.g. functional form, parameter hetero- or homogeneity across country pairs, which observable cost proxies to include or how to estimate each cost proxy's effect, all potentially have a (crucial) effect on the results of any empirical NEG study.

In the empirical NEG literature, the measurement of trade costs is only a means to an end, and as a result the relevance of the preferred trade cost specification for the conclusions with respect to the NEG hypotheses under consideration is typically ignored. Virtually all studies just pick a trade cost specification and do not (or only marginally) address the sensitivity of their results to their chosen approximation of trade costs. This chapter aims to overcome this lack of attention by systematically estimating and comparing trade cost functions that have been used in the empirical NEG literature. We use various trade cost functions to estimate a standard NEG wage equation for 80 countries and look into the importance of the trade cost specification for the relevance of market access, the central NEG variable when it comes to inter-regional spatial interdependencies, as a determinant of the difference in gdp per capita between countries. It turns out that the way trade costs are proxied has substantial effects when it comes to the conclusions about the relevance of market access. Trade costs matter not only in terms of the size (and sometimes also the significance) of the market access effect, but also in terms of the spatial reach of economic shocks. The main message of this chapter is that the empirical specification of trade costs really matters for the conclusions reached with respect to the empirical relevance of NEG models.

The chapter is organized as follows. In the next section we first introduce the basic NEG model with a focus on the equilibrium wage equation and the role of trade costs. Next we discuss the two estimation strategies that have been used in the literature to estimate the wage equation, both of which require the specification of a trade cost function. Section 2.3 discusses the main conceptual difficulties involved in the approximation of trade costs by specifying a trade cost function. Given these difficulties, section 2.4 introduces a third way to approximate trade costs, which does not require a trade cost function but instead infers bilateral trade costs directly from bilateral trade data. Section 2.5 introduces our data set. In section 2.6 we present our estimation results for the NEG wage equation, hereby focusing in detail on the impact of the choice of trade cost approximation on the key explanatory NEG variable, market access. It turns out that the relevance of market access, or in other words of spatial interdependencies between countries, depends strongly on the choice of trade cost approximation. Section 2.7 concludes.

2.2 TRADE COSTS AND THE WAGE EQUATION IN NEG

The need to have a measure of trade costs when doing empirical work on NEG models immediately becomes clear when discussing the basic NEG model on which most empirical studies are based. This section first develops the theory behind the widely used wage equation that serves as the vehicle for our empirical research too (e.g. Krugman, 1991; Venables, 1996 and Puga, 1999). Our exposition is largely based on the seminal paper by Redding and Venables (2004)[43]. It overlaps to a large extent with the model discussed in chapter 1. But, as Redding and Venables (2004) are not so much interested in the exact analytical properties of the model, it is somewhat more flexible and geared towards empirical use, for example allowing explicitly for productivity differences between countries so that one can neatly incorporate an error term or additional variables that may explain wage differences besides market access. In the second part of this section, we move from theory to empirics by introducing the two different estimation strategies that have to date been used to estimate the parameters of the NEG wage equation (see Head and Mayer, 2004), focusing explicitly on the specification of trade costs.

2.2.1 The basic NEG model

Assume the world consists of $i = 1,...,R$ countries, each home to an agricultural and a manufacturing sector[44]. In the manufacturing sector firms operate under internal increasing returns to scale, represented by a fixed input requirement c_iF and a marginal input requirement c_i. Each firm produces a different variety of the same good under monopolistic competition using the same Cobb-Douglas technology combining three different inputs. The first is an internationally immobile primary factor (labor), with price w_i and input share β, the second is an internationally mobile primary factor with price v_i and input share γ and the third is a composite intermediate good with price G_i and input share α, where $\alpha + \gamma + \beta = 1$.

Manufacturing firms sell their variety to all countries and this involves shipping the goods to foreign markets. This is where the trade costs come in, these are assumed to be of the iceberg-kind and the same for each variety produced, i.e. in order to deliver a quantity $x_{ij}(z)$ of variety z produced in country i to country j, $x_{ij}(z)T_{ij}$ has to be shipped from country i. A proportion (T_{ij}-1) of output 'is paid' as trade costs ($T_{ij} = 1$ if trade is costless). Note that this relatively simple iceberg specification (introduced mainly for ease of modeling purposes) does not specify in any way what trade costs are composed of. It is precisely the need to specify T_{ij} more explicitly in empirical research, see below, that motivates this chapter. Taking these shipping costs into account gives the following profit function for each firm in country i,

[43] See their paper, or Puga (1999) and Fujita et al. (1999) ch.14, for more details on these types of models.

[44] In the theoretical exposition, countries are used as the geographical unit of interest. Instead of countries we could have taken any other geographical level of aggregation, e.g. regions, cities, districts, counties, or provinces.

$$\pi_i = \sum_j^R p_{ij}(z)x_{ij}(z)/T_{ij} - G_i^{\alpha} w_i^{\beta} v_i^{\gamma} c_i [F + \sum_j^R x_{ij}(z)] \tag{2.1}$$

where $p_{ij}(z)$ is the price of a variety produced in country i.

Turning to the demand side, each firm's product is both a final (consumption) and an intermediate (production) good. It is assumed that these products enter both utility and production in the form of a CES-aggregator with σ the elasticity of substitution between each pair of product varieties. Given this CES-assumption about both consumption and intermediate production, it follows directly that in equilibrium all product varieties produced in country i are demanded by country j in the same quantity (for this reason varieties are no longer explicitly indexed by *(z)*). Denoting country j's expenditure on manufacturing goods (coming from both firms and consumers) as E_j, country j's demand for each product variety produced in country i can be shown to be (following utility maximization and cost minimization on behalf of consumers and producers respectively),

$$x_{ij} = p_{ij}^{-\sigma} E_j G_j^{(\sigma-1)} \tag{2.2}$$

where G_j is the price index for manufacturing varieties that follows from the assumed CES-structure of both consumer and producer demand for manufacturing varieties. It is defined over the prices, p_{ij}, of all goods produced in country $i = 1,...,R$ and sold in country j,

$$G_j = \left[\sum_i^R n_i p_{ij}^{1-\sigma} \right]^{1/(1-\sigma)} \tag{2.3}$$

Maximization of profits (2.1) combined with demand as specified in (2.2) gives the well-known result in the NEG literature that firms set their f.o.b. price at p_i, i.e. depending only on the location of production (so that price differences between countries of a good produced in country i only arise from differences in trade costs, i.e. $p_{ij} = p_i T_{ij}$), where p_i is a constant markup over marginal costs:

$$p_i = G_i^{\alpha} w_i^{\beta} v_i^{\gamma} c_i \sigma / (\sigma - 1) \tag{2.4}$$

Next, free entry and exit drive (maximized) profits to zero, which pinpoints equilibrium output per firm at $\bar{x} = (\sigma - 1)F$. Finally combining this equilibrium output with equilibrium price (2.4) and equilibrium demand (2.2), and noting that in equilibrium the price of the internationally mobile primary factor of production will be the same across countries ($v_i = v$ for all i), gives the equilibrium wage of the composite factor of immobile production, i.e. labor,

$$w_i = A G_i^{-\alpha/\beta} c_i^{-1/\beta} \left(\sum_j^R E_j G_j^{(\sigma-1)} T_{ij}^{(1-\sigma)} \right)^{\frac{1}{\beta\sigma}} \tag{2.5}$$

where $A = v^{-\gamma/\beta} \left[(\sigma-1)^{\frac{\sigma-1}{\sigma}} F^{-1/\sigma} / \sigma \right]^{1/\beta}$ is a constant.

Equation (2.5) is the wage equation that lies at the heart of those empirical studies in NEG that try to establish whether, as equation (2.5) indicates, there is a spatial wage structure with wages being higher in economic centers (e.g. Brakman et al., 2006; Knaap, 2006; Redding and Venables, 2004; Mion, 2004 and Hanson, 2005). More precisely, the wage equation (2.5) says that the wage level a country is able to pay its manufacturing workers is a function of that country's technology, c_i, the price index of manufactures in that country, G_i, and so called real market access, the sum of trade cost weighted market capacities[45].

Note that trade costs play a crucial role in (2.5), most visibly in the real market access term. It also plays a role in the price index of manufactures (2.3), i.e. using $p_{ij} = p_iT_{ij}$:

$$G_j = \left[\sum_i^R n_i p_i^{1-\sigma} T_{ij}^{1-\sigma} \right]^{1/(1-\sigma)} \tag{2.6}$$

Wages are relatively high in countries that have easier access to consumer markets in other countries when selling their products and that have easier access to products produced in other countries (producer markets). The lower trade costs, the easier access to both producer and consumer markets abroad, the higher the wages that firms can offer their workers to remain profitable. Trade costs are thus of vital importance in determining the spatial distribution of income.

We now turn to the discussion of the two different ways by which the wage equation has been estimated in the literature so far. Hereby particularly emphasizing the way in which trade costs are dealt with.

2.2.2 *Estimating the wage equation*

Taking logs on both sides of (2.5) gives the following non-linear equation that can be estimated:

$$\ln w_i = \alpha_1 + \alpha_2 \ln G_i + \alpha_3 \ln \left(\sum_j^R E_j G_j^{(\sigma-1)} T_{ij}^{(1-\sigma)} \right) + \eta_i \tag{2.7}$$

where η_i captures the technological differences, c_i, between countries that typically consists of both variables that are correlated (modelled by including e.g. measures of physical geography or institutional quality) and/or variables that are uncorrelated (modeled by an *i.i.d.* lognormal disturbance term) with market and supplier access. The α's are the estimated parameters from which in principle the structural NEG parameters can be inferred (see e.g. Redding and Venables, 2004 or Hanson, 2005). There are basically two different ways in which the wage equation (2.7) has been estimated in the empirical NEG literature.

2.2.2.1 Direct non-linear estimation of the wage equation

The first empirical strategy to estimate the wage equation was introduced by Hanson (2005) and can be discussed rather briefly. It involves direct non-linear estimation of the wage

[45] The actual wage equation estimated may differ slightly from the one presented here in each particular empirical study, but the basic idea behind it is always the same, i.e. wages depend on real market access and the price index of manufactures, which both to a large extent depend on the level of trade costs between countries. Also, comparing (2.5) with (1.14) the differences between the model used here and that of chapter 1 clearly show. In contrast to (1.14) here wages also depend on a country's level of technology, c_i, and the parameters settings are slightly different given that the model used here has one additional production factor in manufacturing production compared to the model in chapter 1.

equation (2.7). Authors that have subsequently followed this direct non-linear estimation strategy include Brakman et al. (2004b; 2006), and Mion (2004)[46].

To deal with the unavailability of directly measurable trade costs, all papers in the 'Hanson-tradition' assume a trade cost function to deal with the need to specify T_{ij} for empirical research (see the next section)[47]. What is important here is that this trade cost function is subsequently directly substituted for T_{ij} in (2.7). Its parameters are jointly estimated along with the parameters of the wage equation. This is rather different from the second estimation strategy.

2.2.2.2 Two-step linear estimation of the wage equation making use of trade data

The second strategy comes from the work by Redding and Venables (2004) and involves a two-step procedure where in the first step the information contained in (international) trade data is used to provide estimates of so-called market and supplier capacity and bilateral trade costs that are subsequently used in the second step to estimate the parameters of the wage equation. Other papers using this strategy include inter alia Knaap (2006), Breinlich (2006), Head and Mayer (2006) and Hering and Poncet (2006).

Instead of directly estimating (2.7), this estimation strategy makes use of the following definition of bilateral trade flows between countries that follows directly from aggregating the demand from consumers in country *j* for a good produced in country *i* (2.2) over all firms producing in country *i*:

$$EX_{ij} = n_i p_i x_{ij} = n_i p_i^{1-\sigma} T_{ij}^{1-\sigma} E_j G_j^{\sigma-1} \qquad (2.8)$$

Equation (2.8) says that exports from country *i* to country *j* depend on the 'supply capacity', $n_i p_i^{1-\sigma}$, of the exporting country that is the product of the number of firms and their price competitiveness, the 'market capacity', $E_j G_j^{\sigma-1}$ of the importing country and the magnitude of bilateral trade costs T_{ij} between the two countries. Taking logs on both sides of (2.8) and replacing market and supply capacity by an importer and exporter dummy respectively, i.e. $s_i = n_i p_i^{1-\sigma}$ and $m_j = E_j G_j^{\sigma-1}$, results in the following equation that is estimated:

$$\ln EX_{ij} = \ln s_i + (1-\sigma)\ln T_{ij} + \ln m_j + \varepsilon_{ij} \qquad (2.9)$$

where ε_{ij} is an *i.i.d.* lognormal disturbance term.

In the second step, the estimated country specific importer and exporter dummies and the predicted value of bilateral trade costs that result from the estimation of (2.9) are then used to construct so-called market and supplier access. These are defined as follows respectively, see Redding and Venables (2004, pp. 61-62) for more details:

[46] The first version of this paper was already available as an NBER working paper (nr.6429) in February 1998. This explains why others have used his methodology and have published their work earlier than Hanson himself.

[47] Besides information on trade costs, T_{ij}, also the data on the price index, G_i, is unavailable at the regional level. Very briefly, the problems with the lack of data on regional price indices are solved by either using, besides the wage equation, other (long run) equilibrium conditions (Hanson, 2005; Brakman et al., 2004b and Mion, 2004) *or* by assuming away the use of intermediates in manufacturing production (α = *0*) and approximating each region's price index by the average wage level in the economic centers that are closest to that region (Hanson, 2005 Brakman et al., 2006), see also Head and Mayer, 2004, p. 2624.

$$MA_i = \sum_j^R E_j G_j^{\sigma-1} T_{ij}^{1-\sigma} = \sum_j^R m_j T_{ij}^{1-\sigma}$$
$$SA_j = \sum_i^R n_i p_i^{1-\sigma} T_{ij}^{1-\sigma} = \sum_i^R s_i T_{ij}^{1-\sigma} \tag{2.10}$$

The predicted values of market and supplier access are subsequently used to estimate the wage equation, i.e. rewriting (2.5), using (2.6) and (2.10) and taking logs on both sides gives:

$$\ln w_i = \alpha_1 + a_2 \ln SA_i + \alpha_3 \ln MA_i + \eta_i \tag{2.11}$$

where η_i, α_1 and α_3 are as specified in (2.7) and a_2 captures a somewhat different combination of structural parameters than α_2 in (2.7).

The problem of the unavailability of a direct measurement of trade costs when using this estimation strategy enters in the first step. All papers solve this problem by assuming a trade cost function (see next section). The parameters of this trade cost function are jointly estimated with the importer and exporter dummies and subsequently used in the construction of the predicted values of market and supplier access. Different from the direct estimation of the wage equation, the parameters of the distance function are thus not jointly estimated with the parameters of the NEG wage equation.

The motivation for Redding and Venables (2004) to use this 2-step strategy, is that "this approach has the advantage of capturing relevant country characteristics that are not directly observable but are nevertheless revealed through trade performance" (Redding and Venables, 2004, p. 75). Still, they have to assume an empirical specification for the trade cost function, and moreover the country dummies may be capturing 'too much' relevant country characteristics (see section 2.3 for more detail). In section 2.4 we, following Head and Ries (2001), will take the idea that actual trade data can be used as a foundation for market and supplier access in the wage equation one step further by letting trade data determine the total trade costs thereby circumventing the need to explicitly specify the trade function T_{ij}. But before doing so, we first discuss the main important assumptions, often implicitly made, that are involved when one approximates T_{ij} by making use of a trade cost function.

2.3 THE TRADE COST FUNCTION

All papers using either the direct or two-step estimation strategy deal with the unavailability of a direct measure of trade costs by specifying a trade cost function. In its most general form the trade cost function is:

$$T_{ij} = f(X_{ij}, X_j, X_i, \upsilon_{ij}) \tag{2.12}$$

The trade costs involved in shipping goods from country *i* to country *j* are a function *f* of cost factors that are specific to the importer or the exporter (X_j and X_i respectively), such as infrastructure, institutional setup or geographical features of a country (access to the sea, mountainness), bilateral cost factors related to the actual journey from *j* to *i*, X_{ij}, such as transport costs, tariffs, sharing a common border, language barriers, membership of a free trade union, etc, and unobservable factors, υ_{ij}. Given the aforementioned unavailability of transport cost data between a sufficient number of countries, these are in turn also proxied by most notably bilateral distance, but sometimes also actual travel times or population weighted distance measures are used.

The trade cost function that is used in estimating the wage equation in NEG studies, is typically chosen on the basis of the 'older' empirical literature on international trade, more specifically on the estimation of the so-called gravity equation of which (2.9) is an example (see Anderson and van Wincoop (2004) for an extensive discussion of the gravity equation). Usually, and probably mostly for ease of estimation (see Hummels, 2001), the trade cost function takes the following (multiplicative) form,

$$T_{ij} = \prod_{m=1}^{M} X_{ij}^{\gamma_m} \prod_{k=1}^{K} \left(X_i^{\gamma_{1k}} X_j^{\gamma_{2k}} \right) \upsilon_{ij} \tag{2.13}$$

where the unobservable part, υ_{ij}, of the trade cost function is modeled by a disturbance term (that is usually assumed to be *i.i.d.*). To give an idea about the type of trade cost function used in the NEG wage equation studies, Table 2.1 below shows the trade cost function used in several NEG papers.

As can be seen from Table 2.1, the trade cost function imposed differs quite a bit between these papers and between the 2 estimation strategies.

Table 2.1 Trade cost functions used in the empirical literature

paper	sample	trade cost function
		Direct estimation
Hanson (2005)	US counties	$T_{ij} = \exp(\tau D_{ij})$
Brakman et al. (2004b)	German regions	$T_{ij} = \tau^{D_{ij}}$
Brakman et al. (2006)	European regions	$T_{ij} = \tau D_{ij}^{\delta}$
Mion (2004)	Italian regions	$T_{ij} = \exp(\tau D_{ij})$
		Two-step estimation
Redding and Venables (2004)	World countries	$T_{ij} = D_{ij}^{\delta} \exp(\alpha B_{ij})$ or $T_{ij} = D_{ij}^{\delta} \exp(\alpha B_{ij}) \exp(\beta_1 isl_i + \beta_2 isl_j +$ $\beta_3 llock_i + \beta_4 llock_j + \beta_5 open_i + \beta_6 open_j)$
Knaap (2006)	US states	$T_{ij} = D_{ij}^{\delta} \exp(\alpha B_{ij})$
Breinlich (2006)	European regions	$T_{ij} = D_{ij}^{\delta} \exp\left(\alpha_1 L_{ij} + \sum_i \alpha_{2i} B_{ij}^{i} \right)$
Hering and Poncet (2006)	Chinese cities	$T_{ij} = D_{ij}^{\delta} \exp(\alpha_1 B_{ij}^{f} + \alpha_2 B_{ij}^{C} + \alpha_3 B_{ij}^{fC})$

Notes: D_{ij} denotes a measure of distance, usually great-circle distance, but sometimes also other measures such as travel times (e.g. Brakman et al., 2004b) or population weighted great-circle distance (e.g. Breinlich, 2006) have been used. B_{ij} denotes a border dummy, either capturing the (alleged positive) effect of two countries/regions being adjacent (e.g. Redding and Venables, 2004; Knaap, 2006) or the (possibly country-specific) effect of crossing a national border (e.g. Breinlich, 2006; Hering and Poncet, 2006).

Or, to quote Anderson and van Wincoop (2004): "A variety of ad hoc trade cost functions have been used to relate the unobservable cost to observable variables (p.706)" and "Gravity theory *(read: new economic geography theory)* has used arbitrary assumptions regarding functional form of the trade cost function, the list of variables, and regularity conditions (p.710, phrase in italics added)".

To a large extent based on Anderson and van Wincoop (2004), our discussion of the (implicit) assumptions underlying the use of a trade cost function concerns six issues: i) functional form, ii) variables included, iii) regularity conditions, iv) modeling costs involved with internal-trade, v) the unobservable component of trade costs, vi) estimating the trade cost function's parameters.

i) *functional form.* All papers in Table 2.1 *have* to assume a specific functional form for the trade cost function. As can be seen from Table 2.1, empirical papers in NEG opt for a functional form as shown in (2.13); all cost factors enter multiplicatively. As in the international trade literature (see Hummels, 2001), the main reason for doing so is probably ease of estimation. Although being by far the most common functional form used in the empirical NEG and the international trade literature, its implications are usually not given much attention. As pointed out by Hummels (2001), the multiplicative form implies that the marginal effect of a change in one of the trade cost components depends on the magnitude of all the other cost factors included in the trade cost function. As this may not be that realistic he argues that a more sensible trade cost function combines the different cost factors *additively*, i.e.

$$T_{ij} = \sum_{m}^{M} \gamma_m X_{ij} + \sum_{k}^{K} \left(\gamma_{1k} X_i + \gamma_{2k} X_j \right) + \upsilon_{ij} \qquad (2.14)$$

where X_{ij}, X_i, X_j and υ_{ij} are defined as in (2.12). Using this specification avoids the above-mentioned problem, as each cost factor's marginal effect does no longer depend on the magnitude of the other cost factors. In estimating the wage equation in section 2.6, we will therefore use both a multiplicative and additive trade cost function and check whether this makes a real difference or not.

Also the specific distance function chosen is of concern. Some papers take an exponential distance function (Hanson, 2005; Brakman et al., 2004b and Mion, 2004), hereby following the theoretical NEG literature (e.g. Fujita et al., 1999 and Krugman, 1995). The other papers shown in Table 2.1 opt for the power function instead, which is also the standard choice in the empirical trade literature. As argued by Fingleton and McCann (2007) the latter function has the virtue of allowing for economies of distances[48], so that transport costs are concave in distance (standard in the transportation and logistics literature, see e.g. McCann, 2001). The exponential distance function implies that transport costs are convex in distance, implicitly imposing a very strong distance decay, which may not be preferred (see Head and Mayer, 2004).

ii) variables included. The number and composition of variables included in the trade cost function differs quite substantially across the papers in Table 2.1. The papers employing

[48] When the estimated distance parameter is between zero and minus one.

the direct estimation strategy only include distance in the trade cost function. The impact of assuming a more elaborate trade cost function when applying the direct estimation strategy is shown in section 2.6. Studies employing the two-step estimation strategy usually also take other bilateral trade cost proxies into account besides distance, see the variables L_{ij} and B_{ij} in Table 2.1, capturing the effect of language similarity and the border effect respectively.[49]

When it comes to the inclusion of potentially relevant variables capturing country-specific trade costs, a drawback of the second estimation strategy as outlined in section 2.2 is that the inclusion of the importer and exporter dummies (recall equations (2.9) and (2.10)) wipes out all importer specific and exporter specific variation so that the effect of country-specific trade cost proxies cannot be estimated. As a result, the constructed market (supplier) access term (2.10) includes only the exporter (importer) specific trade costs and misses those trade costs specific to the importer (exporter)[50]. Implicitly all the papers using the two-step estimation strategy *cum* dummies approach mentioned in Table 2.1 assume that country-specific trade costs are zero. Redding and Venables (2004, pp.76-77) take note of this by also estimating the trade equation (2.9) without capturing the market and supplier capacity terms by importer and exporter dummies but by using importer and exporter GDP instead, hereby allowing for a more elaborate trade cost function (see Table 2.1). We will do the same in our estimations.

Besides the above discussion on which variables to include, also the way to measure a certain included variable differs between papers. The best example is the distance variable that shows up in all the assumed trade cost functions. Usually this is measured as great-circle distance between capital cities (e.g. Redding and Venables, 2004), but others have used great-circle distance between countries'/regions' largest commercial centres or counties'/regions' centroids, population weighted distances (e.g. Breinlich, 2006) or travel times (e.g. Brakman et al., 2004b). It is difficult to give a definitive answer to what measure of distance to include and the same applies for other variables (e.g. the border dummy, proxies of infrastructure quality). However, we think two recommendations can be made.

Regarding the general question which variables to include; the appropriateness of the inclusion of a certain variable can (and should) always be tested by assessing its significance. Second, one should be careful with the inclusion of variables that are very likely endogenous. Examples are travel times, population weighted distance measures, quality of infrastructure,

[49] Even though these papers include some more variables in the trade cost function, many additional variables have been shown to be of importance in the empirical trade literature. Examples are tariffs, colonial ties, quality of infrastructure, degree of openness, being member of a common currency union, the World Trade Organization or some preferential trade agreement (NAFTA, EU, Mercosur) and many more (see Anderson and van Wincoop, 2004).

[50] The estimated exporter/importer dummy would in this case also pick up the exporter/importer specific trade costs so that market and supplier access would look like (in case of a multiplicative trade cost function as in (2.13)):

$$M\hat{A}_i = \sum_j^r \left[\hat{m}_j \prod_{m=1}^{M} X_{ij}^{\hat{\gamma}_m} \right] \text{ and } S\hat{A}_j = \sum_i^r \left[\hat{s}_i \prod_{m=1}^{M} X_{ij}^{\hat{\gamma}_m} \right]$$

The use of exporter/importer dummies implies that m_j only captures the trade costs specific to country j and s_i only the trade costs specific to country i. As a result MA_i and SA_j fail to capture the trade costs specific to country i and country j respectively.

institutional setup or even being member of a free trade union. Especially when estimating the parameters of the NEG wage equation, that is itself already (by construction) plagued by endogeneity issues, adding more endogeneity through the trade cost function should in our view be avoided (or properly addressed but this is usually not so easy). The use of proxy variables such as great-circle distance, border and language variables and countries' geographical features such as having direct access to the sea, that can more confidently be considered to be exogenous, should be preferred.

iii) regularity conditions. All papers in Table 2.1, implicitly or explicitly, make assumptions about the extent to which the impact of each variable included in the trade cost function is allowed to be different for different (pairs of) countries. Most papers assume that the effect of distance, sharing a common border or trading internationally on trade costs is the same for all countries or regions included in the sample. It is however likely that there exists some heterogeneity in the effect of different cost factors (see e.g. Limao and Venables, 2001). Some authors do allow these effects to differ between countries or regions (e.g. Breinlich, 2006 and Hering and Poncet, 2006) but usually do so by imposing ad hoc assumptions regarding the way they are allowed to differ[51]. An advantage of the assumption(s) made about the regularity conditions (compared to e.g. assumptions about functional form) is that they can be tested. This has so far not been done; we argue that this should receive much more attention.

iv) internal trade costs. The modeling of the costs associated with within-country trade is another "problematic" feature in the empirical NEG papers[52]. The need to incorporate some measure of internal trade costs follows directly from the functional form of the wage equation (2.5). There it is the sum of trade cost weighted market capacities (real market access) that consists of on the one hand *foreign* real market access, $\sum_{j\neq i}^{R} E_j G_j^{(\sigma-1)} T_{ij}^{(1-\sigma)}$ but also of *domestic* real market access, $E_i G_i^{(\sigma-1)} T_{ii}^{(1-\sigma)}$, which is a measure of own market capacity weighted by internal trade cost. Theoretically these internal trade costs are usually set to zero ($T_{ii} = 1$). In contrast, all empirical NEG papers proxy internal trade costs by using an internal trade cost function that solely depends on so-called internal distance, D_{ii}, excluding other country specific factors that could influence internal trade costs (see Redding and Venables, 2004, p. 62). More formally:

$$T_{ii} = f(D_{ii}) \text{, where almost exclusively } D_{ii} = 2/3\left(area_i/\pi\right)^{1/2} \qquad (2.15)$$

This often-used specification of D_{ii} reflects the average distance from the center of a circular disk with $area_i$ to any point on the disk (assuming these points are uniformly distributed on the disk). Basically own trade costs are simply a function of a country's or region's area, the larger the country or region, the higher the internal trade costs.

[51] Note that assuming the effect to be the same for all countries/regions is also an ad hoc assumption.

[52] Some empirical papers in the international trade literature also deal with this issue (e.g. Helliwell and Verdier, 2001) focusing largely on how to measure internal distances, but in general internal trade costs are not dealt with in this literature.

Also most papers, regardless of estimation strategy, do not allow internal distance to have a different effect than bilateral distance (an exception are Redding and Venables (2004), who make the ad hoc assumption that the internal distance parameter is half that of the bilateral distance parameter). In section 2.6 we explicitly *estimate* a different parameter on internal and bilateral trade and allow own trade costs to depend on other factors than simply internal distance. This and the use of own trade data gives us some indication into the (un)importance of explicitly modeling internal trade costs.

v) the unobservable component of trade costs. In the direct estimation strategy this component is ignored, thereby implicitly positing that the assumed trade cost function is *the* actual trade cost function (Breinlich, 2006, also notes this). Taking account of the unobserved component using this strategy is not straightforward. Even if the unobservable component is assumed to be of the simplest kind, i.e. distributed *i.i.d.* and uncorrelated with any other component of either the trade cost function or the wage equation, the non-linear fashion in which it enters the wage equation makes it difficult to determine the appropriateness of the inference on the structural parameters when simply assuming it away (or equivalently assuming it is nicely incorporated into the error component of the wage equation itself). Using simulation based inference methods could (and maybe should) be a way to shed more light on this issue.

When using the two-step estimation strategy the unobserved trade cost component(s) is (are) more explicitly taken into account. They are usually assumed to be uncorrelated with the other (observable) trade cost components and to be independent draws from a lognormal distribution, so that they can be incorporated as a (possibly heteroscedastic) normal error term in the first step estimation of the gravity equation (2.9). Next the use of bootstrapped standard errors in the 2nd step estimation aims to take account of the fact that the market and supplier access terms (constructed on the basis of the estimated parameters of the first step) implicitly contain the unobservable trade cost component as well, i.e. they are both generated regressors.

vi) estimating the trade cost function's parameters and dealing with zero trade flows. This is only an issue when using the two-step estimation strategy, where, as explained in section 2.2, the parameters of the trade cost function are estimated in the first step by making use of a gravity-type equation. A well-known problem with the estimation of gravity equations is the presence of a substantial number of bilateral trade flows that are zero (i.e. countries not trading bilaterally at all). To deal with this different estimation strategies have been put forward. These can be grouped into two categories, i.e. those estimating the loglinearized trade equation (2.9) and those estimating the non-linear trade equation (2.8). Because taking logs of the zero trade flows is problematic, the loglinearized version of the trade equation (2.9) is usually estimated using OLS and the non-zero trade flows only, or, by first adding 1 (or e.g. the smallest non-zero trade flow) to all or only the zero trade flows, and subsequently estimating the trade cost function's parameters by OLS or Tobit. When estimating the non-linear trade equation (2.8) instead, either NLS (Coe et al., 2002) or the recently proposed Poisson pseudo maximum likelihood (PPML) estimator (Santos Silva and Tenreyro, 2006) can be used, in this case all trade flows can be used (the zero trade flows can now also be used as there is no need to take logs). Arbitrarily adding 1 (or some other positive number) to trade flows in order to be able to take logs of all (also the zero) trade flows is in our view highly unsatisfactory. The subsequent results obtained depend quite strongly on the actual amount that is added to the zero trade flows. Using the non-linear techniques solves this issue and can therefore be considered as a preferred way to estimate the trade function's parameters. This is why we opt for the estimation of the non-linear trade equation (2.8) using the PPML estimator[53].

To summarize, Table 2.2 lists the issues that one has to face when approximating trade costs by a trade cost function, while also providing possible ways to deal with these issues.

Table 2.2 Trade cost functions and the two estimation strategies

		Ability do deal with issue raised	
issue	**(possible) solution**	**Two-step**	**Direct**
functional form	experiment with different functional forms	+	- (non-linearity)
regularity conditions	test the assumptions	+	+
variable inclusion	significance of inclusion can be tested	+ - (difficulty with exporter /importer specific trade costs when using dummies)	+
variable measurement	include exogenous variables	+	+
internal trade cost	include more than simply area	+ - (unavailability of internal trade data)	+
unobservable component	most hidden issue, deserves more explicit care	+	- (implicitly assumed away)
estimating the parameters	non-linear estimation techniques (PPML, NLS) or two-step estimation	- (non-zero trade flows)	+ (NLS should do the job, given the other assumptions)

Notes : + and – indicates the ability of the corresponding estimation strategy to deal with the issue raised w.r.t. to the choice of trade cost function that is used (compared to the other strategy).

2.4 A THIRD ESTIMATION STRATEGY: IMPLIED TRADE COSTS

Now that we have, at some length, discussed the (implicit) choices one has to make when using a trade cost function to approximate trade costs, T_{ij}, in this section we introduce a third option where the need to specify a trade cost function does not arise. This is based on Head and Ries (2001) who provide a clever way to approximate trade costs. Using the trade equation (2.8) and making two important assumptions (see below), they show that trade costs can directly be inferred from bilateral trade data in the following way:

[53] An alternative method that also takes account of the issue of the zero trade flows is a two-step estimation procedure (see e.g. Helpman, Melitz and Rubinstein, 2007; Bosker, 2007b; and chapter 3). Poisson estimation is known to suffer from the so-called excess zero problem, i.e. it underpredicts the number of zero observations. When faced with a sample containing a large number of zeroes, two-step estimation can be preferable.

$$T_{ij}^{1-\sigma} \equiv \varphi_{ij} = \sqrt{\frac{EX_{ij}EX_{ji}}{EX_{ii}EX_{jj}}} \tag{2.16}$$

where EX_{ij} denotes imports of country j from country i and EX_{ii} denotes the total amount of goods consumed in country i that is also produced in country i. Moreover φ_{ij} is introduced for notational convenience as a measure of the so-called 'free-ness' of trade (see Baldwin et al., 2003). It ranges from 0 to 1, with 0 meaning prohibitive and 1 meaning completely free trade. Head and Ries (2001) use this method to construct implied trade costs for bilateral trade (disaggregated at the industry level) between the US and Canada. They show the gradual decline in trade costs over time and use regression methods to decompose it into a tariff and a non-tariff barrier component. Other papers that have also used (2.16) to construct implied trade costs are Head and Mayer (2004) and Brakman et al. (2006); they subsequently use them as a comparison to the theoretical breakpoints following from NEG models or, as Head and Ries (2001) do, to follow their evolution over time.

We argue that (2.16) can also be used in the estimation of the wage equation. Instead of proxying trade costs by making use of a trade cost function, resulting in the (implicit) assumptions summarized by Table 2.2, implied trade costs provide an interesting alternative. But the use of implied trade costs (unfortunately) also has its problems. First, there is the additional data requirement. As can be readily seen from (2.16) the construction of implied trade costs requires the availability of trade with oneself, EX_{ii}, for all countries in the data set. Own-trade data are usually not readily available, but when both total export and production data are available they can be constructed as a country's or region's total own production minus exports (see e.g. Head and Mayer, 2004; Head and Ries, 2001; Hering and Poncet, 2006). It is only in the complete absence of bilateral trade data, as is typically the case when working with data at the regional level (e.g. in the case of Europe, see Breinlich, 2006), that implied trade costs cannot be calculated at all.

Secondly, and turning to the implicit assumptions made when constructing implied trade costs using (2.16), two assumptions are needed in order for the implied trade cost approach to work. They follow directly from the way these implied trade costs are calculated. Substituting (2.8) for both bilateral and internal trade, we arrive at (2.16) in the following way:

$$\sqrt{\frac{EX_{ij}EX_{ji}}{EX_{ii}EX_{jj}}} = \sqrt{\frac{T_{ij}^{1-\sigma}T_{ji}^{1-\sigma}}{T_{ii}^{1-\sigma}T_{jj}^{1-\sigma}}} \underset{(assumption\ A)}{=} \frac{\sqrt{T_{ij}^{1-\sigma}T_{ji}^{1-\sigma}}}{C} \underset{(assumption\ B)}{=} \frac{T_{ij}^{1-\sigma}}{C} = \frac{\varphi_{ij}}{C} \tag{2.17}$$

Where the following two assumptions are made:

$$\begin{array}{lll} (A) & T_{ii} = C & \forall i \\ (B) & T_{ij} = T_{ji} & \forall i, j \end{array} \tag{2.18}$$

That is to say, internal trade costs are a constant and the same for each country[54] (2.18A) and trade costs involved when shipping from country i to country j are the same as shipping from

[54] Note that if $C = 1$, we get the expression for φ_{ij} only as in (2.16). This is done by Head and Ries (2001), but it requires the additional assumption that internal trade is costless. One does not need the assumption that $C = 1$ in

country *j* to country *i* (2.18B). Whether these assumptions are valid is an empirical matter to which we return in section 2.6 when discussing our estimation results. How do these two assumptions relate to the assumptions made when a trade cost function is used instead?

Table 2.3 Trade cost function vs. implied trade costs

issue	Trade cost function	Implied trade costs
functional form	assumed	not an issue
regularity conditions	ad hoc assumptions are (implicitly) made	**symmetry of bilateral trade costs**
variable inclusion	many candidates, which ones to include?	not an issue
variable measurement	no consensus, choices need to be made	not an issue
internal trade cost	assumed to depend on internal distance	**assumed to be the same for each country (see footnote 49)**
unobservable component	needs explicit care (additional assumptions)	implicitly taken into account
estimating the parameters	choice of estimation method not always straightforward	not necessary

Table 2.3 shows this on the basis of the six issues that were already discussed in section 2.2, with assumptions (2.18A) and (2.18B) in bold in Table 2.3.

The potential advantage of using implied trade costs clearly comes to the fore. But this verdict, of course, depends on the innocence of assumptions (2.18A) and (2.18B). As to the assumption of symmetric bilateral trade costs (2.18B), this assumption is also quite common in the empirical NEG studies that use a trade cost function. All the papers mentioned in Table 2.1 use a trade cost function that (implicitly) assumes symmetric bilateral trade costs (the only exception being the second trade cost specification from Redding and Venables, 2004 in Table 2.1). Arguably the most problematic assumption when using implied trade costs is assumption (2.18A). Although virtually all *theoretical* results in NEG are established using this assumption (they even assume $C = 1$, i.e. internal trade is costless), many authors (Anderson and van Wincoop, 2005; Head and Mayer, 2004) have stressed the importance of dropping this assumption when doing empirical work, as it is hard to believe that internal trade is as expensive in a small country like Luxembourg as it is in a large country like Canada. But given the other above-mentioned virtues of using implied trade costs, combined with the fact that theoretically these internal trade costs are also usually absent, we argue that they should be considered as a "third way" to deal with trade costs in empirical NEG studies.

The remainder of this chapter deals with the impact of using different ways to proxy trade costs when estimating the NEG wage equation using either the direct or two-step estimation strategy. Hereby we focus in particular on the way conclusions about real market access, a key NEG variable, may differ when using different methods to proxy trade costs.

order to be able to use implied trade costs when estimating the NEG wage equation (*C* ends up in the constant term), which is why we prefer to use of *C* in (2.18).

2.5 DATA

Our empirical results are based on a sample of 97 countries (see Appendix 2.A for a complete list of these countries) for the year 1996. In order to be able to estimate the wage equation, we have collected data on gdp, gdp per capita (as wage data is not available for all countries in our sample, we follow Redding and Venables (2004) and use gdp per capita as a proxy) and the price index of gdp (as a proxy for G_i in (2.7)[55]) from the *Penn World Tables*. We also need data to calculate the various trade cost proxies. To this end, we have collected data on bilateral distances, contiguity, common language, and indicators of a country being landlocked, an island nation, or a Sub-Saharan African country. All these variables are chosen because of their exogeneity (at least in terms of reverse causality). Complementing these data, we also need trade data to be able to calculate implied trade costs and to be able to infer the trade cost function's parameter(s) when using the two-step estimation strategy. These we have collected from the *Trade and Production 1976-1999 database* provided by the French institute CEPII[56], which enables the use of both bilateral trade and internal trade data for most of the countries in our sample.

2.6 ESTIMATION RESULTS: TRADE COSTS AND MARKET ACCESS

In line with the discussion so far, we discuss our results in two stages. First, we focus on inferring trade cost proxies from bilateral trade data. We estimate the parameters of the trade cost function and illustrate how the results differ when using different trade cost functions. We also calculate implied trade costs and check how much of these implied trade costs can be explained by a particular trade cost function. In the process we look into some evidence regarding the relevance of the assumptions made when using implied trade costs as a proxy for T_{ij} (recall assumptions (2.18A) and (2.18B) from section 2.4). Subsequently, we turn to our main point of interest, i.e. the way in which a particular trade cost approximation affects conclusions about the relevance of real market access in determining gdp per capita. This is done by estimating the wage equation (2.7) using different versions of market access constructed on the basis of a different trade cost approximation, and observing whether or not the size and significance of the parameter on market access (α_3) depends on the particular trade cost proxy used. Moreover, we look at how the choice of trade cost specification affects the spatial reach of economic shocks to market access by simulating the effects of a 5% gdp shock in Belgium on gdp per capita in other countries.

Besides implied trade costs, we distinguish between four different types of trade cost functions, Table 2.4 shows these four trade cost functions. The first two trade cost functions are chosen as they are the ones used by the two papers (Redding and Venables, 2004 and Hanson, 2005) that introduced the two-step and direct estimation strategy respectively. The multiplicative function is chosen as it allows trade costs to depend not only on bilateral variables but also on importer/exporter specific trade cost factors, more specifically those associated with being landlocked (llock), being an island nation (isl) and being a Sub-Saharan

[55] Note that theoretically the price index should only refer to that of tradable goods. Using the overall price index as a proxy does also capture the price of nontradables.

[56] http://www.cepii.org/anglaisgraph/bdd/TradeProd.htm

African country (ssa). As mentioned before, such a multiplicative function is quite common in the empirical trade literature (see e.g. Limao and Venables, 2001).

Table 2.4 Trade cost functions used

Abbreviation	trade cost function
RV 2004	$T_{ij} = D_{ij}^{\delta} \exp(\alpha B_{ij})$
Hanson 2005	$T_{ij} = \exp(\tau D_{ij})$
multiplicative	$T_{ij} = D_{ij}^{\delta} \exp(\alpha_1 B_{ij} + \alpha_2 L_{ij}) \exp(\beta_1 isl_i + \beta_2 isl_j +$ $\beta_3 llock_i + \beta_4 llock_j + \beta_5 ssa_i + \beta_6 ssa_j + \beta_7 ssa_{ij})$
additive	$T_{ij} = D_{ij}^{\delta} + \alpha_1 B_{ij} + \alpha_2 L_{ij} + \beta_1 isl_i + \beta_2 isl_j +$ $\beta_3 llock_i + \beta_4 llock_j + \beta_5 ssa_i + \beta_6 ssa_j + \beta_7 ssa_{ij}$

Notes: See Table 2.1 for the definition of the variables.

We also consider the additive function to address the critique (Hummels, 2001) on the use of a multiplicative trade function. We also allow for the distance parameter to be different for bilateral and internal distance, hereby estimating (instead of imposing) the different impact of distance when considering intra- vs. international trade. Estimating instead of assuming the coefficient on internal distance should in our view be preferred compared to making ad hoc assumptions about it.

2.6.1 Inferring trade costs from trade flows

2.6.1.1 Inferring the trade cost function's parameters from trade flows

To infer the trade cost function's parameters from bilateral (and internal) trade flows we estimate equation (2.8) by using the PPML estimation strategy[57]. This estimation strategy is, see section 2.3, able to take account of the zero trade flows in a way that (contrary to NLS) also deals with the heteroscedasticity that is inherently present in trade flow data (see Santos Silva and Tenreyro, 2006). It gives the PPML method an advantage over the heavily used Tobit and/or OLS methods[58]. Table 2.5 shows our estimation results.

To allow for the more elaborate multiplicative and additive trade cost functions we (following Redding and Venables, 2004, p. 76, see their equation (22)) substituted the importer and exporter dummies with importer and exporter GDP[59]. In all specifications the distance coefficient is significant: the further apart two countries, the higher trade costs. Also

[57] Except for the additive specification, where we use NLS due to the inability of the PPML method to readily perform non-linear poisson regressions.

[58] Note that PPML itself also requires assumptions that may not be met when dealing with international trade flows (most notably that the same process drives the zero and the non-zero observations). See e.g. Helpman et al. (2007), Bosker (2007b) or chapter 3 of this thesis.

[59] For sake of comparison we have also estimated the trade equation including importer and exporter dummies while using the corresponding RV trade cost function (eq. 16 in their paper) . The results were very similar to the results shown here when including importer and exporter gdp, also in terms of explanatory power (R^2) and in terms of the implication in the 2nd step estimation on the effect of market access. Results available upon request. For sake of comparison we do show the results when estimating the wage equation using the 1st-step results for the RV trade cost function with trade dummies as an input in section 2.6.2.

sharing a common border (contiguity) significantly lowers trade costs (except in the additive specification), a finding consistent with earlier studies (e.g. Limao and Venables, 2001 and Redding and Venables, 2004). When estimating the multiplicative specification, the results show the importance of also considering country-specific trade cost proxies. Being landlocked or a Sub-Saharan African country raises trade costs, whereas being an island lowers these costs. These findings are very much in line with the results reported in Limao and Venables (2001) and show that these country-specific trade cost proxies cannot a priori be ignored.

Table 2.5 Trade costs functions and trade flows – PPML estimation

	Trade flows			
Trade cost function:	RV	Hanson	multiplicative	additive
distance	-0.721	-0.0002	-0.712	-1.035
	0.000	0.000	0.000	0.000
internal distance	0.034	-0.001	0.050	-0.047
	0.745	0.000	0.650	0.676
contiguity	0.746	-	0.930	0.000
	0.000	-	0.000	0.424
common language	-	-	0.007	0.000
	-	-	0.962	0.523
landlocked importer	-	-	-0.441	0.000
	-	-	0.045	0.295
landlocked exporter	-	-	-0.257	0.000
	-	-	0.003	0.513
island importer	-	-	0.285	0.000
	-	-	0.000	0.284
island exporter	-	-	0.526	0.000
	-	-	0.000	0.307
ssa importer	-	-	-0.801	0.000
	-	-	0.000	0.281
ssa exporter	-	-	-1.052	0.000
	-	-	0.000	0.635
ssa importer and exporter	-	-	0.950	0.000
	-	-	0.004	0.692
gdp importer	0.751	0.743	0.733	1.070
	0.000	0.000	0.000	0.000
gdp exporter	0.851	0.838	0.841	0.976
	0.000	0.000	0.000	0.000
exporter dummies	no	no	no	no
importer dummies	no	no	no	no
own trade dummy	1.599	3.810	2.431	7.330
	0.008	0.000	0.000	0.216
(pseudo) R2	0.953	0.943	0.956	0.982
nr. obs.	8774	8774	8774	8774
importer = exporter?				
- landlocked	-	-	0.386	0.455
- island	-	-	0.192	0.502
- ssa	-	-	0.275	0.343

Notes: p-values underneath the coefficient; importer = exporter? shows the p-value of a test of equality of the importer and exporter variant of a certain country specific variable.

Things are rather different when considering the additive trade cost specification: except for distance no other variable is significant. We think that the increased non-linearity of the additive trade cost function is (at least partly) to blame for this. These estimation difficulties for the additive trade cost function are something to take explicit note of (we will return again to this issue when estimating the wage equation), and it limits in our view the usefulness of such a functional form.

Of explicit interest here is the coefficient on internal distance (the coefficient shown reflects the difference of the internal distance coefficient with that of bilateral distance). When using the RV, the multiplicative or the additive specification we find no evidence that internal distance affects trade costs significantly different than bilateral distance does. Only when using the Hanson specification we find that internal distance affects trade costs significantly different from bilateral trade, the estimated coefficient suggests that internal distance increases trade costs to a much larger extent than bilateral distance does, which is contrary to what one would expect.

2.6.1.2 Implied trade costs and trade cost functions

Next we turn to the alternative way to infer trade costs from trade data, namely by calculating implied trade costs as shown by equation (2.17) in section 2.4. Doing this for our sample leaves us with no less than 3808 observed bilateral φ_{ij}'s. How much of these implied trade costs can be accounted for by the four different trade cost functions used in the previous sub-section? Or, to put it differently, does the use of implied trade costs provide a proxy for trade costs that differs from the proxy obtained using the trade cost functions in Table 2.5? And what about the (ir)relevance of the underlying assumptions when calculating implied trade costs, see (2.18)?

Table 2.6 Trade cost functions and implied trade costs

	Implied trade costs (phi)			
Trade cost function:	RV	Hanson	multiplicative	additive
distance	-0.854	-0.0002	-0.861	-1.342
	0.000	0.000	0.000	0.000
contiguity	0.707	-	0.865	0.016
	0.139	-	0.016	0.150
common language	-	-	-0.235	0.001
	-	-	0.138	0.436
landlocked	-	-	-1.150	-0.003
	-	-	0.000	0.011
island	-	-	0.241	0.003
	-	-	0.127	0.001
ssa	-	-	-1.148	0.000
	-	-	0.000	0.624
ssa both	-	-	1.222	-0.002
	-	-	0.011	0.244
(pseudo) R2	0.113	0.058	0.144	0.221
nr. obs.	3808	3808	3808	3808

Notes: p-values underneath the coefficient. Landlocked takes the value 0 if neither country is landlocked, 1 if one of the two countries is landlocked and 2 if both countries are landlocked; similarly for island and ssa.

To answer the first question we regressed the bilateral φ_{ij}'s on each of the four different trade cost functions introduced in the previous section. Table 2.6 above shows the results that are again obtained using the PPML estimator to take account of the zeros in implied trade costs. Note that because of assumption (2.18B) we cannot split the country-specific trade cost proxies into an exporter and an importer part, so that each country-specific trade cost proxy enters only once.

As can be readily seen from the (pseudo) R^2 for each of the four regressions, the trade cost functions capture at most 22% of the variation in implied trade costs. In case of the "Hanson" trade cost function this is only 5%! Apparently, approximating trade costs through implied trade costs differs quite a bit from obtaining these costs by estimating the parameters of an a priori specified distance function. Note also that the inclusion of country-specific trade cost proxies does improve the fit[60]. The flexibility of implied trade costs (recall Table 2.3) as compared to the use of trade cost functions provides a useful alternative way to proxy trade costs in our view.

This last conclusion depends, however, crucially on the validity of the two assumptions underlying the calculation of implied trade costs; see (2.18A-B). To shed some light on the (ir)relevance of assumption (2.18B) the last three rows of Table 2.5 are instructive. As we mentioned before, when one believes that only symmetric bilateral trade cost proxies such as distance, contiguity and sharing a common official language matter, assumption (2.18B) is automatically satisfied. But we have been arguing that also country-specific proxies such as being landlocked are important to take into account. Allowing these proxies to have a different effect when engaging in import or export, as we have for example done in Table 2.5, implicitly violates (2.18B). This is so unless one cannot reject that the coefficients on the importer and exporter variant of a variable are the same. The results of performing such tests are shown in the last three rows of Table 2.5 and they provide an indication that indeed the assumption of symmetry as imposed by (2.18B) is not violated in case of the country-specific variables that we have included.

The other assumption, the same constant internal trade costs for all countries (2.18A), is probably a more problematic one. It seems likely that there are some differences across countries in the trade costs involved with internal trade. The results in Table 2.5 all suggest that internal distance can serve as a sufficient proxy for own trade costs. Those results should however be taken with some care. The fact that the observations on own trade are heavily outnumbered by the observations on bilateral trade (by almost a factor of 10) could have its effect on the regression outcomes. To check for this, we estimated each of the trade equations using only the data on internal trade. The results are shown in Table 2.7.

Except in case of the Hanson distance function, none of the included trade cost proxies is found to be significant in explaining the variation in internal trade. In case of the country-specific variables this may not be that surprising (why should it matter for a country's internal trade costs whether or not it is an island or landlocked?). What is most striking is that also internal distance, the widely used proxy for internal trade costs mostly turns out to be

[60] The inclusion of country dummies improves the fit even (results available upon request) further, suggesting that implied trade costs are capturing (unobserved) country-specific trade cost factors.

Table 2.7 Trade cost functions and internal trade

	internal trade			
Trade cost function:	RV	Hanson	multiplicative	additive
internal distance	-0.195	-0.001	-0.176	-0.162
	0.152	0.000	0.240	0.238
landlocked	-	-	-0.025	-0.052
	-	-	0.927	0.638
island	-	-	-0.070	-0.024
	-	-	0.836	0.848
ssa	-	-	-0.256	-0.107
	-	-	0.350	0.413
gdp importer	1.404	1.434	1.382	1.370
	0.000	0.000	0.000	0.000
(pseudo) R2	0.879	0.886	0.880	0.874
nr. obs.	93	93	93	93

Notes: p-values underneath the coefficient.

insignificant (again except in the Hanson specification)[61]. The results shown in Table 2.7 can, of course, not be taken as conclusive evidence that internal trade costs are indeed the same across countries and that assumption (2.18A) can therefore be taken for granted[62]. They do, however, serve as an indication that the way internal trade costs are proxied when specifying them within the trade cost function approach, is also far from straightforward. Proxying internal trade costs by a clever transformation of a region's or a country's area, as is done by virtually all empirical NEG papers, may be just as harmful as assuming them away.

2.6.2 Varying trade costs and the impact on market access

We are now finally in a position to turn to our main point of interest, the way in which the various trade cost approximations affect conclusions regarding the relevance of real market access in determining gdp per capita levels in our sample. To this end, we estimate wage equation (2.7) using both the direct and the 2-step estimation strategies introduced in section 2.2, i.e. to refresh our memory:

$$\ln w_i = \alpha_1 + \alpha_2 \ln G_i + \alpha_3 \ln\left(\sum_j^R E_j G_j^{(\sigma-1)} T_{ij}^{(1-\sigma)}\right) + \eta_i \qquad (2.7')$$

We focus on the size and significance of the parameter on market access (α_3) when using the different trade cost proxies (this section) as well as when looking at the spatial reach of economic shocks (next section). For the direct estimation strategy as developed by Hanson (2005), we use NLS to estimate the parameters whereby we proxy G_i by a country's price index and E_i by a country's gdp level. When using the 2-step estimation method as developed by Redding and Venables (2004), we construct market access as specified in (2.10) on the

[61] Considering that internal distance is merely capturing the area or the size of a country, this may indeed not be so surprising after all. Why should a larger country always face higher internal trade or transport costs (compare transportation within the USA against that of transportation within Sierra Leone)?

[62] For example, we do not take differences in infrastructural quality into account that are likely to have a big influence on internal trade costs and that do differ quite a bit across countries.

basis of the results shown in Table 2.5 and estimate (2.7) by simple OLS, again proxying G_i by a country's price index[63]. We could have instead used more sophisticated GMM or 2SLS techniques that have been used in the empirical NEG literature and/or have proxied G_i by for example a constructed measure of supplier access. But we decided to use OLS and a simple proxy of G_i, to be able to focus entirely on the effect of the trade cost proxy used on the estimated effect of market access. The use of more sophisticated ways of estimating (2.7') would make it far more difficult to ascribe different outcomes to the differences in the way trade costs are proxied. For the same reasons, we also assume that the technological differences between countries as measured by, η_i, can be adequately captured by a simple i.i.d. error term that is uncorrelated with the other regressors instead of also adding additional variables.

The results of the various estimations are shown in Table 2.8a.

Table 2.8a Market access and gdp per capita

Strategy:	2-step	2-step	2-step	2-step	2-step	direct	direct	direct	direct
Trade costs:	RV dum	RV	Hanson	multiplicative	additive	implied	RV	Hanson	multiplicative
α_3	0.509	0.634	0.303	0.642	0.512	0.231	0.262	0.236	0.248
	0.002	0.003	0.047	0.000	0.001	0.003	0.002	0.008	0.000
α_2	0.891	0.969	1.114	0.804	0.896	0.825	0.879	1.090	0.942
	0.000	0.000	0.000	0.000	0.000	0.000	0.002	0.000	0.000
R2	0.648	0.644	0.586	0.731	0.679	0.627	0.706	0.654	0.736
nr obs	80	80	80	80	80	80	80	80	80

Notes: p-values underneath the coefficient in case of 2-step estimation.

Each column gives first the estimation strategy used (2-step or direct) and below the trade cost approximation that was used. So *RV-dum* refers for instance to the Redding and Venables trade cost function with im- and exporter dummies and *RV* to the trade cost function (see Table 2.5) where gdp is used instead of the trade dummies. Similarly, *2-step/Hanson* (column 3 in Table 2.8a) indicates a 2-step estimation of (2.7') with the Hanson trade cost function.

The first thing to note is that *market access is always significant*[64]. But the size of the coefficient differs quite a bit across the trade cost proxies and the estimation strategies! The impact of a 1% increase in a region's market potential on gdp per capita ranges from a minimum of 0.23% when using implied trade costs to 0.64% when using the multiplicative trade cost function. When comparing results for each estimation strategy separately, the differences are smaller but still the impact of a 1% change in market access ranges from 0.23% to 0.26% (0.30% to 0.64%) when using the direct (2-step) estimation strategy[65]. Table 2.8b shows additional evidence on the impact of the type of trade cost proxy used. Here we

[63] Results are very similar when we exclude the price index, G_i, from (2.7').

[64] Notwithstanding differences in the exact specifications used our estimation results for market access in Tables 2.8a and 2.8b are at least for the RV case (columns 1 and 2) similar to those in Redding and Venables (2004), see for instance their Table 3.

[65] Note that the difference in size could also be due to the different ways in which market access is constructed. The thought experiment in the next section, which more explicitly describes the spatial reach of an income shock, implicitly shows, by calculating the marginal effects, that this is probably not the case.

abstracted from the thorny issue of internal trade costs and estimated the effect of only *foreign* market access (FMA), that is MA excluding a region's own internal distance weighted gdp, on a region's gdp per capita level. Hereby focusing more specifically on the way spatial interdependencies between countries matter for an individual country's prosperity.

Table 2.8b Foreign market access and gdp per capita

Strategy: *Trade* *costs:*	2-step RV dum	2-step RV	2-step Hanson	2-step multiplicative	2-step additive	direct implied	direct RV	direct Hanson	direct multiplicative
FMA	0.494	0.425	0.232	0.669	0.528	0.328	0.098	0.102	0.153
	0.022	0.031	0.132	0.001	0.014	0.002	0.133	0.201	0.071
a_2	1.117	1.142	1.203	0.958	1.078	0.839	1.092	1.120	1.042
	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
R2	0.601	0.592	0.571	0.708	0.633	0.672	0.628	0.607	0.645
nr obs	80	80	80	80	80	80	80	80	80

Notes: p-values underneath the coefficient in case of 2-step estimation.

As can be clearly seen from Table 2.8b the estimated impact of foreign market access differs much more than market access itself, a 1% increase in foreign market access raises gdp per capita from a mere 0.1% to 0.67% depending on the trade cost proxy used. This clearly indicates that the choice of trade cost specification makes quite a difference. Its impact is even estimated to be insignificant at the 5% level in 4 out of 8 cases. The latter is especially the case for the direct estimation strategy when estimating the trade cost parameters jointly with the NEG parameters (columns 7-9)[66].On the basis of these estimation results we conclude that both the size and significance of the effect of (foreign) market access on gdp per capita does depend on the type of trade cost approximation used. This is not the only way to illustrate why the empirical specification matters for NEG empirics. As trade costs are key to the strength of spatial interdependencies, the spatial or geographical reach of income shocks can potentially be very different when comparing different trade cost specifications.

2.6.3 Trade costs and the spatial reach of an income shock in Belgium

To address this issue we conduct the following thought experiment. Suppose that Belgium, a country in the heart of Europe, experiences a positive 5% gdp shock, to what extent will this shock, given our estimation results in Table 2.8, spill over to the other countries in our sample through the market access variable? The 5% increase in gdp increases the demand for goods from potentially all countries, however the actual magnitude of this increase in a specific country depends crucially on the strength of the spatial linkages and thus on the measurement of trade costs: i.e. the lower trade costs with Belgium, the larger the impact on a country's gdp per capita.

[66] As mentioned already in the previous section, this probably is to a large extent due to the non-linear estimation process. The use of more elaborate trade cost functions makes it even 'more non-linear' increasing the difficulties with pinpointing the parameters. When using the two-step approach this problem is overcome by using trade data to reveal additional information on the strength of regional interdependencies on the basis of which the trade cost function's parameters can be more easily estimated.

Based on the estimation results from Table 2.8a and the various trade cost approximations, we have calculated the resulting gdp per capita changes as experienced by all other countries in response to the increased demand for their products from Belgium. Table 2.9 shows the correlation between gdp per capita changes obtained using each of the different trade cost proxies and Figure 2.1 visualizes four of these correlations in some more detail.

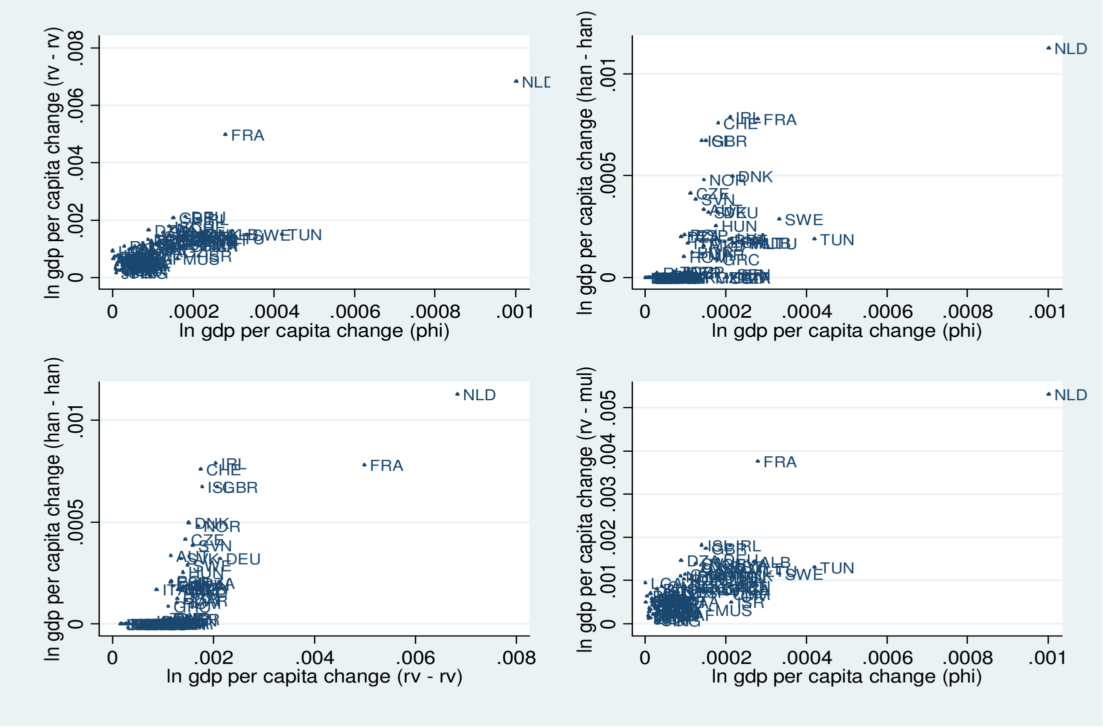
As can be seen from Table 2.9, the correlation between the gdp per capita changes differs markedly across the various different trade cost proxies. For some proxies the correlation is quite high (especially between the "RV" and the "multiplicative" trade cost function), but also in some cases the correlation is rather low or even insignificant. No systematic difference in correlations between estimation strategies used can either be detected.

Table 2.9 Correlations in response to a 5% GDP shock in Belgium

Strategy *-Trade costs:*	2step - RV	2step - Han	2step - multipl.	2step - add	Direct - RV	direct - Han	direct - multipl.	direct - φ
2step – RV	1	-	-	-	-	-	-	-
2step – Hanson	0.423	1	-	-	-	-	-	-
2step – multipl.	0.983	0.490	1	-	-	-	-	-
2step – additive	0.155	0.489	0.243	1	-	-	-	-
direct – RV	0.912	*0.104a*	0.880	*0.068a*	1	-	-	-
direct – Hanson	0.825	*0.422*	0.825	*0.149a*	0.676	1	-	-
direct – multipl.	0.887	*0.031a*	0.852	*0.021a*	0.992	0.655	1	
direct - φ	0.834	0.489	0.817	*0.102a*	0.700	0.671	0.684	1

Notes: - a - means not significant at the 5% level

Figure 2.1 Some correlations visualized

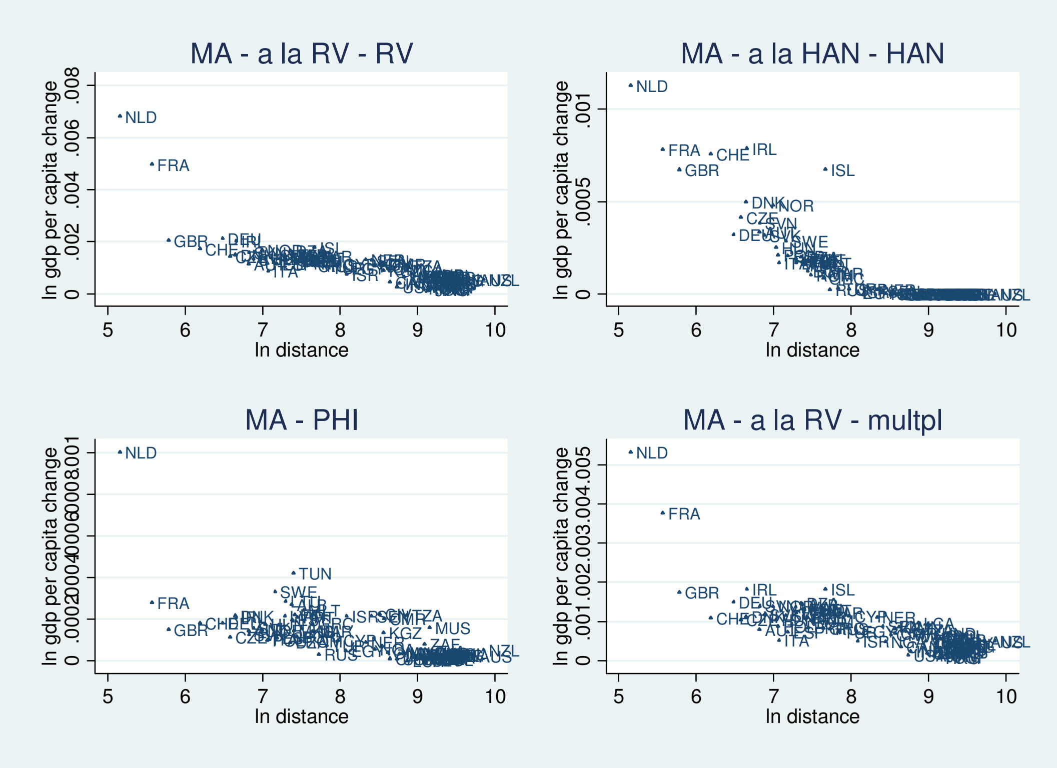


Notes: For the size of the correlations shown see Table 2.9.

Figure 2.1 complements this finding by plotting the different gdp per capita changes for four different trade cost specifications (i.e. Redding Venables (rv) with gdp in the trade cost function, Hanson (han), the multiplicative trade costs function (mul) and the implied trade costs (phi)) against each other. It illustrates in more detail that the spatial impact of a localized gdp shock differs quite a bit across trade cost specifications.

Next, Figure 2.2 focuses explicitly on the spatial reach of the Belgian GDP shock. It plots the percentage gdp per capita change against log distance for the same four different trade cost approximations as shown in Figure 2.1. Again the conclusions differ depending on the type of trade cost proxy used. In for instance the Hanson case (upper right panel) the distance decay is very strong and also the size of the gdp per capita changes is relatively small. Take for instance the case of the Netherlands (NLD). In the Redding and Venables specifications (upper left and lower right panels), the gdp per capita change is about 7 times as large as in the Hanson case.[67] Figure 2.2 also shows that a more elaborate or heterogeneous trade cost specification (see the two lower panels) increases the variation of the gdp shock for countries at a similar distance to Belgium. Moreover this heterogeneity in differences in spatial reach corresponds predictively to the type of trade cost specification used.

Figure 2.2 Spatial impact of 5% GDP shock in Belgium



Notes: The correlation of the shock and ln distance are, going from upper left to lower right, -0.77, -0.84, -0.67, -0.76 respectively.

[67] This seems to re-affirm the conclusion of Head and Mayer (2004, p. 2626) that the strong distance decay in Hanson (2005) "may be a consequence of the functional form of the distance decay function". See also Fingleton and McCann (2007).

When using the Hanson (2005) exponential trade cost function that depends solely on distance, the size and spatial reach of the Belgian income shock also depend heavily on distance. The correlation between the shock and log distance is the largest in this case. Moreover due to the exponential distance function used, the effect of the income shock quickly peters out; there is no discernable effect any more in countries lying farther from Belgium than Egypt (in the RV specification Egypt still experiences a 0.1% increase in gdp per capita; this is about the size of the wage increase in the most heavily affected country, the Netherlands, when using the Hanson specification).

The Redding and Venables (2004) specification also allows contiguity to have its effect on trade costs. Consequently the effect of the Belgian income shock is less correlated with distance and has a larger effect on its contiguous neighbors. The shock now for example has a larger effect on Germany than on the closer (as measured by distance between capital cities) United Kingdom. Also the less extreme distance decay as implied by the estimated distance coefficient (compared to the exponential Hanson specification) ensures that the income shock peters out much slower and affects all countries in the sample in some way (Japan, with only a 0.02% increase in its gdp per capita, is affected the least).

All countries are also affected when allowing for country specific trade cost factors in the multiplicative trade cost function. However, the correlation of the income shock with distance decreases somewhat further. Moreover the estimated positive or negative effect of being an island, landlocked or Sub-Saharan African country has a clear effect on the effect of the Belgian income shock. Landlocked Switzerland, and Sub-Saharan Côte d'Ivoire are relatively much less affected by the income shock for example, whereas island nations such as Ireland, New Zealand experience a relatively larger gdp per capita increase.

Finally when totally abstracting from the use of a trade cost function, by using implied trade costs, the effect of the Belgian income shock becomes even less correlated with distance. Given the way they are calculated the use of implied trade costs has the effect that countries that trade a lot with Belgium relative to their amount of internal trade[68], such as Côte d'Ivoire, Finland or Sweden, experience a larger impact of the Belgian income shock than countries that do not export/import a substantial amount of their total trade to Belgium but to other countries (e.g. Germany or Great Britain).

To sum up, from the evidence in Tables 2.8a, 2.8b and 2.9 as well as in Figures 2.1 and 2.2, we have to conclude that the way trade costs, a crucial element of any NEG model, are approximated when doing empirical work has the potential to shape the overall conclusions about the empirical relevance of NEG and thus of the strength and the geographical reach of spatial interdependencies. The lesson for future NEG research is therefore that the topic of trade costs should be given much more attention. Also the robustness of the results with respect to the use of a particular trade cost proxy warrants much more explicit attention.

[68] Belgian internal trade does also matter of course, but this matters in the same way for all countries. As a result differences in countries' φ_{ij} cannot be ascribed to Belgian internal trade.

2.7 CONCLUSIONS

Trade costs are a crucial element of new economic geography (NEG) models, without trade costs geography does not matter. The size of trade costs crucially determines the strength of regions' spatial interdependencies and thereby the relevance of market access. The unavailability of actual trade cost data hampers empirical research in NEG and it requires the approximation of trade costs. Notwithstanding the importance of trade costs in NEG models, most empirical NEG studies do not pay attention to the ramifications of the particular trade cost specification used. This chapter shows, both theoretically and empirically, that the way trade costs are specified matters. Estimations of an NEG wage equation for a sample of 80 countries shows that the relevance of the key NEG variable, market access, and hence of spatial interdependencies hinges nontrivially upon the trade cost specification. Using two estimation strategies and various trade cost specifications, the main conclusion is that NEG needs to (re-)examine the sensitivity of its empirical findings to the handling of trade costs.

New economic geography (NEG) is, of course, not the only theory around in spatial economics. But compared to urban and regional economics, NEG prides itself on the way it focuses on spatial interdependencies. These interdependencies come to the fore in the crucial role market access and trade costs play in NEG whereas market access (and trade costs) is for example typically given much less attention in urban economics (Combes, Duranton, and Overman, 2005, p. 320). Whether spatial interdependencies and hence market access and trade costs really matter is ultimately an empirical question. This chapter argues that the answer to that, from an NEG perspective, crucial question can depend rather strongly on the empirical specification of trade costs. Future empirical research in NEG should therefore pay far more attention to the way trade costs are dealt with. This chapter does not provide a definitive answer (or a clear-cut empirical test) to the question which of the discussed trade cost approximations should be preferred. It does however provide a useful overview of the ex- and/or implicit assumptions made when approximating trade costs using either of the approximations, hereby facilitating the choice of trade cost approximation when doing empirical work on NEG. A choice that will generally depend on the sample under consideration as the importance of particular trade cost components, and also data availability, differs substantially when considering countries, regions, cities or neighborhoods.

APPENDIX 2.A

Countries included

Albania
Algeria
Argentina
Armenia*
Australia
Austria
Bahamas*
Bangladesh*
Barbados*
Belgium
Bolivia
Brazil
Bulgaria
Cameroon
Canada
Cape Verde*
Chile
China
Colombia
Costa Rica
Côte d'Ivoire
Cyprus
Czech Republic
Denmark
Ecuador
Egypt
El Salvador
Estonia
Ethiopia*
Finland
France
Germany
Greece
Guatemala*
Honduras
Hong Kong
Hungary
Iceland
India
Indonesia
Ireland
Israel
Italy
Japan
Jordan*
Kenya*
Korea
Kuwait*
Kyrgyzstan
Latvia
Lithuania
Macau
Macedonia
Malawi*
Malaysia
Malta
Mauritius
Mexico
Moldova*
Mongolia
Morocco
Nepal*
Netherlands
New Zealand
Niger
Nigeria
Norway
Oman
Pakistan*
Panama
Peru
Philippines
Poland
Portugal
Romania
Russia
Saint Lucia
Senegal
Singapore
Slovakia
Slovenia
South Africa
Spain
Sri Lanka*
Sweden
Switzerland
Taiwan
Tanzania
Thailand
Trinidad and Tobago*
Tunisia
Turkey
United Kingdom
United States of America
Uruguay
Venezuela
Zimbabwe*

Notes: * means the country is excluded in the wage equation estimation.

Chapter 3

Economic geography and economic development in Sub-Saharan Africa[69]

3.1 INTRODUCTION

Sub-Saharan Africa (SSA) is home to the world's poorest countries. Alongside factors such as poor institutional quality, low (labour) productivity and low levels of human capital, the region's geographical disadvantages are often viewed as an important determinant of its dismal economic performance. It is well established that a country's geography may directly affect economic development through its effect on disease burden, agricultural productivity, and the availability of natural resources (see Gallup et al., 1999; Collier and Gunning, 1999; Ndulu, 2007). Geography can also indirectly affect economic development through its influence on institutional quality (Rodrik et al., 2004; Gallup et al., 1999) or by determining a country's transport costs (Limao and Venables, 2001; Amjadi and Yeats, 1995). Recently, the new economic geography (NEG) literature (see Krugman, 1991; Fujita et al, 1999) has, however, highlighted another mechanism through which geography could affect a country's prosperity. The NEG literature emphasizes the relevance of a country's so-called 2nd nature geography (its location relative to other countries) as the main determinant of access to international markets, and this market access in turn determines wages and income levels[70].

Redding and Venables (2004) were among the first to establish empirically that market access indeed matters for economic development. Based on estimation results for a sample of 101 countries, they find[71] for example that were Zimbabwe to be located in central Europe, the resulting improvement in its market access would ceteris paribus increase its GDP per capita by almost 80%. Similarly, halving the distance between Zimbabwe and all its trading partners would boost its GDP per capita by 27%, and direct access to the sea would improve Zimbabwe's GDP per capita by 24%. Following Redding and Venables (2004), several studies have confirmed the positive effect of market access on economic development. These papers all focus on regional economic development. Knaap (2006) finds a strong positive effect of market access on income levels when looking at US states, and Breinlich (2006) finds the same for European regions. Also in case of some developing countries, the positive effect of market access has been confirmed. Amiti and Cameron (2007) show that wages are higher in Indonesian districts that enjoy better market access, and Hering and Poncet (2007) find similar evidence in case of Chinese cities. Moreover, Amiti and Javorcik (2008) find that market access positively affects the volume of FDI in Chinese provinces.

[69]This chapter is an adapted version of Bosker and Garretsen (2007a).

[70] Market access may also indirectly affect income levels through its positive effect on education or skill level (see Redding and Schott, 2004 and also Breinlich, 2006). We will come back to this in section 3.6.

[71] Redding and Venables (2004, p.77-78).

The aim of the present chapter is to find evidence on the importance of market access for the economic development of SSA countries. SSA is only a marginal player on the world's export and import markets. Since 1970, the region's share in global trade (exports plus imports) has declined from about 4% to a mere 2% in 2005 (IMF, 2007). Through their detrimental effect on market access, high trade costs are generally viewed as one of the main causes of this poor trade performance (see Collier, 2002; Foroutan and Pritchett, 1993; Coe and Hoffmaister, 1999; Limao and Venables, 2001; Amjadi and Yeats, 1995; Redding and Venables, 2004). As a result, improving the region's market access by investing in infrastructure, increasing regional integration and providing preferential access to European and US markets, is seen as a vital ingredient for improving the trade potential of SSA and its overall economic performance (IMF, 2007; World Bank, 2007; Collier and Venables, 2007; Buys et al., 2006). We are, however, not aware of studies that have looked at the relevance of market access in explaining differences in GDP per capita levels by explicitly focussing on this particular group of countries. The Zimbabwe example from Redding and Venables (2004) referred to above suggests that market access could be very relevant for SSA countries but their results are based on a sample of 101 countries covering both developed and developing countries (of which only 18 SSA countries). This is unfortunate when it comes to the relevance of the market access-income nexus in case of SSA because the role of trade costs and market access could be rather different for the SSA countries compared to developed countries or the fast-growing economies in South-East Asia.

Against this background the main contributions of this chapter are twofold. *First*, by using bilateral manufacturing trade data involving at least one SSA country over the period 1993-2002, we estimate a trade equation to establish the importance of trade costs and market size as determinants of a country's trade potential. Because SSA countries trade far more with the rest of the world (ROW) than with each other (see e.g. IMF, 2007) and have even been found to undertrade with each other (Limao and Venables, 2001)[72], we focus explicitly on the determinants of intra-SSA trade as well as SSA trade with the rest of the world (ROW). Our results show that poor infrastructure across the continent (see also Amjadi and Yeats (1995), Limao and Venables (2001) and Longo and Sekkat, 2004), the civil unrest experienced by many countries in the region, and the fact that many countries with direct access to the sea (island nations in particular) are much more oriented towards the ROW, are part of the explanation for this 'ROW-bias' in SSA trade.

Second, and following the empirical strategy employed by Redding and Venables (2004), we use the trade estimation results to construct various measures of market access (i.e. intra-SSA, ROW, and total market access) and subsequently estimate the impact of market access on GDP per capita for our 48 SSA countries. A nice feature of our data set is that it allows for the use of panel data estimation techniques. We show that this is quite important when trying to establish the relevance of market access, as cross-section studies are likely to overstate the importance of market access. Overall, our main findings are that market access, and notably intra-SSA market access, has a significant positive effect on GDP per capita.

[72] Although the latter is not undisputed, see e.g. Foroutan and Pritchett (1993) and Subramanian and Tamarisa (2003).

Moreover, and in line with Redding and Schott (2003) and Breinlich (2006), we find evidence of an indirect effect of market access on economic development through its positive effect on human capital. Our results finally show that (policy induced) changes in for instance SSA infrastructure (see also Buys et al. 2006) can indeed have strong positive effects on improving market access and thereby on enhancing economic prosperity across SSA.

The chapter is organized as follows. In the next section we briefly set out the new economic geography model underlying our empirical analysis, focusing on the specification of the wage equation and the trade equation. Section 3.3 introduces our data set. In section 3.4, and as the first step of our estimation strategy, a trade equation will be estimated for bilateral trade involving at least one SSA country. In doing so, and following Helpman et al (2007), we will make use of the Heckman two step estimation procedure that takes explicit account of the (large) number of zero trade flows in our data set. Based on the trade estimation results, we construct our market access variables in section 3.5 and present our baseline results with respect to the impact of market access on GDP per capita for SSA. In section 3.6 we first provide various robustness checks, then analyze the relationship between human capital and market access in some more detail for SSA and finally we conduct some 'policy experiments' to establish how various shocks affect market access and, through market access, GDP per capita in SSA. Section 3.7 concludes.

3.2 THE NEG MODEL: INCOME, TRADE COSTS AND MARKET ACCESS[73]

As mentioned in the introduction, our empirical framework is based on a new economic geography (NEG) model very similar to the one presented in chapter 2. Assume the world consists of $i = 1,...,R$ countries, each being home to an agricultural and a manufacturing sector. As in virtually all NEG models, we focus on the manufacturing sector[74]. Moreover, and in line with e.g. Redding and Venables (2004), Breinlich (2006), Knaap (2006) and Head and Mayer (2006), we restrict our attention to the 'short-run' version of the model. This amounts to, as Redding and Venables (2004) p.59 put it, taking the location of expenditure and production as given and asking the question what wages can manufacturing firms in each location afford to pay its workers? (see also Brakman, Garretsen, and Schramm (2006) on this important assumption when taking NEG models to the data). Thus we are, contrary to most theoretical contributions in geographical economics (see Puga, 1999; Krugman, 1991 and also chapter 1), not so much concerned with the 'long-run' equilibrium distribution of economic activity or its properties.

In the manufacturing sector, firms operate under internal increasing returns to scale, represented by a fixed input requirement c_iF and a marginal input requirement c_i. Each firm produces a different variety of the same good under monopolistic competition using the same

[73] The brief discussion and introduction of the basic NEG ingredients needed to arrive at the equilibrium wage equation, the main vehicle to establish the relevance of market access in our empirical research, is largely taken from chapter 2 but see also Fujita, Krugman, and Venables (1999), Puga (1999), Head and Mayer (2004) and Brakman, Garretsen and van Marrewijk (2001) for more detailed expositions as to how the equilibrium wage equation and consequently market access can be derived from the various basic NEG models.

[74] The agricultural sector uses labor and land to produce a freely tradable good under perfect competition that acts as the numéraire good.

Cobb-Douglas technology combining two different inputs. The first is an internationally immobile factor (labor), with price w_i and input share β, the second is an internationally mobile factor with price v_i and input share γ, where $\gamma + \beta = 1$. In Redding and Venables (2004), Knaap (2006) and also chapter 2, each firm uses a composite intermediate input (made up of all manufacturing varieties) as an additional factor of production, allowing them to also look at the relevance of so-called supplier access for income levels. Since our goal is to establish the relevance of market access we, in line with Breinlich (2006), skip intermediate inputs and thereby ignore supplier access[75]. In this respect our derivation and application of the wage equation is closer to Hanson (2005), see also Head and Mayer (2004, pp. 2622-2624).

Manufacturing firms sell their products to all countries and this involves shipping them to foreign markets incurring trade costs in the process. These trade costs are assumed to be of the iceberg-kind and the same for each variety produced. In order to deliver a quantity $x_{ij}(z)$ of variety z produced in country i to country j, $x_{ij}(z)T_{ij}$ has to be shipped from country i. A proportion (T_{ij}-1) of output 'is paid' as trade costs (T_{ij} = 1 if trade is costless). Note that this relatively simple iceberg specification (introduced mainly for ease of modeling purposes, see Fingleton and McCann, 2007) does not specify in any way what trade costs are composed of[76]. Taking these trade costs into account gives the following profit function for each firm in country i,

$$\pi_i = \sum_j^R p_{ij}(z)x_{ij}(z)/T_{ij} - w_i^\beta v_i^\gamma c_i[F + \sum_j^R x_{ij}(z)] \tag{3.1}$$

where $p_{ij}(z)$ is the price of a variety produced in country i.

Turning to the demand side, consumers combine each firm's manufacturing variety in a CES-type utility function, with σ being the elasticity of substitution between each pair of product varieties. Given this CES-assumption, it follows directly that in equilibrium all manufacturing varieties produced in country i are demanded by country j in the same quantity (for this reason varieties are no longer explicitly indexed by (z)). Denoting country j's expenditure on manufacturing goods as E_j, country j's demand for each product variety produced in country i can be shown to be,

$$x_{ij} = p_{ij}^{-\sigma} E_j G_j^{(\sigma-1)} \tag{3.2}$$

where G_j is the price index for manufacturing varieties that follows from the assumed CES-structure of consumer demand for manufacturing varieties. It is defined over the prices, p_{ij}, of all goods produced in country $i = 1,...,R$ and sold in country j,

$$G_j = \left[\sum_i^R n_i p_{ij}^{1-\sigma}\right]^{1/(1-\sigma)} \tag{3.3}$$

[75] Another reason not to include supplier access along with market access is that when estimating the resulting wage equation it is almost impossible to distinguish between these two concepts. Including both market and supplier access is often impossible as doing so creates severe multicollinearity problems. As a result, also Redding and Venables (2004) mainly focus on market access in their estimations, see also Knaap (2006).

[76] When estimating the effect of several trade cost related variables in section 3.4, we will need to specify a trade cost function (see chapter 2).

Maximization of profits (3.1) combined with demand as specified in (3.2) gives the well-known result in the NEG literature that firms in a particular country set the same f.o.b. price, p_i, depending only on the cost of production in location i, i.e. p_i is a constant markup over marginal costs:

$$p_i = w_i^{\beta} v_i^{\gamma} c_i \sigma / (\sigma - 1) \tag{3.4}$$

As a result, price differences between countries of a good produced in country i can only arise from differences in trade costs, i.e. $p_{ij} = p_i T_{ij}$.

Next, free entry and exit drive (maximized) profits to zero, pinpointing equilibrium output per firm at $\bar{x} = (\sigma - 1)F$. Combining equilibrium output with equilibrium price (3.4) and equilibrium demand (3.2), and noting that in equilibrium the price of the internationally (perfectly) mobile primary factor of production will be the same across countries ($v_i = v$ for all i), gives the equilibrium manufacturing wage:

$$w_i = A c_i^{-1/\beta} \left(\underbrace{\sum_j^R \overbrace{E_j G_j^{(\sigma-1)} T_{ij}^{(1-\sigma)}}^{MA_{ij}}}_{MA_i} \right)^{\frac{1}{\beta\sigma}} \tag{3.5}$$

where A is a constant that contains inter alia the substitution elasticity, σ, and the fixed costs of production, F. Equation (3.5) is the wage equation that lies at the heart of virtually all empirical NEG studies (see e.g. Hanson, 2005; Redding and Venables, 2004; Knaap, 2006 and Amiti and Cameron, 2007). It predicts that the wage level a country is able to pay its manufacturing workers is a function of a country's level of technology, c_i, that determines marginal costs, and, most importantly for our present purposes, so-called real market access MA_i: a trade cost weighted sum of all countries' market capacities[77]. Or put differently, country j's contribution to country i's market access, MA_{ij}, is country j's market capacity weighted by the level of trade costs involved when shipping goods from country i to country j, i.e. $MA_{ij} = E_j G_j^{\sigma-1} T_{ij}^{1-\sigma}$.

As already extensively discussed in chapter 2, two estimation strategies have been proposed to estimate the parameters of the wage equation (3.5). The first strategy follows Hanson (2005) and estimates the wage equation directly either using non-linear estimation techniques or by estimating a linearized version of the wage equation, see e.g. Brakman et al (2004) and Mion (2004). Here, we opt for the second strategy as first introduced in the seminal paper by Redding and Venables (2004)[78]. One of the advantages of this strategy is that it is more flexible (see chapter 2) than the direct estimation strategy in allowing for a more elaborate trade cost function (see equation (3.8) in section 3.4). This strategy involves a two-step procedure where in the first step the information contained in (bilateral) trade data is used to provide estimates of the role of trade costs and market and supplier capacity in determining a country's market access. The connection between bilateral trade and market

[77] Note that the price index of a region does not directly enter the wage equation here (as in equation 2.5) due to the fact that no intermediates are used in the production of manufacturing goods.

[78] Other papers using this strategy include inter alia Knaap (2006), Breinlich (2006), Head and Mayer (2006), Hering and Poncet (2006).

access follows directly from the NEG model. Aggregating the demand from consumers in country *j* for a good produced in country *i*, see (3.2), over all firms producing in country *i*, gives the following aggregate trade equation:

$$EX_{ij} = n_i p_i^{1-\sigma} \underbrace{E_j G_j^{(\sigma-1)} T_{ij}^{(1-\sigma)}}_{MA_{ij}} \tag{3.6}$$

Equation (3.6) says that exports EX_{ij} from country *i* to country *j* depend on the 'supply capacity' of the exporting country, $n_i p_i^{1-\sigma}$, that is the product of the number of firms and their price competitiveness, the market capacity of the importing country, $E_j G_j^{\sigma-1}$, that is the product of its income multiplied by its price index (i.e. its real spending power), and the magnitude of bilateral trade costs T_{ij} between the two countries. As real market access is made up of these market capacities, weighted by bilateral trade costs, see (3.5) and (3.6), one can construct a measure of each country's market access using the estimated parameters of (3.6). This market access variable can subsequently be used in the 2nd step of the estimation procedure to estimate the effect of market access on income levels, making use of the wage equation (3.5).

3.3 DATA SET[79]

To make clear which SSA countries are included in our analysis, Table 3.1 provides a list of the SSA countries in our sample, also indicating (*) for which countries we do not have any information on bilateral manufacturing trade flows.

Table 3.1 SSA countries included in the sample

Angola
Benin
Botswana*
Burkina Faso
Burundi
Cameroon
Cape Verde
Central African Republic
Chad
Comoros
Congo
Dem. Rep. of the Congo
Côte d'Ivoire
Djibouti
Equatorial Guinea
Eritrea
Ethiopia
Gabon
Gambia
Ghana
Guinea
Guinea-Bissau
Kenya
Lesotho*
Liberia
Madagascar
Malawi
Mali
Mauritania
Mauritius
Mozambique
Namibia*
Niger
Nigeria
Rwanda
Sao Tome and Principe
Senegal
Seychelles
Sierra Leone
Somalia
South Africa
Sudan
Swaziland*
Tanzania
Togo
Uganda
Zambia
Zimbabwe

Notes : * denotes not in the trade sample.

The data on SSA manufacturing trade that we use in the first step of the estimation procedure, are collected from CEPII's *Trade and Production Database*[80], which contains information on bilateral manufacturing trade flows from 1976-2002. Within this dataset we focus on bilateral trade flows involving at least one SSA country (exporter or importer). Given poor data

[79] See Appendix 3.A, for a full list of all the variables (including data sources) that we use in our analysis.

[80] http://www.cepii.fr/anglaisgraph/bdd/TradeProd.htm. An explanation of the dataset is given at http://www.cepii.fr/tradeprod/TradeProd_cepii.xls.

availability before 1993 (over this period SSA manufacturing import data are only given for 6 SSA countries[81]), we narrow the time-period of the data to encompass the 10 year period 1993-2002. This leaves us with a data set containing information on bilateral manufacturing trade flows for 44 SSA countries both to and from other SSA countries and to and from 148 countries in the rest of the world (ROW). A nice feature of the data set is also that it contains information on some countries' internal trade, i.e. the amount that a country trades with itself (measured as total production minus total exports). After dropping missing observations, we are left with a total sample of 78748 observations (8574 intra-SSA, 70083 SSA-ROW and 91 internal trade observations).

As determinants of SSA trade that are related to *market and supplier capacity*, recall trade equation (3.6), we collected GDP, % rural population, and % workforce in agriculture and also information on the incidence of civil war and/or conflict. As measures of *trade costs,* we use data on bilateral distances, internal distance, language similarity, sharing a common colonizer, having had a colony - colonizer relationship, being landlocked (i.e. not having direct access to the world's oceans), being an island, sharing a common language, an index of infrastructure quality; and membership of an African regional or free trade agreement. In the second stage of our analysis in section 3.5, we complement the above data with data for 48 SSA countries on GDP per capita that we use as our proxy for wages[82] (see also Redding and Venables, 2004), a human capital measure (adult illiteracy), and a measure for economic density (working population per km^2 of arable land).

3.4 STEP 1: TRADE ESTIMATION RESULTS

Our starting point is the trade equation (3.6) as derived from the NEG model in section 3.2. Rewriting (3.6) in loglinear form and allowing for a year-specific intercept[83] gives:

$$\ln EX_{ijt} = \alpha_0 + \alpha_t + \alpha_1 \ln Y_{it} + \alpha_2 \ln Y_{jt} + \alpha_3 \ln T_{ijt} + \varepsilon_{ijt} \qquad (3.7)$$

where Y_i denotes country i's GDP, and a subscript t is added to denote the year of observation. As has been extensively discussed in chapter 2, the NEG-model does not specify trade costs T_{ijt} in any way (except that they are of the iceberg type), which complicates matters when moving to the empirics. Actual trade cost data are generally missing, especially when also considering the more intangible trade costs such as cultural or linguistic differences. In the absence these actual trade cost data, we follow the modern empirical trade and economic geography literature (see e.g. Anderson and van Wincoop, 2004; Limao and Venables, 2001; Redding and Venables, 2004), and specify trade costs, T_{ijt}, to be a multiplicative[84] function of

[81] South Africa, Kenya, Ethiopia, the Comoros, Malawi and Madagascar.

[82] Using GDP per worker instead invariably gives the same results, but we loose an additional 8 observations in the process. This and the fact that Redding and Venables (2004) also use GDP per capita, made us decide to show the results using GDP per capita. Results using GDP per worker are available upon request.

[83] These time dummies are included to capture worldwide developments affecting the ease of exporting and importing (think of for example increased efficiency and/or technological advancements in transportation).

[84] This is the usual choice in the gravity literature (see e.g. Limao and Venables, 2001; Subramanian and Tamarisa, 2003). See Hummels (2001) for a critique on this, arguing in favor of an additive specification instead. Also we choose to allow for a concave distance function, which is standard in the transport economics literature

the following observable variables that are commonly used in the literature (see Appendix 3.A for more details on each of the variables): bilateral distance (*D)*, sharing a common border (*B*), common language (*CL*), common colonial heritage (distinguishing between sharing a common colonizer (*CC*) and having had a colony-colonizer relationship (*CR*)), being landlocked (*ll*), being an island (*isl*), an index measuring the quality of infrastructure (*inf*) and membership of the same African regional or free trade agreement (*RFTA*). In loglinear form this amounts to the following trade cost specification:

$$\begin{aligned}\ln T_{ijt} &= \chi_1 \ln D_{ij} + \chi_2 \ln B_{ij} + \chi_3 \ln CL_{ij} + \chi_4 \ln CC_{ij} + \chi_5 \ln CR_{ij} \\ &\quad + \chi_6 ll_i + \chi_7 ll_j + \chi_8 isl_i + \chi_9 isl_j + \chi_{10} inf_{it} + \chi_{11} inf_{jt} + \chi_{12} RFTA_{ijt}\end{aligned} \qquad (3.8)$$

Besides including GDP as the 'standard' trade determinants related to countries' economic size, we also included the following additional variables in (3.7) that we think take account of some trade determinants that are to some extent typical of SSA. Given the fact that SSA has been the most conflict-ridden continent over the last few decades, we include two dummy variables that indicate whether a country experienced outbreaks of civil unrest in a specific year: civil conflict (*cconfl)* or civil war *(cwar)*; where civil war indicates more intense fighting than civil conflict. Next, we included the share of people living in rural areas (*%rural).* Manufacturing activity is usually located in or near urban centers; higher urbanization increases a country's capacity to im- and export these goods. Also, Ancharaz (2003) shows that higher urbanization shares increase the likelihood of trade policy reform, and moreover, see Sahn and Stifel (2003), the welfare level of the urban population is generally higher than that of the rural population in SSA, resulting in higher demand for imported manufacturing products[85].

Some related studies do not have to choose the specific variables that capture a country's market capacity by opting instead for the inclusion of *importer and exporter fixed effects* when estimating their version of (3.7), see e.g. Breinlich (2006) and Knaap (2006). We, however, decided to explicitly specify country-specific determinants of trade for the following three reasons. First, as explained in chapter 2 of this thesis, these importer and exporter fixed effects also capture all trade cost related variables that are country-specific. As a result, the constructed market access term only includes the importer fixed effects, hereby ignoring the exporter-specific trade costs (see footnote 50). Second, as pointed out by Redding and Venables (2004, p.75), using importer and exporter fixed effects does not allow one "to quantify the effects on per capita income of particular country characteristics (for example, land locked or infrastructure), since all such effects are contained in the dummies" (Redding and Venables, 2004, p. 75). As a result, recommendations regarding the effect of country-specific policies (see section 3.6.3) aimed at e.g. lowering trade costs, are impossible

(see McCann, 2001 and Fingleton and McCann, 2007). See chapter 2 for a much more extensive discussion on the choice of trade cost function.

[85] Other authors have included GDP per capita as a measure of welfare to the gravity equation. We choose to use % urban population instead. % urban population is highly correlated with GDP per capita, and results are very similar when using GDP per capita instead.

to make[86]. Third, as bilateral trade data are missing for four SSA countries (see Table 3.1), we are not able to estimate the importer and exporter effect for these four countries. As a result these countries would also not be considered when constructing the various market access variables. When using country-specific characteristics instead, this can be avoided: we can use the estimated parameters from the first step in combination with these four countries' characteristics to construct the market access contributions for each of these countries even in the absence of these countries' bilateral trade data (of course given that we do have data on these country characteristics).

Finally, we also include a dummy for intra-SSA and internal trade (α_{ssa}, α_{own}), so that the trade specification that we estimate in the 1st step of our analysis is:

$$\begin{aligned}\ln EX_{ijt} = \alpha_0 + \alpha_{SSA} + \alpha_{own} + \alpha_t + \alpha_1 \ln Y_{it} + \alpha_2 \ln Y_{jt} + \ln T_{ijt} + \alpha_4 \% rural_{it} \\ + \alpha_5 \% rural_{jt} + \alpha_6 cconfl_{it} + \alpha_7 cconfl_{jt} + \alpha_8 cwar_{it} + \alpha_9 cwar_{jt} + \varepsilon_{ijt}\end{aligned} \tag{3.9}$$

with $\ln T_{ijt}$ as in (3.8). The actual estimation of (3.9) is, however, not without problems. In particular, the presence of zero trade flows complicates matters. As shown in Bosker (2007b), about 50% of the observed SSA manufacturing trade flows are zeroes. As taking the log of zero is impossible, one has to choose between several different estimation strategies that all deal with these zero observations in different ways. Referring to Bosker (2007b) for a detailed exposition of which estimation strategy to use, we use a Heckman 2-step estimation strategy (see also Helpman et al., 2007) to estimate the parameters of (3.9). This has the advantage of neither having to discard the zero observations (as using OLS on the non-zeroes would imply) nor having to a priori assume that the exact same model explains both the probability to trade and the volume of trade as using Tobit or estimating (3.9) in its non-linear form using NLS or Poisson techniques, see (3.6), would imply[87].

The Heckman 2-step procedure amounts to first estimating, by using probit, how each of the variables affects the probability to trade. Next, in the second stage, the effect of each variable on the volume of trade is estimated, including the inverse Mills ratio (that is constructed using the results from the first step) to control for endogenous selection bias that would plague the results when simply discarding the non-zero observations (see for instance ch.17 in Wooldridge, 2003). The results of this 2-step procedure are only convincing when one uses an exclusion restriction, i.e. having at least one variable that determines the probability to trade but *not* the volume of trade (see p.589 Wooldridge, 2003 and Bosker (2007b)). To this end, we decided to use the percentage of the labor force employed in agriculture. We assume that this variable (after correcting for the other included variables) does not affect the volume of trade but only the probability to trade. The reasoning for using this variable and again referring to Bosker (2007b) for a much more detailed exposition of both the econometric validity and economic validity of our choice, is that it has been shown

[86] This is also why we decided not to approximate trade costs by calculating implied trade costs, as discussed in chapter 2.

[87] Especially when the number of zero observations is substantial (here 50%), Poisson and NLS techniques face the 'excess zero problem', i.e. they substantially underpredict the number of true zeroes (see Appendix C of Bosker, 2007b).

by inter alia Temple and Wößmann (2006) and Poirson (2001) that a lower share of the labor force employed in agriculture increases aggregate total factor productivity. When looked upon as a proxy of the economy's aggregate productivity, the fact that the percentage of the labor force in agriculture only affects the probability that a SSA economy exports or imports manufacturing goods, can be viewed as being consistent with trade theories of the Melitz (2003) type emphasizing productivity as a major determinant of the probability to trade (see also e.g. Hallak, 2006 and Helpman et al., 2007 for a similar line of reasoning).

Table 3.2 shows the estimation results, where the *marginal effects* give the overall effects of each of the included variables on the volume of trade (after taking the 1st stage into account) and the results for *0/1 trade* refer to the 1st stage probit estimations. To explicitly allow for a different effect of a particular variable on intra-SSA and SSA trade with the rest of the world, we have interacted several variables with a dummy-variable taking the value of 1 when considering intra-SSA trade. In Table 3.2, the addition "*ssa*" to a certain variable denotes that variable interacted with an intra-SSA dummy; significance of an *"ssa"*-variable indicates a different effect of that particular variable on intra-SSA trade compared to SSA trade with the ROW.

Table 3.2 The Trade Equation Estimates

dependent variable			ln trade		
Estimation method			Heckman - 2step		
time period			1993-2002		
Variable	marginal effects	0/1 - trade	variable	marginal effects	0/1 - trade
ln distance	-1.478	-0.395	island exp	0.555	0.166
ln internal distance	0.869*	0.430	island exp ssa	-2.177	-0.690
ln distance ssa	-0.033**	-0.140	island imp	0.393	0.228
ln gdp imp	1.385	0.464	island imp ssa	-1.504	-0.748
ln gdp imp ssa	-0.427	-0.060	ln infrastructure exp	0.459	0.126
ln gdp exp	1.531	0.436	ln infrastructure exp ssa	1.053	0.277
ln gdp exp ssa	-0.250	-0.031	ln infrastructure imp	0.168	-0.013**
colony – colonizer	2.470	1.702	ln infrastructure imp ssa	0.833	0.223
common colonizer	1.094	0.309	RTA or FTA	-0.471**	0.013**
common colonizer ssa	-0.079**	-0.011**	RTA or FTA ssa	1.489	0.085**
Contiguity	-1.241	-0.272**	civil conflict imp	-0.538	-0.196
Contiguity ssa	2.645	0.821	civil conflict exp	-0.599	-0.062
common off language	0.858	0.289	civil war imp	-0.905	-0.322
common off language ssa	-0.553	-0.164	civil war exp	-1.447	-0.368
landlocked exp	-0.226	-0.229	% rural population imp	-0.343	-0.030**
landlocked exp ssa	-0.765	-0.066**	% rural population exp	-2.985	-0.651
landlocked imp	-0.294	0.039*	% in agriculture imp	-	-0.159
landlocked imp ssa	0.334	-0.006**	% in agriculture exp	-	-0.119
dummy ssa trade	13.021	3.371	dummy internal trade	2.413**	4.982
nr observations			74492		
Mills ratio [p-value]			2.930 [0.000]		

Notes : ** (*) denotes **not** significant at the 5% (1%) respectively.

Also, as argued in chapter 2, we allow distance to have a different effect when considering internal trade, hereby explicitly estimating the possibly different effect of distance on internal trade[88] instead of simply postulating a difference (as in Redding and Venables, 2004) or assuming no difference (as in e.g. Breinlich (2006), and Knaap, 2006).

The main estimation results reported in Table 3.2 are as follows (see Bosker, 2007b for a more elaborate exposition of the trade results). We first look at the outcomes for the non-trade cost related variables. Importer and exporter GDP both have the expected positive sign; interestingly, the trade-stimulating effect of an increase in GDP is much lower when considering intra-SSA trade, suggesting that as SSA countries get richer the focus of their manufacturing trade activity shifts away from other SSA countries in favor of countries in the ROW. Civil unrest negatively affects trade, and the more so the more violent civil unrest becomes (compare the parameters of the civil war dummies to those of the civil conflict dummies). Also as expected, a higher degree of urbanization results in more exports and imports of manufacturing goods. The effect on manufacturing exports is however much larger than on imports. Given that manufacturing in SSA is mostly located in urban areas and (unskilled) labor-intensive, an explanation for this last finding could be that a high level of urbanization suppresses wages due to increased supply of unskilled labor, that lowers firms' production costs, making it easier for them to be competitive on world markets.

Next, we turn to the estimation results regarding the effect of the different trade cost variables on SSA manufacturing trade. First we discuss the results regarding the bilateral trade cost variables. Distance negatively affects the volume of trade between countries. In line with Foroutan and Pritchett (1993), but contrary to Limao and Venables (2001), we do not find evidence that the penalty on distance is higher for intra-SSA trade. Also, the results clearly show the advantage of explicitly allowing for a different effect of distance on internal trade: the distance penalty is about 60% lower for internal trade compared to bilateral trade. For intra-SSA trade we find a clear positive effect of sharing a common border on trade flows (see e.g. Limao and Venables, 2001; Subramanian and Tamirisa, 2003 and Foroutan and Pritchett, 1993). For SSA-ROW trade we, however, find (surprisingly) a negative effect. When we take into consideration that the only SSA countries that border non-SSA countries are those bordering North African countries, this simply indicates that these SSA countries trade less with their North African neighbors than with non-African countries (see also IMF, 2007). Sharing a colonial history has a strong positive effect on the volume of trade in manufactures. Especially SSA trade with its former colonizer(s) is much higher than trade with other countries in the world. Having a common colonizer also boosts bilateral trade and we find no indication that the effect is different for intra-SSA trade compared to trade with the ROW. Sharing a common language stimulates both intra-SSA and SSA-ROW trade (see also Foroutan and Pritchett (1993) and Coe and Hoffmaister (1999)). The trade facilitating effect of language similarity is much larger for trade with the rest of the world however (the common border and common colonizer variable may already be capturing some of the language effect in case of intra-SSA trade). A bilateral trade cost variable that is of particular interest is the variable capturing the effect of being a member of the same African regional or free trade agreement (RFTA). Intra-SSA trade in manufactures substantially benefits from

[88] Similar to the *ssa*-variables, the coefficient on internal distance denotes the *difference* in the impact of internal distance on internal trade compared to the effect of bilateral distance on bilateral trade.

having an RFTA, providing evidence in favor of those who argue for increased African integration (one of the explicit goals of e.g. the African Union). The finding that having an RFTA does not significantly affect SSA-ROW trade is not surprising since the only non-SSA countries being part of an all-African RFTA are some of the North African countries (see the results regarding the common border variable).

Next, we discuss the results for the three *country-specific* trade cost variables, i.e. being landlocked, being an island, and the quality of a country's infrastructure. We find that being landlocked depresses both SSA imports and exports of manufacturing goods to the ROW, corroborating the findings in Coe and Hoffmaister (1999). When looking at intra-SSA trade, being landlocked affects intra-SSA exports even more negatively; on the contrary, being landlocked slightly increases the amount imported from other SSA countries. This difference is quite interesting, as it indicates that landlocked countries in SSA are more dependent on imported manufacturing goods from other SSA countries compared to the SSA countries that do have direct access to the sea. Being an island nation increases trade with the ROW, confirming findings by e.g. Limao and Venables (2001). However intra-SSA is much lower for these same island nations. Apparently, the island nations of SSA (Mauritius, Comoros, Cape Verde and Sao Tome and Principe) are oriented away from the African mainland when it comes to trade. The findings on these two 'location' variables suggest that SSA countries are much more oriented towards the ROW than towards other SSA countries: island nations trade much less with the African mainland and countries with direct access to the sea import more manufacturing goods from the ROW than from other SSA countries (see also the results on the GDP variables).

The final trade cost related variable that we consider is the quality of a country's infrastructure, which arguably is the most interesting variable from a policy perspective given the large amounts of funds currently allocated by donors to (co-)finance infrastructure improvements in SSA ($7.7 billion by members of The Infrastructure Consortium for Africa alone[89]). In line with the results in Limao and Venables (2001), we find that improved quality of infrastructure has large positive effects on the volume of trade. Even more interestingly, improving the quality of infrastructure has a much larger positive effect on intra-SSA trade than on SSA trade with the ROW. These findings show that the current focus on improving SSA infrastructure (see e.g. the aim of the Sub-Saharan African Transport Policy Program[90] and The Infrastructure Consortium for Africa[89]) is warranted.

3.5 STEP 2: CONSTRUCTING MARKET ACCESS AND BASELINE RESULTS

3.5.1 Constructing market access from the trade estimations

Having estimated the effects of both the trade-cost related and the market capacity related variables on the volume of trade, we are now in a position to construct market access for our sample of 48 SSA countries. More specifically, we use the estimated coefficients shown in

[89] See http://www.icafrica.org/fileadmin/documents/AR2006/ICA_Annual_Report_-_Volume_1_-_FINAL_March_2007.pdf.

[90] For more info see: http://web.worldbank.org/WBSITE/EXTERNAL/COUNTRIES/AFRICAEXT/EXTAFRREGTOPTRA/EXTAFRSUBSAHTRA/0,,menuPK:1513942~pagePK:64168427~piPK:64168435~theSitePK:1513930,00.html?

Table 3.2, and the relationship between the trade equation in (3.6) and market access in (3.5), to construct market access. In doing so, we distinguish explicitly between the contribution of internal market access, SSA market access and ROW market access to a country's total market access:

$$MA_{it} = MA_{it}^{own} + MA_{it}^{SSA} + MA_{it}^{ROW} \tag{3.10}$$

where $MA_{it}^{SSA} = \sum_{j \in SSA, j \neq i}^{R} MA_{ijt}^{SSA}$, $MA_{it}^{ROW} = \sum_{j \notin SSA}^{R} MA_{ijt}^{ROW}$ and $MA_{it}^{own} = MA_{iit}$

and

$$\begin{aligned}
MA_{iit}^{own} &= e^{\hat{\alpha}_{own}} \left(Y_{jt}\right)^{\hat{\alpha}_2} \left(\%rural_{jt}\right)^{\hat{\alpha}_5} \left(D_{ii}\right)^{\hat{\chi}_1 + \hat{\chi}_1^{own}} \\
&\quad e^{(\hat{\chi}_6+\hat{\chi}_7)ll_i + (\hat{\chi}_8+\hat{\chi}_9)isl_i + (\hat{\chi}_{10}+\hat{\chi}_{11})inf_{it} + (\hat{\alpha}_6+\hat{\alpha}_7)cconfl_{it} + (\hat{\alpha}_8+\hat{\alpha}_9)cwar_{it}} \\
MA_{ijt}^{row} &= \left(Y_{jt}\right)^{\hat{\alpha}_2} \left(\%rural_{jt}\right)^{\hat{\alpha}_5} \left(D_{ij}\right)^{\hat{\chi}_1} e^{\hat{\chi}_2 B_{ij} + \hat{\chi}_3 CL_{ij} + \hat{\chi}_4 CC_{ij} + \hat{\chi}_5 CR_{ij} + \hat{\chi}_{12} RFTA_{ijt}} \\
&\quad e^{\hat{\chi}_6 ll_i + \hat{\chi}_7 ll_j + \hat{\chi}_8 isl_i + \hat{\chi}_9 isl_j + \hat{\chi}_{10} inf_{it} + \hat{\chi}_{11} inf_{jt} + \hat{\alpha}_6 cconfl_{it} + \hat{\alpha}_7 cconfl_{jt} + \hat{\alpha}_8 cwar_{it} + \hat{\alpha}_9 cwar_{jt}} \\
MA_{ijt}^{ssa} &= e^{\hat{\alpha}_{ssa}} \left(Y_{jt}\right)^{\hat{\alpha}_2{}^{ssa}} \left(\%rural_{jt}\right)^{\hat{\alpha}_5} \left(D_{ij}\right)^{\hat{\chi}_1{}^{ssa}} e^{\hat{\chi}_2{}^{ssa} B_{ij} + \hat{\chi}_3{}^{ssa} CL_{ij} + \hat{\chi}_4{}^{ssa} CC_{ij} + \hat{\chi}_{12}{}^{ssa} RFTA_{ijt}} \\
&\quad e^{\hat{\chi}_6{}^{ssa} ll_i + \hat{\chi}_7{}^{ssa} ll_j + \hat{\chi}_8{}^{ssa} isl_i + \hat{\chi}_9{}^{ssa} isl_j + \hat{\chi}_{10}{}^{ssa} inf_{it} + \hat{\chi}_{11}{}^{ssa} inf_{jt} + \hat{\alpha}_6 cconfl_{it} + \hat{\alpha}_7 cconfl_{jt} + \hat{\alpha}_8 cwar_{it} + \hat{\alpha}_9 cwar_{jt}}
\end{aligned} \tag{3.11}$$

where $\hat{\alpha}_k^{ssa}$ and $\hat{\chi}_k^{ssa}$ capture the estimated effect of a variable on intra-SSA trade (i.e. the coefficient on a variable plus the coefficient on that variable interacted with the intra-SSA dummy), $\hat{\chi}_1^{own}$ captures the possibly different effect of distance on internal trade, and D_{ii} denotes internal distance as defined in (2.15).

Using (3.10) and (3.11), we construct total market access (MA), SSA market access (SSA-MA), ROW market access (ROW-MA) and own market access (own-MA) for each of the 48 SSA countries and for each year in our sample period 1993-2002. Table 3.3 shows average (log) total market access along with the share of each of its subcomponents, and the average SSA GDP per capita for each of the years in our sample.

Table 3.3 Market Access (shares) and GDP per capita over time

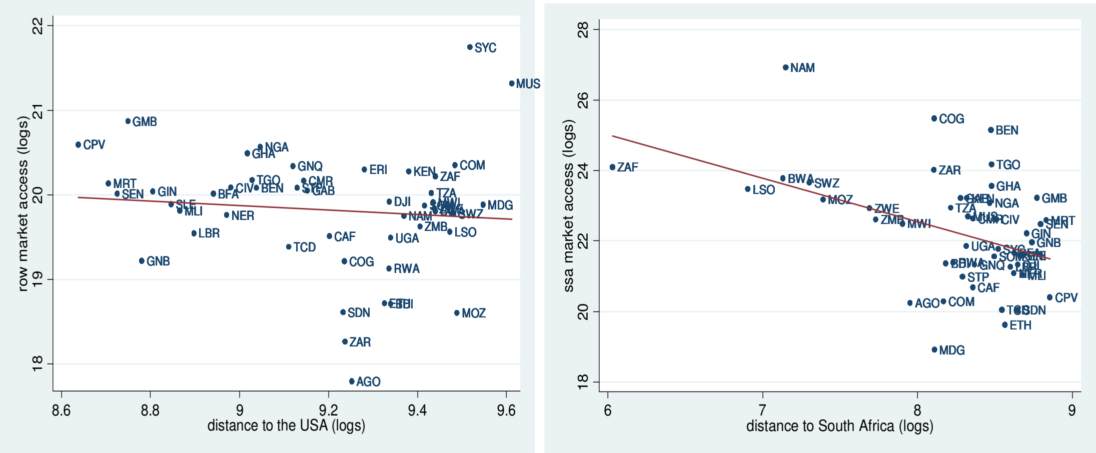
year	ln MA	% row	% ssa	% own	GDP per capita
1993	24.23	8.9	69.7	21.4	2342
1994	24.24	9.0	70.2	20.7	2334
1995	24.33	9.3	69.9	20.9	2353
1996	24.12	9.3	69.0	21.7	2394
1997	24.03	10.2	67.6	22.2	2462
1998	24.04	8.9	68.4	22.7	2534
1999	24.11	9.6	67.6	22.8	2567
2000	24.12	12.1	62.5	25.3	2633
2001	24.49	11.3	63.4	25.3	2559
2002	24.48	11.4	63.3	25.4	2661
% change 1993-2002	28.61	2.5*	-6.5*	4.0*	13.64
average yearly % change	1.40	0.3*	-0.7*	0.4*	1.30

Notes: * denotes percentage *point* changes.

Average market access has improved at an annual rate of 1.4%, slightly higher than the average annual growth rate of GDP per capita. Looking at the three subcomponents of market access shows that SSA market access dominates total market access, and this reflects in part the high penalty on distance in SSA trade and the positive border effect (see Table 3.2). The share of SSA market access has, however, decreased from around 70% in 1993 to about 63% in 2002. ROW and own market access have both gained in importance in total market access. The fact that the ROW's share in market access has risen partly reflects that the ROW experienced (on average) higher GDP growth than SSA. Combined with the higher coefficient on ROW GDP (see Table 3.2), this has increased ROW market access faster than SSA market access. The increase in own market access can also be ascribed to the higher coefficient on GDP in own market access compared to that in SSA market access, but also the much smaller penalty on internal distance (see Table 3.2) plays a role here.

Figure 3.1 illustrates in some more detail the role of distance to major markets as a determinant of market access. The left panel plots for each SSA country its ROW-market access against the distance to the United States, and the left panel plots for each country its SSA-market access against the distance to South Africa.

Figure 3.1 Market Access and distance to major markets



Notes: the raw correlation between log distance to the USA and log ROW market access is (p-values in brackets): -0.09 (0.05) and that between log distance to South Africa and log SSA market access is –0.42 (0.00).

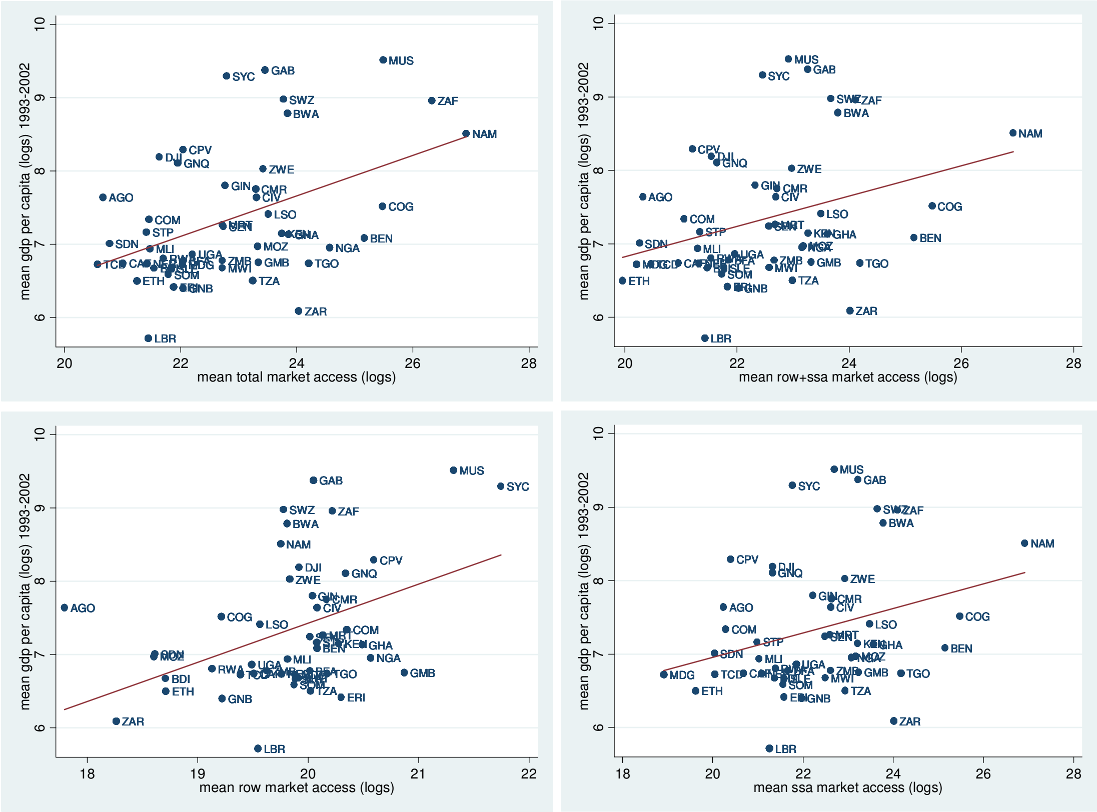
It shows that in both cases market access is lower for those countries at greater distance from the United States and South Africa respectively. This effect is, however, much more pronounced when considering SSA market access and the distance to South Africa, which is in part due to the fact that besides the USA also Europe (and increasingly also Asia) constitutes a large market for SSA products. Figure 3.1 also shows that some of SSA countries with the worst access to markets in the ROW are landlocked SSA countries (e.g. Chad, the Central African Republic, Rwanda), whereas the island nations (e.g. Seychelles, Mauritius, Comoros and Cape Verde), tend to have the best access to non-SSA markets. When considering SSA market access, these same island nations are doing much worse. Also landlocked countries are again among the countries with the worst SSA market access.

Besides distance to South Africa also countries close to Africa's second largest economy, Nigeria, tend to have higher SSA market access. Finally we note that those countries experiencing civil conflict or, worse still, civil war, e.g. Sudan, the Democratic Republic of Congo, Ethiopia, or Angola, tend to be among the countries with the lowest SSA as well as ROW market access.

3.5.2 The relevance of market access for GDP per capita in SSA: baseline results

Now that we have constructed the various measures of market access for all 48 SSA countries in our sample, we can assess the effect of market access on GDP per capita. Before we turn to the actual estimation of the wage equation (3.5), Figure 3.2 plots *mean* market access (TOTAL, ROW+SSA, ROW, and SSA market access) for the period 1993-2002 against *mean* gdp per capita over that same period.

Figure 3.2 Market Access and GDP per capita in SSA



Notes: the raw correlations of each of the market access variants are (p-values in brackets): total: 0.44 (0.00); row + ssa: 0.31 (0.00); row: 0.38 (0.00); ssa: 0.28 (0.00).

Figure 3.2 shows a clear positive relationship between gdp per capita and market access. Also, SSA-market access seems somewhat less important compared to ROW market access (see the reported correlations below Figure 3.1).

By taking logs on both sides of the wage equation (3.5) from the NEG model in section 3.2, we arrive at the log-linear relationship between market access and wages that we will estimate using panel data techniques:

$$\ln w_{it} = \beta_0 + \beta_1 \ln MA_{it} + \eta_{it} \tag{3.12}$$

In line with Redding and Venables (2004, p.63) and Breinlich (2006), we proxy wages (the price of the immobile factor of production) by GDP per capita. The error term η_{it} includes c_i, a country's level of technological efficiency. Again following Redding and Venables (2004), we start by assuming that these cross-country differences in technology are captured by an idiosyncratic error term (implicitly also allowing for other variables determining technological efficiency that are *un*correlated with our market access measure), and estimate (3.12) using pooled OLS. The results are shown in the first four columns of Table 3.4[91].

Table 3.4 Market access and gdp per capita – first estimation results

dep: log gdp per capita	ols	ols	ols	ols	ols	ols	ols	ols
log tot ma	0.258	-	-	-	0.063	-	-	-
robust	0.000	-	-	-	0.005	-	-	-
bootstrapped	0.000	-	-	-	0.010	-	-	-
log ssa+row ma	-	0.186	-	-	-	0.031	-	-
robust	-	0.000	-	-	-	0.050	-	-
bootstrapped	-	0.000	-	-	-	0.073	-	-
log row ma	-	-	0.461	-	-	-	0.032	-
robust	-	-	0.000	-	-	-	0.321	-
bootstrapped	-	-	0.000	-	-	-	0.373	-
log ssa ma	-	-	-	0.153	-	-	-	0.031
robust	-	-	-	0.000	-	-	-	0.036
bootstrapped	-	-	-	0.000	-	-	-	0.049
p-value country FE	no	no	no	no	0.000	0.000	0.000	0.000
p-value time FE	no	no	no	no	0.000	0.000	0.001	0.000
nr observations	477	477	477	477	477	477	477	477
R2	0.190	0.094	0.147	0.077	0.966	0.965	0.951	0.965

Notes : p-values below coefficients. Bootstrapped p-values on the basis of 200 replications. Results for the constant and the time- and country fixed effects are not shown to save space.

The estimated market access coefficient is positive and significant for each of the four measures of market access, indicating a positive effect of market access on GDP per capita across SSA. An increase of total market access by 1% would increase gdp per capita by 0.25%. For only foreign market access (ROW+SSA market access excludes own market access), we find a positive (but somewhat smaller) effect of market access. When considering the effect of only SSA or only ROW market access, we find an interesting difference between the two. The estimated coefficient on ROW market access is much higher than that on SSA market access, and also ROW market access on its own explains about twice as much of the SSA-variance in gdp per capita than SSA market access does. This suggests that it is above all improved market access to non-SSA countries that will boost economic development in SSA,

[91] Following Breinlich (2006), Knaap (2006) and Redding and Venables (2004), we show both robust and boostrapped standard errors for all our estimation results. The bootstrapped standard errors take explicit account of the fact that our measures of market access are all generated regressors, see Redding and Venables (2004, p. 64) for more details.

thereby seemingly vindicating those studies that proclaim that intra-SSA economic linkages are too weak and underdeveloped to be of importance to SSA countries.

The estimation results in the first 4 columns of Table 3.4 are, however, only valid under the earlier-mentioned assumption of idiosyncratic differences in countries' technological efficiency that are uncorrelated with market access. As this assumption is likely to be violated, we subsequently make use of the panel data nature of our data set. We include country fixed effects to capture country-specific variables affecting a country's technological efficiency that do not vary over time, most notably physical geography (climate, soil quality, etc) and institutional quality[92] (see also Breinlich, 2006). And by including also time (year) fixed effects, we take account of shocks that are affecting all countries similarly, such as the availability of technological innovations made in developed countries (the introduction of mobile phones, which have rapidly spread all over SSA, is a prime example). The last 4 columns of Table 3.4 show the results of these fixed effect estimations.

As can be seen from Table 3.4, the inclusion of fixed effects is quite important: the effect of total market access on gdp per capita is still positive and significant but the size of the market access coefficient is much lower: a 1% increase in a country's total market access, now 'only' increases gdp per capita with 0.06%. In addition, when we split total market access in ROW+SSA-MA, ROW-MA and SSA-MA, we now observe that the coefficient on SSA market access is *not* different from that on ROW market access. Even more strikingly, when considering ROW market access only, it no longer has a significant impact on GDP per capita, whereas SSA market access still does have a significant effect. The significance of ROW+SSA-MA and also that of total MA seems to be largely due to market access to other SSA countries. Also note that both the country and time fixed effects are both significant and including both of them substantially increases the explained percentage of the variance in SSA's GDP per capita.

The inclusion of these two fixed effects may still not provide us with accurate estimates of the effect of market access however, as they only control for time-invariant country-specific or country-invariant time-specific variables. It is possible that a country's technological efficiency is also determined by time- ànd country-varying variables that are correlated with market access. If this were the case, we would still obtain biased estimates of the coefficient on market access, even when allowing for country- and year fixed effects. Following Breinlich (2006), we therefore also include two variables, namely the adult illiteracy rate as a measure for a country's human capital[93], and the working population density per km^2 of arable land[94], that both have been shown to affect a country's productivity level and to be correlated with market access (see e.g. Ciccone and Hall, 1996; Ciccone ,2002;

[92] Of course, institutional quality may change over time, but given our relatively short time span of 10 years we are quite confident that we are capturing institutional quality by allowing for country fixed effects.

[93] In section 3.6.2, we focus in more detail on the relationship between human capital, market access and income levels.

[94] We use arable land, instead of total land because large parts of almost each SSA country are quite hostile to human settlement (e.g. the Sahara and Kalahari desert and the dense jungles in central Africa).

Redding and Schott, 2003 and Breinlich, 2006) [95]. Table 3.5 shows the corresponding estimation results when we add these two control variables (with and without fixed effects).

Table 3.5 Adding Human Capital and Employment Density: Baseline Results

					BASELINE			
dep: log gdp per capita	ols	ols	ols	ols	ols	ols	ols	ols
log tot ma	0.160	-	-	-	0.076	-	-	-
robust	0.000	-	-	-	0.001	-	-	-
bootstrapped	0.000	-	-	-	0.004	-	-	-
log ssa+row ma	-	0.088	-	-	-	0.053	-	-
robust	-	0.000	-	-	-	0.004	-	-
bootstrapped	-	0.000	-	-	-	0.007	-	-
log row ma	-	-	0.352	-	-	-	0.083	-
robust	-	-	0.000	-	-	-	0.038	-
bootstrapped	-	-	0.000	-	-	-	0.043	-
log ssa ma	-	-	-	0.072	-	-	-	0.050
robust	-	-	-	0.000	-	-	-	0.003
bootstrapped	-	-	-	0.000	-	-	-	0.005
adult illiteracy	-0.018	-0.022	-0.022	-0.022	-0.021	-0.019	-0.017	-0.018
robust	0.000	0.000	0.000	0.000	0.048	0.074	0.111	0.084
bootstrapped	0.000	0.000	0.000	0.000	0.073	0.105	0.167	0.133
log working pop / km2 arable land	0.108	0.091	0.047	0.089	0.294	0.300	0.320	0.304
robust	0.004	0.011	0.206	0.013	0.050	0.044	0.032	0.042
bootstrapped	0.004	0.000	0.331	0.029	0.063	0.079	0.058	0.046
p-value country FE	no	no	no	no	0.000	0.000	0.000	0.000
p-value time FE	no	no	no	no	0.320	0.395	0.419	0.385
nr observations	369	369	369	369	369	369	369	369
R2	0.401	0.356	0.412	0.352	0.966	0.965	0.965	0.966

Notes : p-values below coefficients. Bootstrapped p-values on the basis of 200 replications.

With regard to the different impact of each of the components of total market access, we find that all three components are significant and positively contributing to gdp per capita. ROW market access still has a larger impact on gdp per capita, although the difference with SSA market access is much smaller than in the first four columns of Table 3.4 or Table 3.5. Given the fact that SSA-MA's contribution to total MA is much larger than that of ROW-MA (see Table 3.3), so that the variation in SSA-MA effectively swamps the variation in ROW-MA, the coefficient on ROW+SSA market access is about the same as that on SSA market access. Overall, a 1% increase in total market access increases gdp per capita by 0.08%, and when focussing only on SSA, ROW or foreign (SSA+ROW) market access the effect of a 1% increase in the corresponding market access term increases gdp per capita by 0.05% in case of SSA and SSA+ROW market access and by 0.08% in case of ROW market access[96].

[95] Also controlling for natural resource dependence (by including a dummy for oil exporting countries and/or the percentage of ores and metals in merchandise trade) leaves the results on market access unaffected. Results are available upon request.

[96] Note that all results that we present are robust to the exclusion of South Africa from the sample.

3.6 ADDITIONAL RESULTS: ROBUSTNESS CHECKS, HUMAN CAPITAL AND POLICY SHOCKS

3.6.1 Robustness of the results

The last four columns of Table 3.5 constitute our baseline results. One could still raise several issues that would invalidate these results. First, even though we have corrected for fixed time and country effects and have added two additional control variables, there is the issue of endogeneity. The assumption under which our baseline results are valid is that, after controlling for fixed effects and our two included controls, the remaining error term is uncorrelated with our measures of market access. One way in which this may be violated is when the error term still contains other variables influencing a country's GDP per capita that are correlated with market access. Another way is reverse causality; when market access not only influences GDP per capita but GDP per capita in turn also influences market access, the error term would by construction be correlated with market access and thus give biased estimates of the effect of our measures of market access.

Table 3.6 Robustness of the results – IV, lagged MA, and 1st differences (FD)

					ols	ols	ols	ols				
dep: log gdp per capita	IV	IV	IV	IV	lagged	lagged	lagged	lagged	FD	FD	FD	FD
log tot ma	0.427	-	-	-	0.077	-	-	-	0.033	-	-	-
robust	0.000	-	-	-	0.003	-	-	-	0.011	-	-	-
bootstrapped	-	-	-	-	0.005	-	-	-	0.013	-	-	-
log ssa+row ma	-	0.392	-	-	-	0.053	-	-	-	0.027	-	-
robust	-	0.000	-	-	-	0.013	-	-	-	0.024	-	-
bootstrapped	-	-	-	-	-	0.010	-	-	-	0.024	-	-
log row ma	-	-	0.232	-	-	-	0.102	-	-	-	0.035	-
robust	-	-	0.059	-	-	-	0.013	-	-	-	0.214	-
bootstrapped	-	-	-	-	-	-	0.022	-	-	-	0.241	-
log ssa ma	-	-	-	0.376	-	-	-	0.048	-	-	-	0.025
robust	-	-	-	0.000	-	-	-	0.014	-	-	-	0.024
bootstrapped	-	-	-	-	-	-	-	0.024	-	-	-	0.031
adult illiteracy	-0.007	-0.012	-0.023	-0.012	-0.025	-0.023	-0.022	-0.022	-0.008	-0.008	-0.006	-0.007
robust	0.035	0.000	0.000	0.006	0.029	0.044	0.054	0.051	0.676	0.681	0.746	0.695
bootstrapped	-	-	-	-	0.040	0.069	0.085	0.054	0.674	0.688	0.732	0.703
log working pop / km2 arable land	0.175	0.173	0.054	0.180	0.258	0.264	0.286	0.268	0.066	0.060	0.051	0.060
robust	0.000	0.000	0.147	0.000	0.068	0.061	0.044	0.059	0.727	0.751	0.787	0.751
bootstrapped	-	-	-	-	0.095	0.106	0.043	0.076	0.748	0.774	0.796	0.774
p-value country FE	no	no	no	no	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
p-value time FE	0.838	0.927	0.999	0.937	0.849	0.891	0.828	0.894	0.345	0.356	0.407	0.352
nr observations	369	369	369	369	369	369	369	369	328	328	328	328
R2	0.990	0.988	0.992	0.9871	0.961	0.960	0.960	0.960	0.082	0.076	0.070	0.076
F-statistic	41.45	35.38	66.33	57.63	-	-	-	-	-	-	-	-
p-value F-test	0.000	0.000	0.000	0.000	-	-	-	-	-	-	-	-
p-value overID-test	0.845	0.631	-	-	-	-	-	-	-	-	-	-

Notes : p-values below coefficients. Bootstrapped p-values on the basis of 200 replications.

To control for both possible sources of endogeneity[97], we used an instrumental variable approach by using the distance to the USA and South Africa as instruments (see Figure 3.1)

[97] It also controls for the third way by which endogeneity issues may be raised, i.e. measurement error.

for our measures of market access[98]. The first four columns of Table 3.6 show that our results remain unaffected (also note that the overidentification and instrument relevance tests indicate that our instruments seem to be valid): all variants of market access still positively and significantly affect gdp per capita. Note however that the instruments are time-invariant, which precludes the use of country-fixed effects. Comparing our results to the first four columns of Table 3.5 (that also excludes country-fixed effects) thus provides the best insight into the effect of controlling for endogeneity by using our instruments. We observe that the coefficient on SSA (and total and ROW+SSA, that are largely made up of SSA market access, see Table 3.3) is much larger, whereas the coefficient on ROW market access is much lower. The results on human capital and density are largely unaffected.

Given the choice of instruments, the inability to control for fixed country effects thus constitutes a drawback of the IV-estimates. Columns 5-8 of Table 3.6 hence show the results when one includes each market access measure lagged one period (the human capital and density measure are also lagged one period). This arguably controls for one of the main reasons of possible endogeneity problems when estimating the NEG wage equation, namely reverse causality, while still allowing for the inclusion of country-fixed effects. Comparing these results to the last four columns in Table 3.5 shows that reverse causality does not seem to be a major issue, and, most importantly, we still find a positive effect of each of the market access variables on SSA's GDP per capita. Our final robustness check again concerns the way we try to control for unobserved country-specific variables that are correlated with our measures of market access. In our baseline results we capture these by including country-fixed effects. Another standard way of doing this is by estimating (3.12) in first differences. Compared to the fixed effect estimation, first differencing requires less strict assumptions regarding the exogeneity of lagged error terms (fixed effect requires strict exogeneity, i.e. any lagged error is uncorrelated with the included explanatory variables, whereas first differencing requires this for the first lag of the error process only). The last four columns show the results of estimating the wage equation (3.12) in first differences. The effect of market access, although slightly lower than when using fixed effects, is still significant and positive. The only substantial difference lies in the fact that ROW-market access is no longer significant (see also Table 3.4). Overall, the positive effect of SSA-MA that we find is most robust, indicating that current efforts to improve SSA-MA in particular, such as the Trans-Africa highway network that is being developed by among others the United Nations Economic Commission for Africa, the African Development Bank and the African Union (see also e.g. Buys, et al., 2006), are likely to contribute positively to SSA economic development.

3.6.2 Human capital and market access

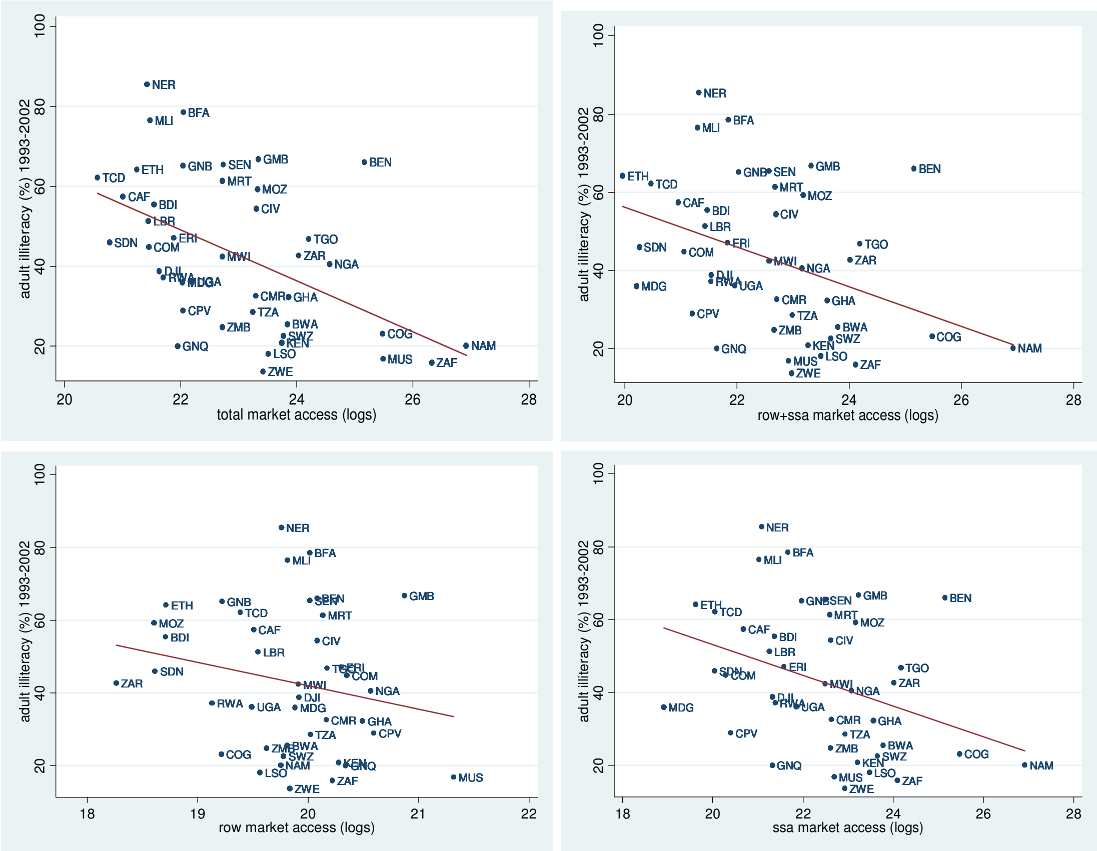
Besides the direct effect of a country's market access, Redding and Schott (2003) argue that it may also have an indirect effect through the accumulation of human capital. They show

[98] Distance to major markets is often used as an instrument in empirical studies in NEG, see e.g. Redding and Venables (2004) and Breinlich (2006). When considering total or SSA+ROW market access we use both distances as instruments, when considering SSA or ROW market access by itself, we use only distance to South Africa in case of SSA and only distance to the USA in case of ROW market access.

theoretically that if manufacturing is relatively skill and trade cost intensive, countries with better access to international market will experience increased incentives to invest in the education of their workforce. They also provide empirical evidence to back up their claim using a sample of 106 countries. Breinlich (2006) finds similar empirical evidence of a positive effect of market access on human capital when considering European regions.

Our baseline results in Table 3.5 showed that controlling for human capital still leaves a positive and significant direct effect of market access on gdp per capita. In this subsection we are concerned with a possible indirect effect of market access on income levels through its effect on human capital. Figure 3.3 shows that we also find a strong positive correlation between market access and human capital (as measured by the adult illiteracy rate) when considering our sample of 48 SSA countries. Interestingly, the correlation with ROW market access is somewhat weaker than with SSA (and also ROW+SSA and total) market access.

Figure 3.3 Market Access and Human Capital



Notes: the raw correlations of each of the market access variants with the adult illiteracy rate are (p-values in brackets): total: -0.46 (0.00); row + ssa: -0.35 (0.00); row: -0.19 (0.00); ssa: -0.33 (0.00).

Complementing Figure 3.3, Table 3.7 provides estimation results to assess the relevance of market access for human capital. It shows the results of regressing a logistic transformation of adult illiteracy (following Redding and Schott (2003) this makes sure that illiteracy is bounded between 0 and 1) on our four measures of market access while controlling for the positive effect of income per capita on human capital.

Table 3.7 Human capital and Market Access

dep: adult illiteracy	ols	ols	ols	ols
log tot ma	-0.143	-	-	-
robust	0.000	-	-	-
bootstrapped	0.000	-	-	-
log ssa+row ma	-	-0.110	-	-
robust	-	0.000	-	-
bootstrapped	-	0.000	-	-
log row ma	-	-	0.048	-
robust	-	-	0.307	-
bootstrapped	-	-	0.324	-
log ssa ma	-	-	-	-0.091
robust	-	-	-	0.000
bootstrapped	-	-	-	0.000
log gdp per capita	-0.488	-0.549	-0.630	-0.559
robust	0.000	0.000	0.000	0.000
bootstrapped	0.000	0.000	0.000	0.000
nr observations	369	369	369	369
R2	0.389	0.373	0.342	0.369

Notes : p-values below coefficients. Bootstrapped p-values on the basis of 200 replications.

The results confirm the positive relationship between total market access and human capital levels (remember that our dependent variable is adult illiteracy) in SSA, even after controlling for income levels. Strikingly, we do not find evidence of such a positive effect when considering only ROW market access. The significant positive effect of total (and ROW+SSA) market access seems to be entirely driven by the cross-country variation in SSA market access.

Finally, and to check whether these results regarding human capital depend on our measure of human capital, we also collected data on three different measures of human capital, namely youth (= under 25) illiteracy, the gross secondary enrolment rate and the primary completion rate. As can be seen in column 2 of Table 3.8, the coverage of the last two variables is, however, much poorer than both illiteracy variables. Together with the fact that adult illiteracy is highly correlated with the other three measures of human capital (see column 1 of Table 3.8), this is the main reason for us to include adult illiteracy in our baseline estimates of the wage equation in the previous section.

Table 3.8 Other Human Capital variables

	correlation with			total MA effect	HC effect
human capital variable	adult illiteracy	nr. observations	effect of total MA	in baseline	in baseline
adult illiteracy	1	369	-0.143 (0.00)	0.076 (0.00)	-0.021 (0.05)
youth illiteracy	0.97 (0.00)	369	-0.197 (0.00)	0.064 (0.00)	0.016 (0.10)
gross 2nd enrolment	-0.67 (0.00)	277	0.289 (0.00)	0.055 (0.02)	0.002 (0.12)
primary completion rate	-0.82 (0.00)	242	0.196 (0.00)	0.105 (0.01)	-0.001 (0.58)

Notes: p-values in brackets.

When regressing a logistic transformation of any of the other three human capital measures, column 4 of Table 3.8 also shows that we always find a positive effect of market access on human capital. In addition, as shown in the last two columns of Table 3.8, when substituting either of the other three human capital measures for adult illiteracy in our baseline regression (see Table 3.5), we always find a positive effect of market access, whereas the human capital variable is not always significant any more.

3.6.3 Policy experiments

Our results clearly show the importance for the SSA countries of improving their market access, both with the rest of the world as well as (or even more importantly) with other SSA countries. Our estimation results also help to gain insight into the relative effect of different policies or shocks aimed at improving a country's market access. As has already been discussed in section 3.4 (see also Redding and Venables (2004), section 7), the inclusion of country-specific variables allows us to perform policy experiments for both country-specific (e.g. infrastructure improvements, or efforts to end civil conflict) and country-pair specific variables (e.g. entering a regional or free trade agreement). The extent to which these policy measures improve a country's market access can first be inferred from the 1st step of our analysis, the estimation of the trade equation. Next, the effect of the resulting improvement in market access on GDP per capita easily follows from the estimated coefficient of market access shown in our baseline estimation results (Table 3.5). Table 3.9 and Figure 3.4 show the results of six such "policy" experiments. Four experiments focus on conflict-ridden Sudan and two on landlocked Ethiopia.

Table 3.9 Policy experiments

policy measure:	+ 1 s.d. infrastructure	end to civil war	all distances halved	RFTA with South Africa	no longer landlocked
country:	Sudan	Sudan	Sudan	Sudan	Ethiopia
		% increase in market access			
total	64.0	144.7	85.2	5.6	40.9
ROW+SSA	81.1	144.7	104.3	8.9	80.9
ROW	27.1	144.7	102.4	0.0	22.6
SSA	89.4	144.7	104.7	10.6	99.1
		resulting % increase in gdp per capita			
total MA	4.9	11.0	6.5	0.4	3.1
ROW+SSA-MA	4.3	7.7	5.6	0.5	4.3
ROW-MA	2.3	12.0	8.5	0.0	1.9
SSA-MA	4.4	7.2	5.2	0.5	4.9

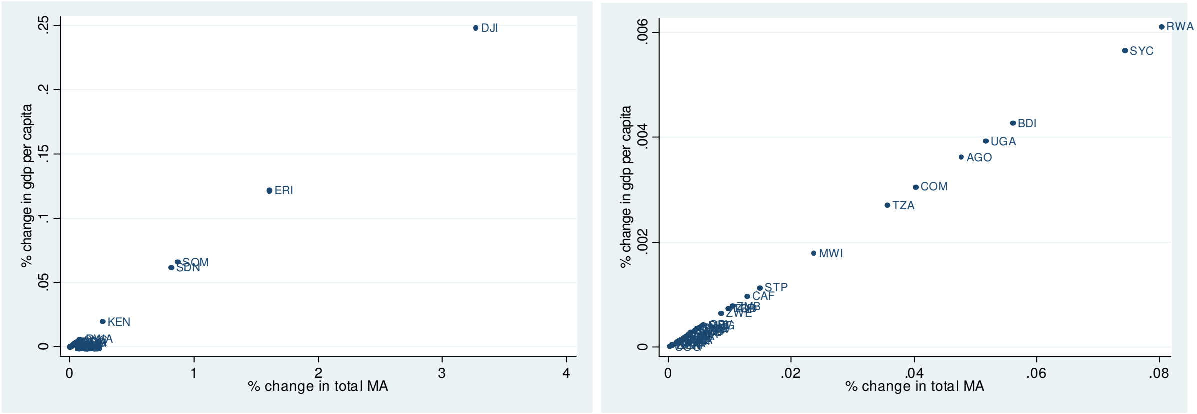
First the results for Sudan. Ending the civil war (Darfur) in that country would increase its market access the most[99] and raises its GDP per capita by around 10% depending on the measure of market access (and subsequent estimate of its effect on gdp per capita). This shows the devastating impact of civil unrest on SSA's economic development in general. Hypothetically halving Sudan's distance to all its trading partners also increases market access substantially and would increase GDP per capita by about 6.5%. Of particular interest are the effects of investments in infrastructure and the establishment of regional free trade

[99] Note that the % increase in total market access is the same as for each of its subcomponents as we do not allow the coefficient on civil war to be different when considering intra-SSA trade or SSA trade with the ROW when estimating the trade equation (see Table 3.2).

agreements. The results show that improving Sudan's infrastructure by one standard deviation (resulting in a quality of infrastructure comparable to Namibia) would raise GDP per capita by about 5%, whereas forming a bilateral RFTA with South Africa would only raise its GDP per capita by 0.4%. The reason for this difference is that improvements in infrastructure affect all trading partners alike, whereas the establishment of a bilateral RFTA only affects one trading partner. This gives a clear policy recommendation: policies aimed at improving a country's ability to trade will have a much higher pay-off when they aim at general improvements that affect as many of that country's trading partners as possible.

An example of such an 'all-trade-partners-affecting' experiment is the one shown for Ethiopia in Table 3.9. When Eritrea officially became independent in 1993, Ethiopia lost its direct access to the sea. Table 3.9 shows that if this had not taken place Ethiopia's market access would ceteris paribus have remained much better, resulting in an improvement of its GDP per capita of about 3 to 4%.

Figure 3.4 The spatial reach of a 10% positive GDP shock in Ethiopia



The final experiment that we conducted, is not so much concerned with improving a country's market access by trade cost reducing policies, but instead it focuses on the spatial reach a one-time positive exogenous shock in Ethiopia's GDP of 10%. This improves other countries' market access through the increased demand from Ethiopia for their products, and the more so, the lower trade costs with Ethiopia. Given the estimated penalty on distance and the positive border effect (see Table 3.2), Figure 3.4 clearly shows that Ethiopia's neighbors (Djibouti, Eritrea, Somalia and Sudan) will benefit the most from the increased demand from Ethiopia, improving their market access by more than 1% and resulting in an improvement of GDP per capita in the range of 0.1% to 0.25%[100]. By zooming in on the least affected countries the right panel of Figure 3.4 shows that, by zooming in on the least affected countries a clear spatial pattern remains visible with the more distant countries affected the least.

[100] The overall effect is small compared to some of the trade cost experiments; this is again due to the fact that for all the affected countries, Ethiopia constitutes only one of many trading partners (and mostly also a relatively unimportant one).

3.7 CONCLUSIONS

The main message of this chapter is that market access is important for economic development in Sub-Saharan Africa (SSA). Based on a sample of 48 SSA countries for the period 1993-2002 and controlling for various econometric and economic issues like the role of human capital, we invariably find that improved market access positively affects income per capita. Building on the empirical framework used by Redding and Venables (2004) that is firmly grounded in new economic geography theory, we first estimated a trade model by using bilateral trade data, explicitly allowing for a different impact of trade costs on intra-SSA trade and SSA trade with the rest of the world (ROW). Second, we used the trade estimation results to construct measures of market access for each SSA country and look at the impact of market access on GDP per capita, again distinguishing explicitly between intra-SSA and ROW market access. The fact that, among our market access measures, intra-SSA market access has the most robust and a relatively large impact on economic development suggests that current policies aimed at improving SSA infrastructure and at encouraging further economic integration of SSA countries are likely to pay off in the future. This is further strengthened by our finding of a possible additional indirect effect of market access on income levels through improvements in human capital (as argued by Redding and Schott, 2003). More generally, and in line with claims by for instance Collier and Venables (2007), policies aimed at improving Sub-Saharan African access to international markets for manufactured goods seem to be important, but would constitute a shift away from policies (exclusively) aimed at promoting growth through agricultural goods. Above all, see also Henderson, Shalizi and Henderson (2001), our results are a reminder that distance or relative geography matters for economic development. Despite room for (policy-induced) improvements in market access, the (economic) remoteness of Sub-Saharan Africa remains a main deterrent to its economic development.

APPENDIX 3.A DATA DEFINITIONS AND SOURCES

GDP (also per capita and per worker)
Gross Domestic Product (also per capita and per worker), from Penn World Tables 6.2, if not available (for Angola, Haiti, French Polynesia, New Caledonia, Azerbaijan, Armenia, Belarus in selected years) from World Bank Development Indicators 2003 or World Bank Africa Database 2006.

Distance
Great circle distance between main cities, from CEPII.

Internal distance
This often-used specification of D_{ii} reflects the average distance from the center of a circular disk with $area_i$ to any point on the disk (assuming these points are uniformly distributed on the disk). It is calculated on the basis of a country's area: $D_{ii} = 2/3\left(area_i/\pi\right)^{1/2}$.

Contiguity
Dummy variable indicating if two countries share a common border, from CEPII.

Common official language
Dummy variable indicating if two countries share a common official language, from CEPII.

Common colonizer
Dummy variable indicating if two countries have been colonized by the same colonizer, from CEPII.

Colony – Colonizer relationship
Dummy variable indicating if two countries have ever had a colony-colonizer relationship, from CEPII.

Landlocked
Dummy variable indicating if a country has no direct access to the sea.

Infrastructure index
Following Limao and Venables (2001), the index constructed as the unweighted average of four variables (each normalized to have a mean of 0 and standard deviation 1 over the whole sample period as well as in each year). The four components are:

- Roads
Km road per km2, from World Bank Development Indicators 2003, World Bank Africa Database 2006 and Canning (1998).

- Paved roads
Km paved road per km2, from World Bank Development Indicators 2003, World Bank Africa Database 2006 and Canning (1998).

- Railways
Km railways per km2, from Canning (1998).

- Telephone main lines
Telephone main lines per 1000 inhabitants, from World Bank Development Indicators 2003, World Bank Africa Database 2006 and Canning (1998).

As Limao and Venables (2001) I ignore missing values, making the implicit assumption that the four variables are perfect substitutes to a transport services production function.

Island
Dummy variable indicating if a country is an island.

Regional or Free trade agreement
Dummy variable indicating if two countries are both a member of one of the following regional or free trade agreements: ECOWAS, ECCAS, COMESA, SADCC, UEMOA, CEMAC (or UDEAC), EAC, IGAD or CENSAD.

Civil conflict
Dummy variables indicating if a country experienced the use of armed force between two parties, of which at least one is the government of a state, that results in at least 25 and at most 999 battle-related deaths, from the International Peace Research Institute, Oslo.

Civil war
Dummy variables indicating if a country experienced the use of armed force between two parties, of which at least one is the government of a state, that results in at least 1000 battle-related deaths, from the International Peace Research Institute, Oslo.

% rural population
Share of the population living in rural areas, from World Bank Development Indicators 2003 and World Bank Africa Database 2006.

% labor force in agriculture
Average proportion of the total labor force recorded as working in agriculture, hunting, forestry, and fishing (ISIC major division 1) over the period 1993-2002. Labor force comprises all people who meet the International Labour Organization's definition of the economically active population, from World Bank Development Indicators 2003 and World Bank Africa Database 2006.

Adult illiteracy
The percentage of the population that is 25 years and older that cannot read or write, from World Bank Development Indicators 2003 and World Bank Africa Database 2006.

Youth illiteracy
The percentage of the population under 25 that cannot read or write, from World Bank Development Indicators 2003 and World Bank Africa Database 2006.

Gross secondary enrolment
Gross enrolment ratio is the ratio of total enrolment, regardless of age, to the population of the age group that officially corresponds to the level of education shown. Secondary education completes the provision of basic education that began at the primary level, and aims at laying the foundations for lifelong learning and human development, by offering more subject- or skill-oriented instruction using more specialized teachers, from World Bank Development Indicators 2003 and World Bank Africa Database 2006.

Primary completion rate
Primary completion rate is the percentage of students completing the last year of primary school. It is calculated by taking the total number of students in the last grade of primary school, minus the number of repeaters in that grade, divided by the total number of children of official graduation age, from World Bank Development Indicators 2003 and World Bank Africa Database 2006.

Working population per km2 of arable land
Data on the working population and the km2 of arable land are separately taken from the World Bank Development Indicators 2003 and the World Bank Africa Database 2006.

Oil exporter
Dummy variable indicating whether a country is exporting oil. From the World Bank Africa Database 2006.

Exports of ores and metals
Exports of ores and metals as a percentage of total merchandise trade. From the World Bank Africa Database 2006.

Part II

Chapter 4

Regional income evolution in South Africa after Apartheid[101]

4.1 INTRODUCTION

The end of Apartheid in 1994 changed South Africa's economic landscape considerably. It marked the end of its exclusion from the international markets and released the largest part of its population from severe restrictions on both economic and physical mobility. Subsequently the economy has been growing steadily and by March 2007 it experienced 29 quarters of uninterrupted expansion. At times in 2005 and 2006 the real growth rate accelerated to an annualized rate of 5 per cent. Clearly the national economy has been doing well, but the robust growth of the national economy masks significant differences in economic performance across regions.

Indeed, a characteristic of (the development of) South Africa's regional economic activity is its spatial unevenness. Over the period 1996 to 2004 only 53 regions grew at rates faster than the national average of 2.9 per cent. Of those, only eight recorded growth in per capita GDP in excess of 5 per cent per annum. At the same time, 81 regions grew by less than one per cent per annum and in 78 regions GDP per capita even fell. So that in 2004 only 48 regions out of 354 (or 13 per cent) produced approximately 50 per cent of total South African output, 71 regions produced 60 per cent of the country's output and the top 100 regions produced 70 per cent of GDP. The changes in South Africa's economic landscape are clearly not affecting all regions in the positive way suggested by the overall national growth rate.

This is significant for the decentralization of economic decision-making that has taken place since democratization in 1994. A recent report by the Center for Development and Enterprise shows that "middle South Africa" is characterized by slow, jobless economic growth, little external or local investment and emigration or internal migration of the young and educated (Bernstein and McCarthy, 2005). Most of the country's local governments have struggled to cope with their constitutional responsibility to develop their areas, as was clearly marked by protests against poor service delivery in the run up to the 2006 local elections.

A first step to addressing the challenges of low growth, poverty and inequality is to examine spatial economic growth patterns determining which are the fast and slow growing regions, what are their characteristics and whether the poorer regions are catching up or falling behind. Porter (2003) stresses the need to carefully examine regional economic performance as these regional economies are of paramount importance to the performance of the overall national economy. Recent work by Naudé and Krugell (2003; 2006) has already examined the determinants of economic growth at the sub-national level in South Africa. Their results, obtained using dynamic panel data techniques, suggest slow conditional beta

[101] This chapter is an adapted version of Bosker and Krugell (2008) that is forthcoming in the Journal of Regional Science.

convergence (Naudé and Krugell, 2003). On the contrary, an exploratory distribution dynamics approach suggested that divergence occurred (Krugell, Koekemoer and Allison, 2005).

This chapter builds on the latter approach and uses (spatial) Markov chain techniques to describe the evolution of the entire South African regional income distribution in terms of its intra-distributional dynamics. Estimating Markov chains (first introduced by Quah, 1993a) is an attractive empirical method to model the evolution of (regional) income inequality by virtue of its ability to accommodate shocks, discontinuities and ongoing turbulence in the growth process (Fingleton, 1999). This is especially useful as such features of the growth process correspond to predictions made by both the new growth (see e.g. Aghion and Howitt, 1998) and the new economic geography (see e.g. Fujita, Krugman and Venables, 1999) literature. The approach can reveal aspects of the (regional) growth process that may remain hidden when using the more standard growth regressions in the spirit of Barro and Sala-i-Martin (1991). Moreover, recent contributions in the field of spatial econometrics (Rey, 2001; Le Gallo, 2004 and Mossi et al., 2003) allow for the explicit incorporation of the spatial nature of our regional dataset in the Markov Chain analysis. These techniques provide a closer look at the effect of economic conditions in neighboring regions on regional economic growth, so that predictions regarding the effect of trade openness and/or increased labor mobility on the distribution of economic activity that follow from the geographical economics literature (see e.g. Krugman, 1991; Krugman and Livas Elizondo, 1996; Puga (1999) or chapter 1 of this thesis), and/or the regional economic literature (see e.g. Ciccone and Hall, 1999 and Ciccone, 2002), can be looked at in more detail.

The chapter is structured as follows. Section 4.2 presents a brief history of South Africa's regional economy, highlighting the role that trade and natural resources played in the initial location of economic activity and how this was later reinforced by policy. Section 4.3 presents results of a diverging regional income distribution using the above-mentioned Markov Chain techniques. Section 4.4 explicitly takes account of the spatial nature of the data providing evidence of localized growth poles using recently developed spatial Markov Chain techniques. Section 4.5 identifies possible determinants of the observed evolution of South African regional incomes, describing the characteristics of the 25 fastest and slowest growing regions, and linking this chapter's results back to both theory and to earlier regression estimates in the literature. Finally, section 4.6 concludes.

4.2 A BRIEF HISTORY OF THE SOUTH AFRICAN SPATIAL ECONOMY

To understand the recent spatial patterns of economic growth in South Africa it is necessary to provide some background information on factors such as geography and policy that have shaped an economy characterized by an unequal regional distribution of economic activity.

4.2.1 The spatial economy up to 1994

The port cities of Cape Town and Durban were first developed in the 17th and 18th centuries as trading posts on the shipping route between Western Europe and Asia. During the 19th century this role changed with the discovery of diamonds and gold in the interior, and they developed from being stopover and service points providing shipping services, to being ports through which commodities were handled. Mineral wealth determined the location and growth of the other two dominant cities, Johannesburg and Pretoria. The location of minerals as well as the extraction technology required in mining then influenced the pattern of South Africa's inland development. Where railways and electric power were provided for mining, these also contributed to the development of the manufacturing sector.

Next, Apartheid left its mark on regional economic development. The Apartheid era extended from 1948 to 1994 and was marked by a number of policies that shaped the spatial economy. The homelands policy restricted black communities to their traditional tribal areas and aimed to limit migration to the cities. To this end incentives were provided for industrial development and job creation in remote rural areas, which were referred to as decentralized growth points. In related policy, the Group Areas Act restricted the settlement and movement of black people in urban areas. This left them in informal settlements on the periphery of cities and towns.

Overall, trade and extraction, along with Apartheid left South Africa with a spatial economy characterized by excessive transport costs, segmented labor and consumption markets and great disparities in regional development (Krugell and Naudé, 2005) [102].

4.2.2 The spatial economy after Apartheid

Since the new democratically elected government came to power in 1994, South Africa has been marked by the lifting of the sanctions imposed by the international community (opening it up to the world economy), fiscal adjustment, a fluctuating exchange rate and moderate economic growth rates. Along with the lifting of the severe restrictions on economic and physical migration posed by Apartheid on the largest part of the population, which resulted in increased rural-urban migration, government policy shaped the spatial economy after Apartheid in two other ways.

The first is industrial policy. In 1994 the new government started out with an 'a-spatial' industrial policy. For example, in 1997 a tax holiday scheme was introduced to encourage industrial development throughout the country. In recent years, however, thinking has swung back to spatially focused policies in the form of Spatial Development Initiatives (SDI) and Industrial Development Zones (IDZ) (see Nel, 2002). Government seeks to encourage investment, manufacturing and other economic activities along a series of defined transport corridors. It does however not yet appear that the initiatives have brought about dramatic economic transformation in their areas.

Along with the spatial focus of industrial policy, South Africa's regional economy has recently also been shaped by the local government transition process that has reduced the number of municipalities and has made local government the primary institutional vehicle for economic development in South Africa (Naudé and Krugell, 2006). Most of the country's municipalities have struggled to cope with their constitutional responsibility to economically develop their region, as clearly marked by protests against poor service delivery in the run up

[102] See Krugell (2005) for a detailed account of the history.

to the 2006 local elections.

In conclusion, this brief historical overview shows the complexity of issues, in terms of geography, history and policy, that have all had their effect on the location of economic activity in South Africa. The aim of the next two sections is to provide a detailed picture of the evolution of regional GDP per capita across 354 South African regions during the turbulent post-Apartheid period from 1996 to 2004 that provides useful guidelines when looking for explanations (and potential solutions) to South Africa's regional disparities.

4.3 THE EVOLUTION OF THE REGIONAL INCOME DISTRIBUTION

Recent empirical work that has looked at the development of regional South African GDP per capita differences, focused largely on the use of (un)conditional growth regressions in the spirit of e.g. Barro and Sala-i-Martin, 1991 and Mankiw, Romer and Weil, 1992. Naudé and Krugell (2003; 2006) examine the determinants of regional economic growth in South Africa, over the period 1998 to 2002. Their results suggest slow conditional beta convergence, with the poorer regions very slowly catching up with the richer ones.

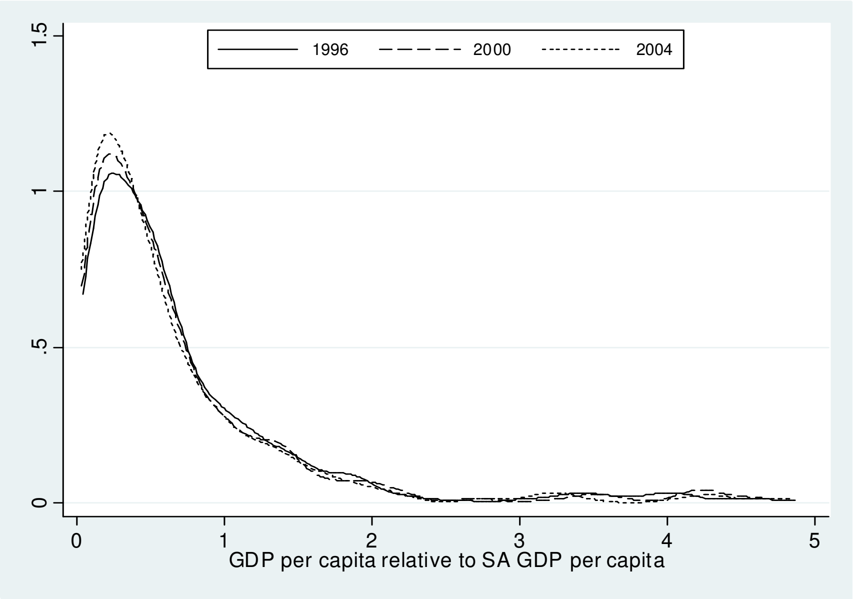
The use of these growth regressions has however not remained free of criticism. Besides raising some econometric issues such as endogeneity and heterogeneity problems, these growth regressions may be plagued by Galton's fallacy of regression to the mean (see Friedman, 1992 and Quah, 1993b). Arguably even more problematic is the fact that, when performing a standard growth regression, one assumes that the growth rate of the production efficiency (or technological development) is both region and time invariant. Hereby it does not allow for a process of technology adaptation and/or catch up (see also Lee, Pesaran and Smith, 1998), which lies at the heart of some of the recent theoretical developments in the field of economic growth (see Aghion and Howitt 1998 for a good overview). Also, recent insights from the new economic geography (NEG) literature (see Fujita et al., 1999 and Baldwin, Martin and Ottaviano, 2001) indicate that the growth process is less smooth than predicted by the neoclassical growth models in the spirit of Solow (1956). Instead, it can even be discontinuous and depends to a large extent on the developments in neighboring regions: economic development does not take place in isolation (see also Fingleton and López-Bazo 2006). Related to these issues is the fact that the main focus of a growth regression is on the economic performance of the representative regional economy, which leaves one unable to say something about the dynamics of the entire cross-sectional distribution.

The above-mentioned caveats of performing (un)conditional growth regressions have led to the development of a different empirical method to look at the evolution of regional income differences over time. First introduced by Quah (1993a; 1993b; 1996a; 1996b), this method involves modeling the evolution of the entire cross-section income distribution in terms of a homogenous Markov Chain process. By quantifying the evolution of the income distribution in terms of its intra-distributional dynamics, this approach describes the evolution of regional incomes during a certain period very accurately. It is an attractive empirical method to model the evolution of (regional) income inequality by virtue of its ability to accommodate shocks, discontinuities and ongoing turbulence in the growth process (Fingleton, 1999). Also, it provides the researcher with empirical results that allow him or her to assess the relevance of predictions following from both the neoclassical and the new growth theory without making some of the restrictive assumptions when performing growth regressions (see Fingleton, 1997 and 1999). A drawback of the method is however that while giving a very accurate description of the evolution of regional incomes, the approach is unable to provide statistical evidence regarding the important factors that are behind the observed evolution. This shortcoming is addressed in more detail in section 4.5, where we link our findings to both theory and recent estimates obtained using more standard growth regressions (Naudé and Krugell, 2003 and 2006).

4.3.1 The regional per capita income distribution

Figure 4.1 provides a first look at (the evolution of) South Africa's regional GDP per capita distribution[103].

Figure 4.1 South Africa's regional income distribution



Notes: Over the sample period the SA GDP per capita rose from R19875 to R22125 (or in PPP adjusted 2000 I$ from I$ 7308 to I$ 9146).

We have collected the data from the Regional Economic Explorer database compiled by Global Insight Southern Africa (see *www.globalinsight.co.za*) that combines sub-national economic information from South Africa's census, government departments, development agencies and Regional Services Councils. The sub-national regions that we analyse are magisterial districts[104]. To control for general South Africa wide trends and business cycle

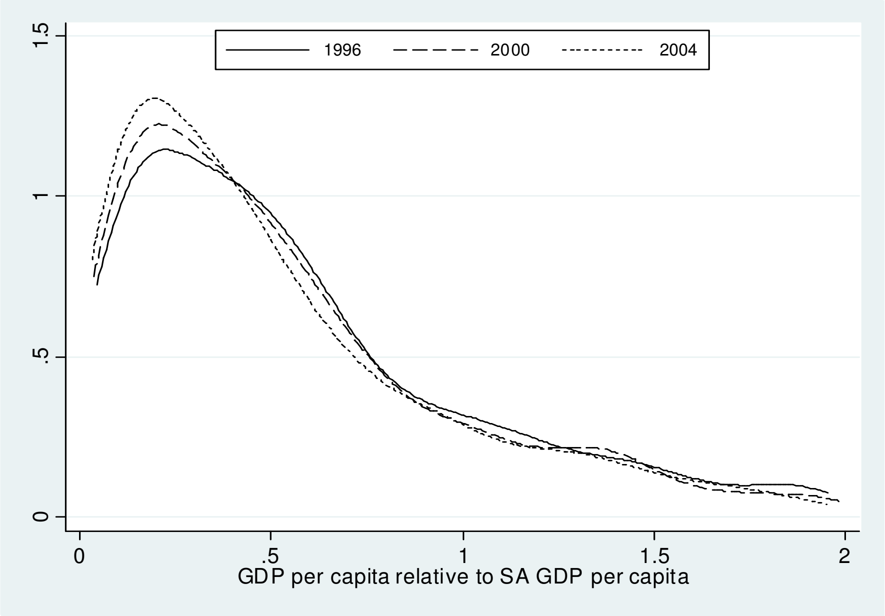
[103] The distributions are obtained by kernel estimation methods using a Gaussian kernel with the optimal bandwidth chosen using the method proposed in Silverman (1986).

[104] The choice of this spatial unit of analysis comes down to the availability of data over the period 1996 to 2004. Before 2000 there were 843 municipalities in South Africa, which were then reduced to 237 to constitute so-called wall-to-wall local government. This change in demarcation complicates analysis over an extended period. At the lower level of magisterial district, data are available over the whole period. Also, the magisterial districts define the location of cities and towns, whereas the municipalities are governing bodies of larger areas. Throughout the article the terms regions or areas are used to refer to these sub-national economies.

effects, regional incomes are measured relative to the overall South African GDP per capita. This means that 1, (> 1), [< 1] on the horizontal axis denotes a region with a GDP per capita equal to, (large than), [smaller than] the South African GDP per capita respectively.

The estimated distributions immediately show that South Africa is characterized by many regions that have a GDP per capita far below the national level, i.e. smaller than 0.5 times SA GDP per capita, and relatively few regions with a GDP per capita above the national level. Also, some regions are more than three or four times as rich as the average South African region. Due to the wide dispersion observed in Figure 4.1, it is difficult to clearly see the evolution of the distribution. Therefore, Figure 4.2 below zooms in on the left (0 – 2 times South African GDP per capita) of the estimated regional per capita income distribution[105]. It gives a somewhat clearer view of the dynamics of the external shape of the distribution. Note that between 1996 and 2004 the distribution gains mass in the left tail, i.e. the number of regions with a GDP per capita level below 0.4 times the national level increases. Also (but less pronounced) the distribution loses mass in the range 0.4 – 0.8 times the national average over that period.

Figure 4.1 South Africa's regional income distribution – zoomed in on the left



Notes: Over the sample period the SA GDP per capita rose from R19875 to R22125 (or in PPP adjusted 2000 I$ from I$ 7308 to I$ 9146).

The overall picture emerging from this first look at (the evolution of) the regional income distribution already shows some interesting things. First, the South African economy is characterized by large regional income inequalities, confirming the notion of substantial regional disparities in economic development mentioned in section 4.2. The extent of these disparities is much larger than in Europe (Le Gallo, 2004; Magrini, 1999) and the USA (Rey,

[105] The result when zooming in on the right of the distribution is not shown as this does not provide a clearly discernable pattern over time. This is partly due to the relatively few regions with such high GDP per capita levels, making the estimated distribution more susceptible to GDP per capita changes in only a few regions.

2001; Johnson, 2000) but also than other countries that are at a similar stage of economic development as South Africa, e.g. India (see Bandyopadhyay, 2004), Brazil (see Mossi et al., 2003) or China (see Tianlun et al., 1996 and Aziz and Duenwald, 2001). Furthermore, it seems to be the case that over the period 1996 to 2004 more and more regions have become relatively poorer, suggesting divergence in regional income levels (something also found in Krugell et al., 2005). This contrasts with the results obtained from standard cross-region growth regressions in e.g. Naudé and Krugell (2003) that showed evidence in favor of (be it very slow) conditional convergence in income levels. Looking at the evolution of the entire regional income distribution thus offers a much more pessimistic picture.

4.3.2 Markov chain analysis

The evolution of the external shape of the regional per capita GDP distribution does not give any idea about the movement of regions within this distribution. In other words, one does not know if the increase in the number of poorest regions is due to richer regions becoming poor, or already poor regions becoming even poorer. It is these intra-distributional movements that are of particular interest. In order to quantify them a Markov chain analysis is performed.

This draws upon the work by Quah (1993a) and allows one to quantify the dynamics of the regional income distribution as a whole in terms of the intra-distributional dynamics of the individual regions making up this distribution. The use of Markov chain techniques requires the distribution to be discretized, i.e. each region is assigned to one of a pre-specified number of groups based on its relative GDP per capita level. Letting f_t denote the vector of the resulting discretized distribution at period *t* and making the assumption that the distribution follows a homogenous, stationary, first order Markov chain process, the evolution of the discretized distribution can be characterized as follows:

$$f_{t+x} = M\ f_t \tag{4.1}$$

where *M* is the so-called *x*-period transition matrix that maps the distribution at time *t* into period *t+x*. Each element of the transition matrix, m_{ij}, denotes the probability that a region moves from income group *i* in period *t* to income *j* in period *t+x*.

Given the number of regions and time periods for which income data is available, the number of income groups used to discretize the distribution is chosen to be five (see also Le Gallo (2004) and Quah (1996a)). To assign each region to one of the five income groups on the basis of its South Africa relative GDP per capita, the boundaries of each income group need to be chosen. We follow the recommendation by Quah (1993a) and choose these such that the initial number of regions in each of the income groups is the same. However, we are aware that the particular discretization chosen could in principle lead to very misleading results (see Bulli, 2001). Discretization very likely results in the loss of the Markov property (that the transition probability of moving from state *i* to state *j* does not depend in any way on how a region reached state *i* in the first place). Different discretization methods than the one we use have been proposed (Bulli, 2001; Magrini, 1999). We, however, choose to stick with the ‘simple’ choice of cutoffs as proposed by Quah (1993a), and subsequently applied by e.g. Rey (2001), Le Gallo (2004), Lopez-Bazo et al. (2004) and Bosker (2006), for the following reasons.

First, we think that our discretization is an adequate approximation of the underlying continuous regional growth process (which is all that matters according to Bartholomew 1981). Our results are qualitatively robust to the use of different (sensible) cut-offs and the discretization of the continuous distribution into seven instead of five income groups[106]. Second, the use of a different method proposed by Magrini (1999) and also used in Cheshire and Magrini (2000) to select the (number of) cut-offs on the basis of the ability of the resulting discretized distribution to approximate the continuous distribution, can in our view result in very misleading results as this method is much more sensitive to outliers[107]. And third, the estimation of stochastic kernels (see Figures 4.A1 and 4.A2 in Appendix 4.A) in the spirit of Quah (1997), Johnson (2000) and Fingleton and Lopez-Bazo (2003) and the calculation of a polarization index (see Figure 4.3) support and/or complement the main results obtained using our chosen discretization. We focus on the results obtained using the discretized distribution, referring to Appendix 4.A for the stochastic kernel estimates, as they are easier to interpret and allow for the computation of many interesting indices and statistics (see Bulli, 2001). Also, discretizing the distribution alleviates the problem of measurement error (and outliers), which may plague stochastic kernel estimates and too finely discretized distributions.

Our choice of discretization results in the following five relative income groups: regions with a GDP per capita (1) less than 17%, (2) between 17% and 36%, (3) between 36% and 57.5%, (4) between 57.5% and 105% and (5) larger than 105% of the South African GDP per capita. Having thus discretized the income distribution we can estimate each of the elements of the transition matrix, *M*, which given that the data set contains income data on a yearly basis is chosen to be the 1-year (*x =1*) transition matrix, by maximum likelihood, i.e.

$$\hat{m}_{ij} = \frac{\sum_{t=1}^{T-1} n_{it,jt+1}}{\sum_{t=1}^{T-1} n_{it}} \tag{4.2}$$

where $n_{it,jt+1}$ denotes the number of regions moving from income group *i* in year *t* to income group *j* in year *t+1*, and n_t the number of regions in group *i* in year *t*. Table 4.1 shows the resulting estimate of this 1-year transition matrix. It also shows which of the estimated transition probabilities are at least significant at a 10% level, following the suggestion made in Bosker (2006) who argues that also the standard errors of the estimated transition probabilities should be estimated[108].

[106] All results are available upon request.

[107] In Magrini (1999) for example the use of eleven income groups may approximate the continuous distribution best but his method of boundary selection leads to having income groups, those in the tails of the distribution, containing very few observations (sometimes less than three), shedding serious doubts on the results found in, and the conclusions drawn from, his subsequent Markov chain analysis.

[108] Standard errors are calculated as follows: $\hat{\sigma}_{ij} = \sqrt{\hat{m}_{ij}(1-\hat{m}_{ij})/N_i}, \quad N_i = \sum_{t=1}^{T-1} n_{it}$.

Table 4.1 Estimated Markov 1-year transition probabilities

x SA GDP/cap	< 0.17	0.17 < < 0.36	0.36 < < 0.575	0.575 < < 1.05	> 1.05
< 0.17	0.997	*0.003**	0	0	0
0.17 < < 0.36	0.028	0.961	0.011	0	0
0.36 < < 0.575	0	0.048	0.931	0.021	0
0.575 < < 1.05	0	0	0.048	0.936	0.016
> 1.05	0	0	0	0.026	0.974

Notes: Italic transition probabilities marked with a * are not significant at the 10% level. Each of the numbers in the table shows the probability that a region in a particular income group moves to another income group the next year. For example, a region in the lowest income group has a probability of 0.3% to move one income group up, whereas a region in the 2nd highest income group has a probability of 1.6% to make the transition to the highest income group. The p-value likelihood ratio test for time homogeneity (see Bickenbach and Bode, 2003) of the transition probabilities (splitting the sample period in half) is 0.604.

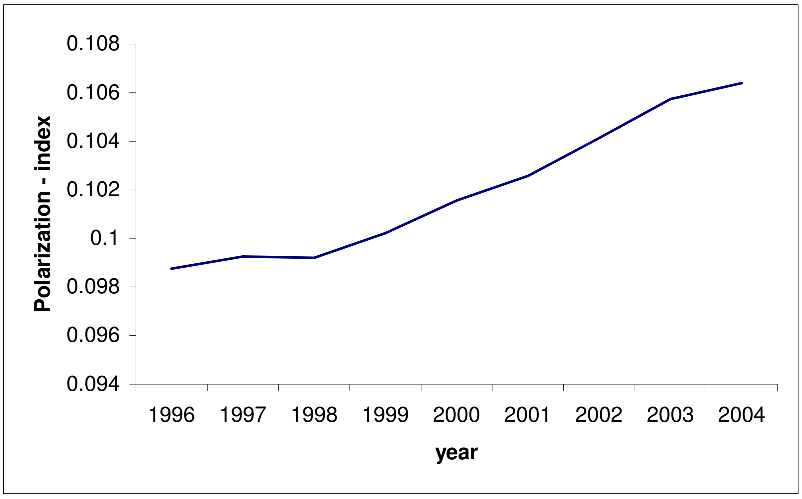
The estimated probabilities show some interesting features. First note that the probability of one of the poorest regions to move up in the income distribution is not significantly different from zero, indicating that the probability of these very poor areas to improve their economic condition(s) is very small; they might be caught in a regional poverty trap (see Ravallion, 1996). Furthermore, when comparing the off-diagonal elements of the matrix with each other, it is quite striking to note that, except for the richest regions with a GDP per capita higher than the national level (who are the least likely to move a group down in the discretized income distribution), all regions are more likely to experience a fall in their South Africa relative GDP per capita resulting in a move to a lower income class than an increase in income level resulting in a move up in the discretized distribution (Figure 4.A1 showing the corresponding stochastic kernel estimate confirms this finding). This suggests that during the post-Apartheid period, the richest regions have consolidated or increased their lead in terms of economic development over the other regions in the country, despite active government involvement in promoting economic development of backward regions (see Nel, 2002). Also, following the suggestion in Bickenbach and Bode (2003), the p-value of a likelihood ratio test for the assumed time homogeneity of the transition probabilities is shown. Time-homogeneity is not rejected, giving confidence in the estimated transition probabilities and the conclusions drawn from them regarding the degree of intra-distributional mobility (see the next sub-section(s)) within the regional income distribution.

It is however possible that the discretization of the income distribution hides some aspects of the evolution of the income distribution due to the discretization chosen (see the earlier discussion on the choice of discretization). A somewhat extreme example can illustrate this quite clearly: suppose that the resulting estimated transition matrix would be the identity matrix, suggesting that the income distribution is quite stable. At the same time, however, it could be the case that all regions in the highest income group earn twice as much, and all regions in the lowest income groups twice as little (with the same number of regions in each of the income groups), such that overall South African GDP per capita does not change. This, although clearly resulting in a diverging income distribution, would not be picked up by the Markov chain analysis, as no region shifts between income groups. Of course this example is quite extreme and would surely show up in the kernel estimates of the regional income distribution, but it serves illustrational purposes. Such patterns not picked up by the Markov chain analysis can be picked up by looking at the evolution of mean income in each of the income groups. More specifically one can calculate a measure of polarization following Esteban and Ray (1994). Their proposed polarization measure is a weighted sum of the difference in log GDP per capita between all possible income groups, i.e.:

$$ER = \sum_{i=1}^{k} \sum_{j=1}^{k} f_i^{1+\alpha} f_j \mid \overline{y}_i - \overline{y}_j \mid \qquad (4.3)$$

The index is the sum of the absolute difference in the log of the conditional means of GDP per capita, $\overline{y}$, between all combinations of income groups of the discretized income distribution $i, j \in k$ (k denoting the number income groups) weighted by their respective frequency in the distribution, *f*. α determines how heavy polarization is weighted with higher values of α resulting in a heavier weighting of polarization. Being a weighted sum of the absolute difference in average income between income groups, the higher the ER index, the more polarized the income distribution is. Figure 4.3 below shows the evolution of this index over the period 1996-2004. The polarization index shows that after being quite stable up to 1998, it gradually increases over the period after that (i.e. 1998-2004), suggesting that besides the movement shown in the discretized distribution (more regions in the poorer income classes) the difference between groups in terms of average GDP per capita has also increased over the sample period.

Figure 4.3 Polarization between income groups



Notes: α is set to 1.5 when calculating the ER-index (see (4.3)). This follows Le Gallo, 2004 and Bosker, 2006 and puts a high weight on polarization.

This further strengthens the notion of a (heavily) diverging South African regional income distribution characterized by a widening of the, beforehand already present, and large, per capita income differences between regions during the post-Apartheid period[109].

[109] A concern may be that the degree of regional inequality provides a distorted picture of the severity of the problem of income inequality in South Africa. It could be the case that the observed regional inequality overstates the degree of personal income inequality. This would happen when most of the poor regions are sparsely populated, so that the largest part of the population resides in the high income per capita regions. In our

Given the estimated transition probabilities, one can also calculate several mobility indices that provide interesting information about the degree of regions' intradistributional mobility and the speed at which the regional income distribution is evolving. The Shorrocks' (1978) index is calculated as:

$$SI = \frac{k - tr(M)}{k - 1} \tag{4.4}$$

where k denotes the number of discretized income groups and $tr(M)$ the trace of the estimated transition matrix, M. It takes on values on the interval $[0, k/(k-1)]$ with lower values indicating less mobility between income groups. The half-life on the other hand tells something about the speed at which the income distribution is changing over time[110]; it is calculated as follows:

$$HL = -\frac{\ln(2)}{\ln|\lambda_2|} \tag{4.5}$$

where λ_2 denotes the second largest eigenvalue of the estimated transition matrix, M. The higher the half-life, the slower the distribution is evolving. The two indexes are shown in Table 4.2.

Table 4.2 Mobility indices

Shorrock's index	Half life
0.050	58.28

They show that, although the regional income distribution tends to evolve quite slowly over time (half-life of 58 years) towards a situation characterized by more regional inequality (suggesting some time for policymakers to turn the tide), the low degree of intra-distributional mobility indicated by the Shorrocks' index shows that the problems facing the poorest regions are probably quite substantial to overcome (see section 4.5 for more details on these problems). The regions that are relatively rich (poor) remain rich (poor).

Summing up the evidence provided in this section, it is very likely that the high degree of regional income inequality observed today remains or even slowly increases in the future, with only few regions being able to reverse their fortune.

4.4 THE SPATIAL EVOLUTION OF THE REGIONAL INCOME DISTRIBUTION

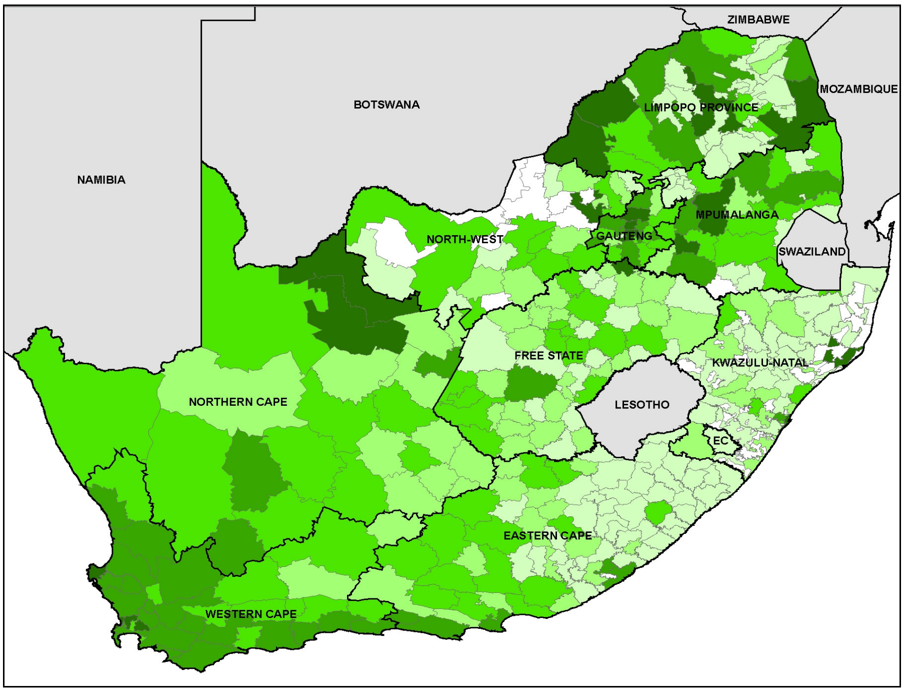
So far the analysis has ignored the spatial context of the data set at hand, treating regions as if being isolated islands having no direct influence on each other's development. Recent theoretical insights from the new economic geography literature (see e.g. Krugman, 1991;

case, 30% of the population lived in the richest regions (with GDP per capita above the national level) in 1996 and this share declined to 27% in 2004. At the same time, 21% of the population lived in the poorest regions (with GDP per capita below 0.17 times the national level) and this share increased to 26% in 2004. (The other three income groups' population share went from (from the lowest to highest income group) 20%, 12% and 17% in 1996 to 18%, 13% and 16% in 2004 respectively.) This clearly indicates that the high degree of regional inequality in South Africa (or its increase during the post-Apartheid period) is not due to the existence of many sparsely populated poor areas. Indeed, if anything, the evolution of regional inequality may even understate the evolution of personal income inequality, as the share of the population in the poorest regions is on the rise.

[110] More specifically it denotes the number of periods it takes for the distribution to move halfway towards its long run steady state.

Puga, 1999 and Baldwin et al., 2001) and the regional science and urban economics literature (Fingleton and Lopez-Bazo, 2006; Hu, 2002; Ciccone and Hall, 1999; Hanson, 1998 and Boarnet, 1998) however view the locational aspect of a region, and the spatial interdependencies that follow from this, as one of a region's central features. Trade between regions, technology and knowledge spillovers, market access and labor (im)mobility are very convincing reasons why the relative location of a region matters for its economic performance. Also empirically, the need to take explicit account of the so-called second nature geography aspect of a spatial data set has become evident in recent years (see Rey and Janikas (2005) and Abreu, de Groot and Florax (2005) for a good overview). Several studies have already made use of spatial econometric techniques that take explicit account of the spatial aspect of a regional data set when looking at the evolution of income per capita. Examples are Rey and Montouri (1999), Le Gallo (2004), Le Gallo and Dall'Erba (2006), Dall'Erba (2005), Fingleton (1999), Fingleton and Lopez-Bazo (2006), Lopez-Bazo et al. (2004), Mossi et al. (2003) and Bosker (2007a), all showing that a region's location vis-à-vis other regions substantially affects its own economic development.

Figure 4.4 Regional GDP per capita as a proportion of the national average in 2004



Notes: Coloring goes from dark, denoting high GDP per capita, to light, denoting low GDP per capita. Namibia, Botswana, Zimbabwe, Mozambique, Swaziland and Lesotho are South Africa's neighboring countries; the other names denote South Africa's main provinces.

In this section we therefore take explicit account of the spatial context of our data set in the Markov chain analysis and look for evidence on the effect that economic development in a region's neighbors has on a region itself. Instead of confining our attention to NEG-based explanations for the interaction between regions as in chapters 1-3, where we estimated a structural equation describing regions' interdependencies, we take a more general approach here that does not specify the exact reason why regions' economic development may be interdependent[111]. Before going into the details and showing the results of these space-incorporating methods, Figure 4.4 above gives a preliminary look at the spatial aspect of the distribution of regional incomes in South Africa. The darker colored a region on the map, the higher GDP per capita compared to the national average. By looking at the map one can immediately see patterns of concentration of economic activity, specifically in Gauteng (with Johannesburg and Pretoria as main cities) and the Western Cape (with Cape Town as its main city), already suggesting some degree of spatial interdependence.

To formally test for spatial autocorrelation, that is the clustering of regions with a (dis)similar realization of a random variable (here upward or downward mobility in the discretized income distribution), in the sample of South African regions, we calculate the Cliff and Ord (1981) BB-statistic for both upward and downward mobility in the income distribution. It checks if, when a region moves up (down) in the discretized distribution, its neighbors are more likely to move up (down) than other regions, indicating positive spatial autocorrelation, or are less likely to do so than other regions, indicating negative spatial autocorrelation. The statistic is calculated as follows:

$$BB = \frac{1}{2}\sum_i \sum_j w_{ij} d_i d_j \qquad (4.6)$$

where $d_i = 1$ if a region has moved up (down) in the discretized income distribution when testing for spatial autocorrelation in upward (downward) mobility and 0 otherwise. w_{ij} measures the 'strength' of the spatial interaction between regions *i* and *j*. Following most other papers using spatial econometric techniques, e.g. Fingleton and Lopez-Bazo (2006), Le Gallo (2004) and Bosker (2006; 2007a), these w_{ij} are chosen to depend on the bilateral distances between the regions in the sample. This reflects the fact that transport costs and also the extent of knowledge spillovers and trade are empirically found to depend on distance (see e.g. Hummels 2001 and Audretsch and Feldman 1996). Instead of distance, other measures, such as for example trade shares or the information from input-output tables, can be argued to more accurately reflect spatial interactions between regions. These are however not readily available at the regional level, making bilateral distance an attractive alternative. Distances are also clearly exogenous giving it an advantage from an econometric point of view over weights constructed on the basis of for example GDP or trade shares (see Anselin 1988). More formally the w_{ij} are constructed as follows:

$$w_{ij} = \begin{cases} D_{ij}^{-1} / \sum_k D_{ik}^{-1} & \text{if} \quad i \neq j \text{ and } D_{ij} < D_{\max} \\ 0 & \text{else} \end{cases} \qquad (4.7)$$

where D_{ij} is the distance between the centroids of two regions and the direct dependence between two regions is limited to regions closer than D_{max}, the lower quartile distance of all bilateral distances between the regions in our sample (= 332 km). D_{ij}^{-1} is chosen as distance

[111] This is usually done in papers using spatial econometric techniques. It comes with the drawback of not being able to clearly link the results back to (preferably one) economic theory, but has the advantage of being able to use econometric techniques that do not suffer from endogeneity problems resulting from the way regions' interdependencies are modelled (e.g. the use of an endogenous measure of regions' interdependencies such as infrastructure or trade shares), or from the inherent endogeneity problems that result from the inclusion of a spatially lagged dependent variable (see e.g. Anselin, 1988).

decay function, a choice quite common in the empirical literature on trade and economic geography[112]. Finally the weights are row-standardized, so that the spatial interaction of a region with another region depends on that other region's relative closeness compared to the other regions in the sample.

Using this constructed measure of spatial interaction between the regions in the sample, the BB-statistic is calculated. Table 4.3 below shows the BB-statistic in case of both upward and downward mobility along with the corresponding 2.5% upper and lower bound critical values, obtained by bootstrapping the empirical distribution of the statistic. When the BB-statistic is smaller (larger) than the 2.5% lower (upper) bound this means that a region is significantly less (more) likely to move up (in case of the BB-up) or down (in case of the BB-down) when its neighbor(s) make a similar move.

Table 4.3 BB-statistics and critical values for upward and downward mobility

	BB-statistic	2.5% critical value	97.5% critical value
UP	0.033	0.095	5.738
DOWN	5.872	0.006	4.673

The results show something very interesting, the upward statistic is significant and smaller than the 2.5% lower bound indicating negative spatial autocorrelation in upward moves, i.e. if your neighbor(s) moves up in the distribution, you are less likely to make a similar move. For the downward statistic the opposite holds, if your neighbor moves down the distribution, you are significantly more likely to do so than a more distant region (significant positive spatial autocorrelation in downward moves). This is quite different from earlier studies looking at e.g. Europe or the USA (see e.g. Le Gallo, 2004; Rey and Montouri, 1999 and Bosker, 2006) who find significant positive spatial autocorrelation in both upward and downward moves or no significant spatial autocorrelation at all. This finding for South Africa suggests the presence of localized growth poles, with some regions (relatively far away from each other) showing strong economic performance reflected in rising per capita income levels, leaving nearby regions behind in relative poverty. Lall and Shalizi (2003) and Ying (2000) also present significant negative spatial autocorrelation in regional growth rates in case of the Brazilian North-East and Chinese provinces respectively, which may be tentative evidence that during the development process of the national economy as a whole, regional disparities are likely to arise which corresponds to the theoretical predictions in Puga and Venables (1996), who show (their focus is at the country level) that regions' economic development may come in waves during the growth process of the national economy as a whole, spreading from a few, high wage, regions to the rest of the country. When looking at the maps in Figures 4.4 and 4.7 this finding shows itself in the high GDP per capita levels in regions surrounding for example Johannesburg, Cape Town and Durban (in Kwazulu Natal).

The finding of significant spatial autocorrelation also implies that it is not correct to view the regions in the sample as isolated islands (as in the previous section) and calls for the

[112] Taking another value for D_{max} and/or a different distance decay function (both arbitrary choices) does not alter any of the results qualitatively. Results are available upon request.

need to use econometric techniques that take explicit note of the spatial dimension of our data. In the next two sub-sections the spatial dimension of the data set is incorporated in the Markov chain analysis in two different ways:

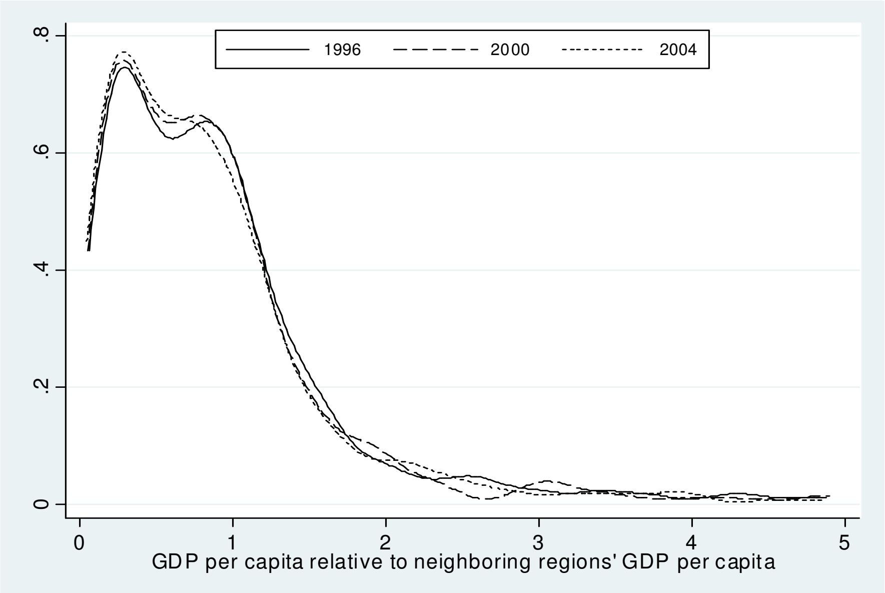
1. Following the evolution of a regionally conditioned instead of the SA-wide conditioned relative income distribution (see Quah, 1996b; Mossi et al., 2003; Bosker, 2006 and Le Gallo 2004).
2. Estimating spatial Markov Chains (see Le Gallo, 2004; Rey 2001; Mossi et al. 2003 and Bosker 2006).

4.4.1 Regionally conditioned Markov chain analysis

Instead of looking at the evolution of regions' GDP per capita relative to the South African level of GDP per capita, one can look at the evolution of regions' GDP per capita relative to 'regional GDP per capita', i.e. a distance weighted sum of neighboring regions' GDP per capita[113]. This was first suggested by Quah (1996b) and can give interesting insights in the relevance of a region's relative location for its economic performance. More specifically the regionally conditioned income distributions can be interpreted as the part of GDP per capita that cannot be explained by location-specific factors.

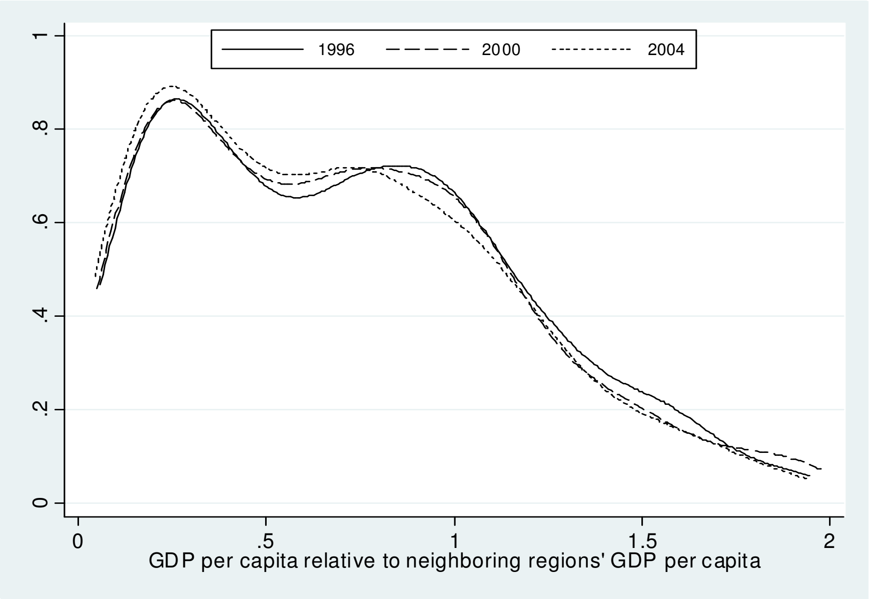
Figures 4.5 and 4.6 show this regionally conditioned distribution for South African regions, both as a whole and zoomed in on the left tail respectively[114].

Figure 4.5 The regionally conditioned income distribution



[113] The same weights as in (11) are used to construct this weighted sum of neighboring regions' GDP per capita.
[114] The result of zooming in on the right is not shown here as this does not provide any additional insights over Figure 4.5.

Figure 4.6 The regionally conditioned income distribution – zoomed in on the left



Interestingly the distribution is less skewed to the left than the South African-relative GDP per capita distribution, most clearly seen by comparing the zoomed in left part of the distribution. This indicates that a substantial number of regions have income levels similar to that of their neighbors. However, note also that a small number of regions earn much more than their neighbors; the pattern in the right tail is quite similar to that of the South African-relative distribution.

Merely comparing the two differently conditioned distributions does however only give suggestive evidence about whether the regions that are rich compared to their neighbors are also the regions having a level of per capita income that is higher than the overall South African level of per capita income. As in the case of Markov chain transition probabilities, this can be more formally quantified by estimating a transition matrix. Not a matrix describing the evolution over time in this case, but a matrix that tells what the probability of a region is to be rich compared to its nearby neighbors, given its income level compared to the South African total (see Quah, 1996b). Table 4.4 shows this matrix (with the regional distribution cut-off points set in the same way as for the South African-relative distribution):

Table 4.4 The South Africa-relative vs. the regionally conditioned income distribution

x REG GDP/cap	< 0.27	0.27 < < 0.59	0.59 < < 0.9	0.9 < < 1.3	> 1.3
< 0.17	0.798	0.202	0	0	0
0.17 < < 0.36	0.172	0.491	0.287	0.050	0
0.36 < < 0.575	0	0.508	0.349	0	0.144
0.575 < < 1.05	0	0.070	0.298	0.455	0.177
> 1.05	0	0	0.061	0.234	0.706

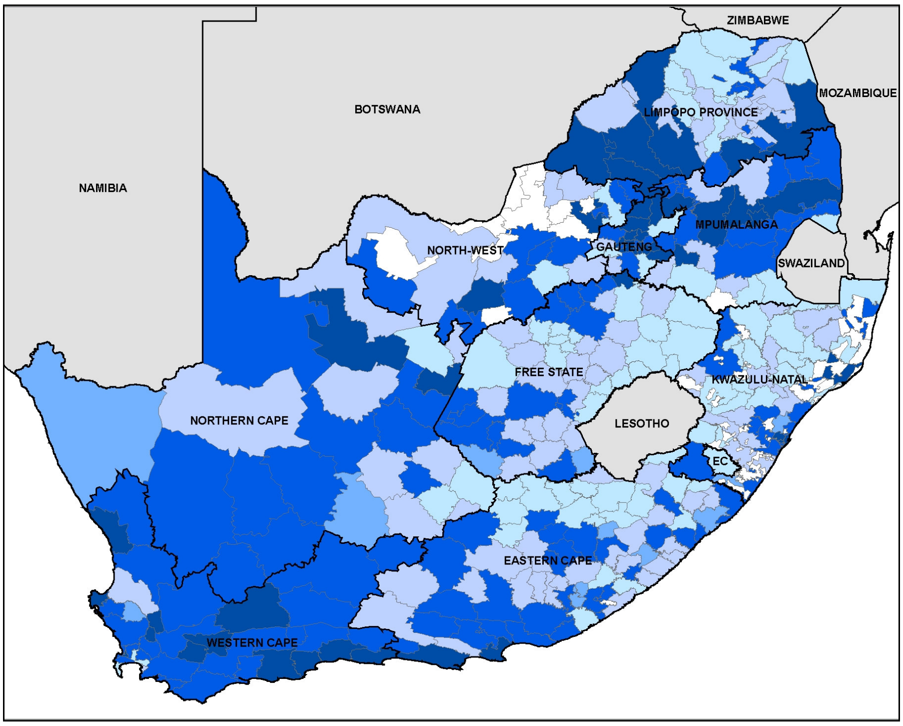
Notes: All probabilities are significant at least at the 10% level. Each of the numbers in the table shows the probability that a region in a particular SA-relative income group is also located in a particulate regional-relative income group. For example, a region in the lowest SA-relative income group has a probability of 80% to also be in the lowest regional-relative income group, whereas a region in the 2nd lowest SA-relative income group has a probability of 20% to be in the lowest regional-relative income group.

Several interesting things can be said about these probabilities. First, only 6% of the richest South African-relative regions have income levels below 90% that of their neighbors. Second, all of the poorest South African-relative regions have income levels below 60% that of their neighbors.

This corroborates the suggestive evidence provided by Figures 4.5 and 4.6 of regions with high levels of GDP per capita being scattered around the country surrounded by relatively poorer areas. Also interesting is the fact that 14% of the regions with income levels between 0.36 and 0.575 times the national GDP per capita level have income levels that are more than 1.3 times that of their neighbors giving some evidence that within the clusters of poorer regions there are also some that stand out as being 'less poor'[115].

Combining this with the evidence from the calculation of the BB-statistics in the previous section, strengthens the notion of a regional economy characterized by localized regional growth poles, being those regions that are currently already enjoying a higher GDP per capita level than their (immediate) neighbors.

Figure 4.7 Regional GDP per capita growth between 1996 and 2004



Notes: Coloring goes from dark, denoting high GDP per capita growth, to light, denoting low GDP per capita growth. Namibia, Botswana, Zimbabwe, Mozambique, Swaziland and Lesotho are South Africa's neighboring countries; the other names denote South Africa's main provinces.

Figure 4.7 shows that such localized growth poles (the darker colored a region, the higher GDP per capita growth), can be found in e.g. Gauteng around Johannesburg and Pretoria, in the Western Cape around Cape Town and around the port cities of Durban and Richard's Bay in KwaZulu Natal. There are also some inland regions in e.g. Limpopo province, Northwest

[115] Again, estimating the corresponding stochastic kernel, shown in Figure 4.A2 of Appendix 4.A, corroborates these findings.

Province and the Northern Cape that show relatively high growth rates, mainly due to the presence of natural resources.

The existence of such local growth poles is consistent with theoretical predictions from NEG theory (Krugman, 1991; Fujita et al., 1999; Puga, 1999) arguing that agglomerations, once established, continue to attract people and firms alike in search of higher wages or higher profits, eventually resulting in a strong core-periphery pattern. Agglomerations offer workers and firms higher real wage and profit prospects respectively due to the positive externalities associated with agglomerations such as a better matching of jobs and workers, the presence of specialized inputs, knowledge spillovers and better market access (see e.g. Rosenthal and Strange, 2001; Ciccone and Hall, 1999 and LaFountain, 2005).

4.4.2 Spatial Markov chain analysis

The results in the previous sub-section provide clear evidence on the regional spread of economic activity in South Africa (i.e. its clustered nature). They do not provide insights into the relevance of a region's spatial setting for the evolution of its GDP per capita level over time. The new economic geography (Puga, 1999; Fujita et al., 1999) and regional science (Fingleton and Lopez-Bazo, 2006; Ciccone and Hall, 1999; Ciccone, 2002) literature suggest however that a region's spatial setting has a substantial impact on its subsequent economic development. A high GDP level in neighboring regions is mostly expected to have a positive effect, resulting in more trade due to increased demand for a region's products (Krugman, 1991), a higher degree of knowledge spillovers (Baldwin et al., 2001; Fingleton and Lopez-Bazo, 2006) and localized geographical externalities (Ciccone and Hall, 1999). However it could also have a negative impact on the evolution of a region's own income level as firms (Blonigen et al., 2007 provide evidence for this in case of US outward FDI in Europe) and workers (Crozet, 2004 and Pons et al., 2007 provide evidence on this) choose to locate in the richer (more agglomerated) neighboring region, strengthening the core-periphery pattern (Krugman, 1991; Baldwin et al., 2001) in the process.

In order to say something about this in case of the South African regional economy, this sub-section estimates so-called spatial Markov chains as introduced by Rey (2001). These spatial Markov chains estimate the dynamics of South Africa's regional income distribution conditional on the distance weighted GDP per capita in neighboring regions. It gives clear insights into the influence of high/low GDP levels in a region's neighboring regions (frequently referred to as a region's market potential) on the evolution of GDP per capita in that region itself. To estimate these spatial Markov chains one first groups all regions according to their neighbor-relative GDP per capita[116]. Next, given the ordering based on neighbor-relative GDP per capita, one looks at the evolution of South African-relative GDP per capita for only those regions within a certain spatial income group. This results in five (given the five neighbor-relative income groups) 1-year transition matrices that are of

[116] Constructed in the same manner as in the previous section. As argued in Bosker (2006) conditioning on GDP per capita relative to a region's neighbors gives some more interesting results regarding the relevance of the economic condition(s) in a region's immediate surroundings, than conditioning on the absolute GDP level of a region's neighbors as done by Rey (2001).

dimension 5x5 (given the five South African-relative income groups): Tables 4.5a-4.5e below.

Tables 4.5a - 4.5e Spatial Markov Chains

a. Neighbor-relative income group 1 (GDPcap < 0.27 that of its neighbors)

x SA GDP/cap	< 0.17	0.17 < < 0.36	0.36 < < 0.575	0.575 < < 1.05	> 1.05
< 0.17	0.998	_0.002*_	0	0	0
0.17 < < 0.36	0.087	0.913	0	0	0
0.36 < < 0.575	.	.	.	.	.
0.575 < < 1.05	.	.	.	.	.
> 1.05	.	.	.	.	.

Notes: Italic transition probabilities marked with a * are not significant at the 10% level.

b. Neighbor-relative income group 2 (0.27 < GDPcap < 0.59 that of its neighbors)

x SA GDP/cap	< 0.17	0.17 < < 0.36	0.36 < < 0.575	0.575 < < 1.05	> 1.05
< 0.17	0.991	_0.009*_	0	0	0
0.17 < < 0.36	0.025	0.971	_0.004*_	0	0
0.36 < < 0.575	0	0.068	0.917	_0.015*_	0
0.575 < < 1.05	0	0	0	1	0
> 1.05	.	.	.	.	.

Notes: Italic transition probabilities marked with a * are not significant at the 10% level.

c. Neighbor-relative income group 3 (0.59 < GDPcap < 0.9 that of its neighbors)

x SA GDP/cap	< 0.17	0.17 < < 0.36	0.36 < < 0.575	0.575 < < 1.05	> 1.05
< 0.17	.	.	.	.	.
0.17 < < 0.36	0	0.981	0.019	0	0
0.36 < < 0.575	0	0.044	0.931	0.025	0
0.575 < < 1.05	0	0	0.096	0.886	0.018
> 1.05	0	0	0	0.091	0.909

Notes: Italic transition probabilities marked with a * are not significant at the 10% level.

d. Neighbor-relative income group 4 (0.9 < GDPcap < 1.3 that of its neighbors)

x SA GDP/cap	< 0.17	0.17 < < 0.36	0.36 < < 0.575	0.575 < < 1.05	> 1.05
< 0.17	.	.	.	.	.
0.17 < < 0.36	0	0.926	_0.074*_	0	0
0.36 < < 0.575	0	0.061	0.914	0.025	0
0.575 < < 1.05	0	0	0.019	0.973	_0.008*_
> 1.05	0	0	0	0.047	0.953

Notes: Italic transition probabilities marked with a * are not significant at the 10% level.

e. Neighbor-relative income group 5 (GDPcap > 1.3 that of its neighbors)

x SA GDP/cap	< 0.17	0.17 < < 0.36	0.36 < < 0.575	0.575 < < 1.05	> 1.05
< 0.17	.	.	.	.	.
0.17 < < 0.36	.	.	.	.	.
0.36 < < 0.575	0	0	0.988	*0.013*	0
0.575 < < 1.05	0	0	0.059	0.901	0.040
> 1.05	0	0	0	0.013	0.987

Notes: Italic transition probabilities marked with a * are not significant at the 10% level.

The estimated transition probabilities show several interesting things. First, for regions with a GDP per capita level of only 60% or lower that of their neighboring regions (Tables 4.5a and 4.5b), the probability of moving up in the South African-relative income distribution is not significantly different from zero at the 10% level. On the other hand all regions that are richer than the South African-average have a lower probability of moving to a lower income group the richer they are compared to their neighboring regions (compare the last row of Table 4.5a-4.5e with each other). In general, it is mostly the case that the richer (poorer) a region compared to its neighbors, the higher (lower) probability it has to move up in the South African-relative income distribution. The result of the calculation of a likelihood ratio test for spatial dependence as suggested by Bickenbach and Bode (2003), i.e. $\chi^2(23)$ = 70.18 with corresponding p-value: 0.00, shows that these differences are significant, strengthening the notion that a region's transition probability significantly differs depending on its neighboring regions' income levels (or market potential).

This again confirms the earlier evidence given by the regionally conditioned distribution in Figures 4.5 and 4.6, Table 4.4 and the BB-statistics, that the South African economy is characterized by regional growth poles. Several rich regions, scattered around the country, absorb economic activity from their neighbors, leaving them behind with relatively lower income levels. This supports the theoretical prediction made by theories stressing the advantage of agglomerations (Ciccone and Hall, 1999; Rosenthal and Strange, 2001 and Ciccone, 2002) in the process of economic development, with the more prosperous regions attracting people and economic activity alike from the periphery establishing a strong core-periphery pattern that as predicted by new economic geography theories is not so easy to break by active government policies (Baldwin et al. 2003).

4.5 EXPLAINING THE OBSERVED REGIONAL INCOME DISTRIBUTION

As already mentioned in section 4.3 the results obtained from applying (spatial) Markov Chain techniques, although giving a very clear picture of the post-Apartheid evolution of South African regional incomes, are unable to (statistically) say anything about the important factors that are driving the observed evolution. Besides looking at theory for an explanation, this section addresses this is issue by comparing some characteristics of the fastest and slowest growing regions over the period 1996 to 2004. These characteristics are closely linked to the factors that Nel (2002) regards as the main influences on the location of economic activity in South Africa since democratization in 1994. They also coincide with some of the determinants of South African regional income growth identified in previous empirical work (Naudé and Krugell, 2003 and 2006).

The characteristics of the 25 fastest and 25 slowest growing regions in South Africa over the period 1996 to 2004 are shown in Table 4.A1 and 4.A2 in Appendix 4.A[117]. This presents a diverse profile. A number of the fast growing places form part of the large metropolitan areas, that are characterized by high literacy rates and a large share of value added that is exported. Three of the other fast growers are located further away from the large

[117] A detailed description of the fast and slow growing places can be found in Krugell (2005).

hubs, but have benefited from growth of specific industries, specifically platinum mining. Then there are the smaller areas where the economies have grown fast, particularly those in the Western Cape. These places typically have smaller populations and lower population growth rates; they tend to export less, but are attractive tourist destinations. A closer look at a sectoral breakdown of gross value added (GVA) shows that these regions have particularly experienced growth in construction, retail trade, and services sectors such as postal and telecommunication services, finance and insurance, and real estate activities. The remaining fast growing places are typically quite small and have grown fast off a small base, with growth often ascribable to a single industry.

Turning to the characteristics of the slowest growing regions in Table 4.A2 shows that these regions generally have smaller populations, are less urbanized, have lower literacy rates and higher poverty rates, and little is exported from these regions. There are broadly two main groups to distinguish. First there are the regions that have been exposed to the declining fortunes of the South African gold and coal mining industry. Second, there are the regions where manufacturing contracted. Specifically, this occurred in the so-called decentralized growth points that received industrial promotion subsidies under the Apartheid government and were hard hit when these were suspended in the post-Apartheid period.

Overall, the profile of the fast and slow growers adds to the notion of localized growth poles, specifically in Gauteng and the Western Cape (Nel, 2002), that, by virtue of their agglomeration benefits (see Ciccone and Hall, 1999; Rosenthal and Strange, 2001; Fujita, et al., 1999 and Puga and Venables, 1996), are leaving the peripheral regions behind in (literal) poverty. The larger metropolitan areas, offering a better economic climate to both workers in terms of both income and unemployment, and to firms, in terms of human capital and export facilities, are prospering. The opening up of South Africa to the world economy and the increased labor mobility that followed the lifting of the severe mobility restriction imposed on the largest part of the South African population, that both resulted from the end of the Apartheid era, seem to have exacerbated this process of increased spatial disparities even further; exporting regions are performing well in line with predictions by e.g. Hu (2002), Mansori (2003) and Gianetti (2002) and people are migrating towards regions offering them better economic prospects. These conclusions also corroborate earlier evidence provided by Naudé and Krugell (2003) who, by using more standard growth regressions, found that the faster growing regions are those with a better educated population, better market access (measured by the distance to Johannesburg) and with a greater share of exports in total output. Interestingly, the post-Apartheid regional growth process seems to largely confirm Rogerson's (1991) prediction of a return to a spatial economy initially shaped by trade, minerals and energy, and dominated by the large metropolitan centers.

4.6 CONCLUSIONS

The demise of Apartheid in 1994 introduced an era of opportunity for the South African economy. It marked the end of years of international sanctions, opening up the economy to the rest of the world, and it provided (economic) freedom to its previously disadvantaged black population. Since the end of Apartheid the South African national economy has shown strong performance but this masks significant spatial inequality at the regional level. Using (spatial) Markov chain techniques, this chapter provides clear evidence of a heavily diverging regional income distribution. Relatively poor regions are likely to remain poor or become even poorer and the richest regions will maintain or increase their lead in terms of income levels. Explicitly taking account of space furthermore shows that these richest regions are scattered around the country, acting as local growth poles absorbing economic activity from the nearby periphery. It is these few core regions that have driven recent economic growth in South Africa while ever more places on the periphery are producing less and less. The result is a strong core-periphery pattern in line with predictions from the new economic geography literature (see e.g. Fujita et al., 1999). Increased labor mobility, location, trade, education and the variable fortune of the gold mining industry seem to be important determinants of the observed evolution.

When considering both the government's past record in regional development (Nel, 2002) and some of the predictions made by new economic geography theory (Fujita et al., 1999), this clearly sets a daunting, if not impossible challenge, for local governments in the periphery that have to fulfill their constitutional responsibility to economically develop their areas. Theory suggests that public investments in transportation, communication and education (Boarnet, 1998; Démurger, 2001; Mansori, 2003; Lall, 2007; Ravallion, 1996; Hu, 2002) could provide some of the answers.

APPENDIX 4.A

Table 4.A1 Fast growers, 1996-2004

	Annual growth rate 96-04	GDP-R per capita	Total population	Population growth rate	Poverty rate	Functional literacy	Urbanization rate	Unemployment rate	% Exports
Phalaborwa	8.2%	141,580	37,118	1.2%	15.5%	71.0%	53.4%	3.6%	14.0%
Volksrust	8.0%	20,170	36,854	1.7%	53.6%	60.8%	85.2%	27.6%	1.2%
Lower Umfolozi	7.1%	48,166	286,902	1.5%	41.9%	67.2%	30.0%	35.8%	78.3%
Randburg	6.7%	117,434	445,772	1.6%	14.0%	89.5%	93.3%	19.0%	21.8%
Warmbad	6.4%	20,278	58,941	1.4%	33.1%	68.1%	64.8%	22.3%	5.7%
Rustenburg	6.1%	72,116	446,234	1.6%	26.7%	69.1%	46.0%	27.5%	30.6%
George	5.5%	32,499	130,990	1.5%	17.6%	78.8%	89.8%	20.6%	1.4%
Knysna	5.3%	31,398	71,905	1.7%	18.5%	80.5%	88.9%	19.9%	5.9%
Waterberg	4.9%	15,031	67,255	1.4%	43.7%	61.2%	50.6%	9.9%	0.8%
Pretoria	4.8%	102,671	825,868	1.1%	11.7%	92.9%	89.5%	14.6%	51.3%
Wellington	4.7%	27,054	49,660	1.0%	15.6%	80.0%	81.3%	14.3%	28.1%
Sekhukhuneland	4.7%	3,403	473,456	1.3%	56.4%	56.5%	4.2%	60.1%	0.0%
Boksburg	4.7%	35,961	324,282	1.8%	19.6%	88.7%	98.1%	31.0%	17.1%
Kempton Park	4.6%	39,238	551,602	2.0%	29.8%	85.6%	97.6%	42.7%	18.9%
Tulbagh	4.5%	16,714	33,011	1.0%	22.1%	68.4%	57.9%	18.6%	5.0%
Cullinan	4.4%	8,756	97,601	1.5%	35.9%	80.6%	89.4%	39.8%	2.6%
Potgietersrus	4.4%	29,532	58,867	1.3%	35.9%	66.6%	62.1%	15.8%	0.2%
Nigel	4.4%	18,456	130,017	2.0%	36.1%	78.9%	94.5%	42.6%	7.8%
Laingsburg	4.3%	14,723	6,298	0.7%	30.5%	57.8%	60.8%	15.7%	0.0%
Sasolburg	4.0%	93,625	116,186	0.8%	28.4%	79.2%	89.1%	33.2%	2.6%
Montagu	3.9%	20,140	26,156	1.3%	27.6%	70.1%	75.3%	28.0%	15.8%
Alberton	3.9%	19,740	503,046	2.0%	28.6%	85.6%	100.0%	44.1%	7.6%
Mosselbay	3.9%	28,242	65,955	1.3%	24.6%	81.1%	88.5%	30.5%	3.9%
Vredenburg	3.8%	44,262	51,638	1.0%	9.7%	80.2%	97.4%	17.1%	88.0%
Pinetown	3.8%	38,117	534,444	1.6%	29.0%	83.9%	99.9%	34.0%	7.7%
Average (top 25)	**5.1%**	**41,572**	**217,202**	**1.4%**	**28.2%**	**75.3%**	**75.5%**	**26.7%**	**16.7%**

Table 4.A2 Slow growers, 1996-2004

	Annual growth rate 96-04	GDP-R per capita	Total population	Population growth rate	Poverty rate	Functional literacy	Urbanization rate	Unemployment rate	% Exports
Dannhauser	-7.7%	2,876	84,137	1.0%	90.4%	65.7%	5.6%	68.4%	0.2%
Theunissen	-6.7%	15,202	41,278	0.3%	71.2%	59.5%	69.7%	32.6%	0.6%
Odendaalsrus	-5.2%	7,420	107,252	0.9%	72.7%	69.8%	92.1%	52.0%	0.2%
Virginia	-4.5%	11,336	87,064	0.4%	49.4%	71.7%	74.5%	39.0%	22.1%
Welkom	-3.6%	16,613	294,562	1.1%	35.1%	73.3%	89.4%	35.3%	1.5%
Hennenman	-3.3%	7,514	30,742	1.1%	53.9%	68.1%	86.8%	50.0%	41.3%
Vryheid	-2.9%	8,394	101,760	1.1%	62.6%	63.9%	43.8%	45.7%	0.1%
Oberholzer	-2.8%	27,363	202,120	1.6%	30.8%	76.0%	98.2%	22.7%	0.1%
Klerksdorp	-2.4%	18,958	382,898	0.9%	56.6%	70.1%	88.2%	42.2%	0.6%
Komga	-2.3%	5,357	15,702	0.9%	78.6%	46.7%	48.1%	38.0%	0.1%
Mtunzini	-2.2%	5,665	232,336	1.4%	66.7%	69.5%	26.9%	55.6%	8.9%
Hlabisa	-2.1%	2,614	233,953	1.6%	84.4%	53.3%	5.6%	58.2%	0.6%
Utrecht	-2.0%	6,947	29,329	1.2%	77.7%	46.4%	16.3%	45.7%	1.1%
Aliwal North	-2.0%	12,781	32,142	1.1%	67.8%	64.9%	86.7%	41.5%	0.0%
Harrismith	-1.9%	8,465	70,027	1.0%	68.0%	63.4%	78.5%	42.9%	9.0%
Maclear	-1.9%	4,835	21,404	0.9%	82.9%	50.4%	72.6%	37.8%	0.0%
Ficksburg	-1.8%	7,941	53,516	1.2%	73.1%	67.4%	0.0%	25.3%	1.1%
Westonaria	-1.7%	11,027	193,723	1.6%	29.1%	72.3%	97.8%	29.9%	0.8%
Babanango	-1.7%	686	47,132	1.5%	84.8%	53.1%	8.8%	83.7%	0.0%
Estcourt	-1.7%	3,768	186,681	1.5%	66.8%	66.4%	13.3%	62.7%	2.8%
Bochum	-1.6%	2,654	181,339	1.2%	89.0%	53.3%	0.0%	70.6%	0.0%
Reitz	-1.6%	9,196	31,651	0.6%	74.5%	55.5%	58.4%	27.7%	0.4%
Mahlabathini	-1.6%	3,229	169,788	1.4%	88.6%	55.1%	11.5%	75.5%	0.0%
Wodehouse	-1.5%	6,175	14,595	0.8%	88.3%	50.2%	60.6%	39.2%	0.0%
Thabamoopo	-1.5%	3,583	414,501	1.6%	74.3%	72.1%	14.1%	58.5%	0.0%
Average (bottom 25)	**-2.7%**	**8,424**	**130,385**	**1.1%**	**68.7%**	**62.3%**	**50.0%**	**47.2%**	**3.7%**

Figure 4.A1 Continuous 1-year Markov Chain (stochastic kernel) and corresponding contour plot

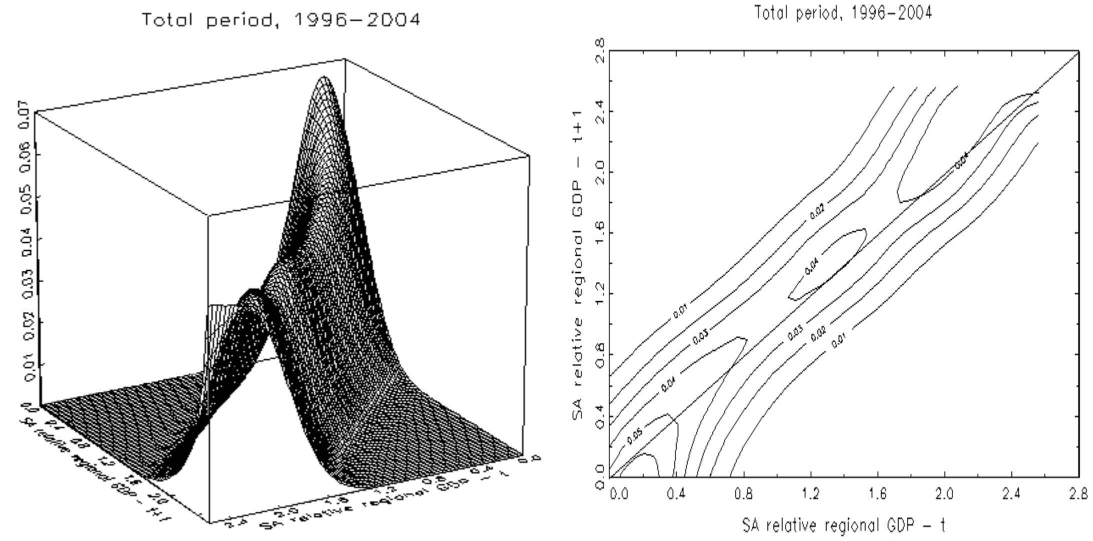
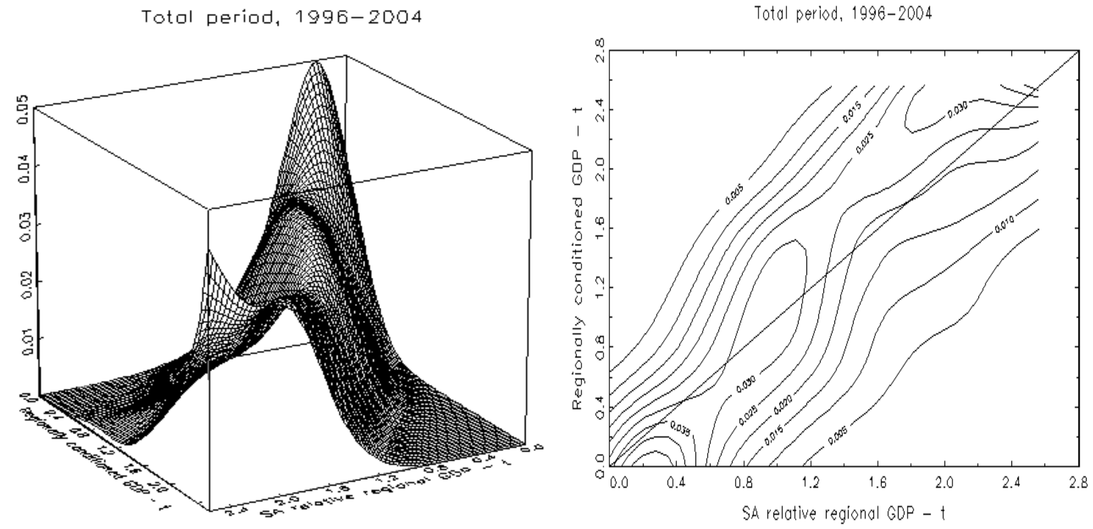


Figure 4.A2 Stochastic kernel South Africa vs. regionally conditioned income distribution and corresponding contour plot



Chapter 5

A century of shocks: the evolution of the German city size distribution 1925-1999[118]

5.1 INTRODUCTION

City size distributions and the underlying city size dynamics have received considerable attention in the urban economic literature in recent years. Empirical studies have in particular produced evidence with respect to three features of city size distributions. First, city size distributions are found to be remarkably stable over time. Second, the hierarchy of the individual cities making up these distributions is also often rather stable, which suggests proportionate city growth, see for example Eaton and Eckstein (1997) and Black and Henderson (2003). The third stylized fact is that city size distributions are very well approximated by a Power Law in the upper tail of the distribution. A special case of which is better known as Zipf's law and has been found to hold for various countries at various points in time, see e.g. Soo (2005) and Nitsch (2005).

These empirical regularities have stimulated the development of city growth models that can explain these features of city size distributions in a coherent economic framework. Modern theories (e.g. Eeckhout, 2004; Rossi-Hansberg and Wright, 2005; Cordoba, 2008) try to explain the evolution of city size distributions in a way that is consistent with the empirical results. Either a stable city size distribution adhering to Zipf's law in the upper tail follows directly from the model (Gabaix, 1999; Eeckhout, 2004), or it is one of the possible outcomes of the model (Rossi-Hansberg and Wright, 2005). These models have benefited substantially from the work of Gabaix (1999) who, building on earlier work by Simon (1955), showed that a stable city size distribution adhering to Zipf's Law in the upper tail naturally results if the individual city size growth process adheres to what is essentially Gibrat's Law of proportional effect, i.e. a city's expected growth rate and the variance of that growth rate are independent of its initial size (also referred to as proportionate city growth). The whole city size distribution would in that case be lognormal (see Eeckhout 2004)[119].

Recent empirical papers on the evolution of the city size distribution focus exclusively on the US experience, e.g. Black and Henderson (2003), Overman and Ioannides (2001), Ioannides and Overman (2003, 2004), Dobkins and Ioannides (2000, 2001), or Eeckhout (2004). Besides some simple Zipf studies that do not look at distributional dynamics nor at evidence for Gibrat's Law of proportional effect, the only two papers we know of that offer a

[118] This chapter is an adapted version of Bosker, Brakman, Garretsen and Schramm (2007c) that is forthcoming in Regional Science and Urban Economics. Also, I thank Peter Koudijs for his excellent research assistance.

[119] The fact that proportionate city growth gives rise to a lognormal city size distribution is itself also frequently referred to as Gibrat's Law (see e.g. Sutton, 1997 and Eeckhout, 2004). To avoid any misunderstanding, our use of Gibrat's Law should hereafter be read as referring to Gibrat's Law of proportional effect, i.e. proportionate city growth (see section 5.4.1 for more detail).

thorough look at the distributional dynamics of city size distributions for other countries than the USA are Eaton and Eckstein (1997) and Anderson and Ge (2005). The former provides evidence for France and Japan confirming the notion of a stable city size distribution, while the latter shows that in case of China the city size distribution has been affected in a predictable way by government policies.

The contributions of this chapter are the following. First, we examine the evolution of the city size distribution for West Germany (from now on referred to as Germany). The interest in the German city size distribution can be dated back to as early as 1913 when Auerbach (1913), as one of the first, noted that the city size distribution could be approximated by a power law. For our empirical analysis we have constructed a unique data set of annual city population data for 62 of the largest cities in Germany over the period 1925-1999. This data set allows us to describe the evolution of the German city size distribution quite accurately. Second, the use of German data provides a specific empirical view on the evolution of the city size distribution, namely it offers a (unique) look at the effect of large shocks to the urban system. In the time period under consideration German cities were subject to a number of large 'quasi-natural experiments' namely the heavy destruction of cities during WWII and the separation from and subsequent reunification with East Germany. Our data set allows us to look at the impact of these 'quasi-natural experiments' in a much more dynamic fashion than previous research on the same topic (Davis and Weinstein, 2002; Bosker et al., 2007a; Redding and Sturm, 2005, or Redding, Sturm, and Wolf, 2007). Third, our annual data set allows us to perform unit root tests for each individual city in order to find evidence on Gibrat's Law of proportional effect[120]. This chapter is among the first to provide (panel) unit root tests for cities that, given our extensive data set, arguably suffer less from low power problems that beset unit root tests based on a limited number of observations, while at the same time adequately controlling for the short-run dynamics in city size.

The results of our analysis of the German urban system can also help to distinguish between some of the proposed urban economic theories that try to explain the shape and evolution of city size distributions. Following Davis and Weinstein (2002), these theories can be grouped into three broad categories, i.e. increasing returns to scale, random growth and locational fundamentals.[121] The models in all three categories predict a stable city size distribution in equilibrium; the reaction to shocks is however quite different. Models exhibiting increasing returns to scale, of which the new economic geography theory is a prominent example, can give rise to a stable distribution which is sensitive to shocks and

[120] As already mentioned in the previous footnote, there seems to be an inconsistency between the use of Gibrat's Law in the modern urban economics literature and its use in the statistical process theory. In the urban economics literature, Gibrat's Law is referred to as a process of proportionate city growth or a random walk in city sizes. In the statistical process theory, the log normality of the city size distribution is referred to as Gibrat's Law and a random walk in city sizes is merely a pre-condition for Gibrat's Law but not Gibrat's Law itself. To avoid any confusion, and as we do not feel ourselves in a position to make a final judgement on this discrepancy, we stick to the definition of Gibrat's Law used in the urban economics literature that serves as a benchmark for our own analysis (see also section 5.4.1), and leave a more in-depth analysis of this issue for future work.

[121] This is of course a rather extreme classification of urban growth models, for instance because many models do not exactly coincide with one of the three categories. However for our present purposes, where we only want to highlight that different urban theories may have quite different implications as to the evolution of urban systems, this three-way classification is quite useful in our view.

which does not necessarily adhere to Zipf's law (see also Chapter 12 of Fujita et al., 1999; Gabaix and Ioannides, 2004, or Brakman, Garretsen, van Marrewijk en van den Berg, 1999). As a result a large shock has the potential to (radically) change both the shape of, as well as individual cities' rank within the city size distribution. Models falling under the random growth category, e.g. Gabaix (1999) or Cordoba (2008), predict that shocks have a permanent effect on city sizes, but given that these shocks are distributed randomly over cities and mean- and variance-independent of city size, they will in the limit result in a city size distribution that adheres to Zipf's law in the upper tail. The effect of a large shock thus has no effect on the limiting city size distribution; it can however have a permanent impact on the relative position of cities within the distribution. Finally, the locational fundamentals approach suggests that the observed city size distribution is the result of fixed underlying locational fundamentals (first nature geography). A large shock will now result in both the city size distribution as a whole and the relative position of cities within this distribution returning to their pre-shock state[122]. Given the three categories' different reaction to large shocks, the 'quasi-natural experiments' that the German urban system was subjected to, provide a way to try and distinguish between these competing theoretical views of city size evolution (see also Head and Mayer (2004), who consider shock sensitivity one of the five 'testable' propositions following from new economic geography theory).

Our first main finding is that the German city size distribution is permanently affected by the World War II shock, more so than by any other shocks. Cities that have been hit relatively hard due to the substantial bombings and the subsequent allied invasion do not recover the loss in relative size. After the war, the city size distribution does not revert to its pre-WWII level, but shifts to one characterized by a more even distribution of population over the cities in the sample. Compared to the impact of WWII, the separation from and later reunion with East Germany has had a much less severe impact on relative city sizes. Our second finding is that, once corrected for the heavy destruction during WWII, (panel) unit root tests that are used to test for the validity of proportionate city growth reject Gibrat's Law for about 75% of all cities. Also we find strong evidence that the locational fundamentals approach does not seem to explain the evolution of Germany's urban system, which is in contrast to the findings of Davis and Weinstein (2002, 2004) for Japan. Overall the evidence does seem to comply best with urban theories exhibiting increasing returns to scale.

5.2 DATA

In constructing our data set, we first had to choose which cities to include in our sample. We choose to include those West-German cities in our data set that either had a population of over 50,000 inhabitants before the beginning of WWII or cities that were over the sample period classified as Großstädte, cities with a population of at least 100,000 people. Cities are defined on a city-proper or administrative basis in Germany. Adjustments to the administrative boundaries (and hence size) of some cities were made during our sample period due to metropolitan developments but we were able to correct for the most significant of these

[122] Note that this assumes that the shock does not change the underlying locational fundamentals themselves.

changes (see Appendix 5.A for more information on the city selection criteria and the boundary adjustments made). This selection process initially resulted in 81 cities, the same sample of West-German cities as used in Brakman et al. (2004a) and Bosker, Brakman, Garretsen and Schramm (2007a). In those two studies the only requirement to be included in the analysis was that for each city, population data had to be available for the years 1939, 1946 and 1963 (and possibly 1933). In the present study, however, we only include cities for which we have annual population data for each year in the 1925-1999 period. In total 16 of the 81 West-German cities in our sample did not meet this requirement and for 3 cities we were unable to adequately correct for changes in their administrative border. We are therefore left with a data set that consists of 62 West-German cities over the period (Appendix 5.A lists the cities included in our sample). The 19 cities that were dropped from our analysis were mostly relatively small. Since our main point of interest will be the upper tail of the city size distribution it can be argued that their exclusion does not matter too much for our analysis.

As to the decision to focus on West-German cities only, there are two main reasons to exclude East-German cities, that is to say cities that were part of the German Democratic Republic (GDR). The first reason is simply data availability. For most of the cities concerned there are too many missing observations during the GDR-period. The second and more fundamental reason is that (see Brakman et al., 2004a) cities in the GDR were not part of the kind of urban system that lies at the heart of all urban location theories where economic agents are free to choose their location. On the contrary, in the centrally planned economy of the GDR, firms and workers were not free to move between cities. In our view this has the implication that any testing of the stability or any other feature of the pan-German city-size distribution is not very useful during our sample period. Obviously, this does not imply that we are not concerned with the split between West and East Germany or the subsequent reunification, but we will deal with this from the perspective of West-German cities only. Finally, with respect to the length of the sample period one could argue that it may be worthwhile to include population data for the pre-1925 period as well so as to be able to deal with for instance the WWI shock. For some cities in our sample we have population data that go as far back as 1871, but the number of cities with annual pre-1925 data is rather small so we decided to take 1925 as our cut-off date.

5.3 THE EVOLUTION OF THE CITY SIZE DISTRIBUTION

We start our analysis by giving a description of the evolution of the West-German urban system. Table 5.1 below shows that during our sample period total population increased by about 70% from about 39 million people in 1925 to 67 million in 1999. During the same period, the share of Germany’s population living in one of our sample cities declined by about 32% (or 13 ppt) suggesting a process of suburbanization over the sample period. The average city size in our sample increased by 16% from 258329 in 1925 to 300295 inhabitants in 1999 but it has been quite stable from 1955 onwards. Comparing the development of mean city size to that of the median city size, which increased by 42%, the impression comes to the fore that smaller cities grew faster than the larger cities in our sample, hinting at a transition towards a more equally spread population over the cities in our sample.

Table 5.1 Sample description

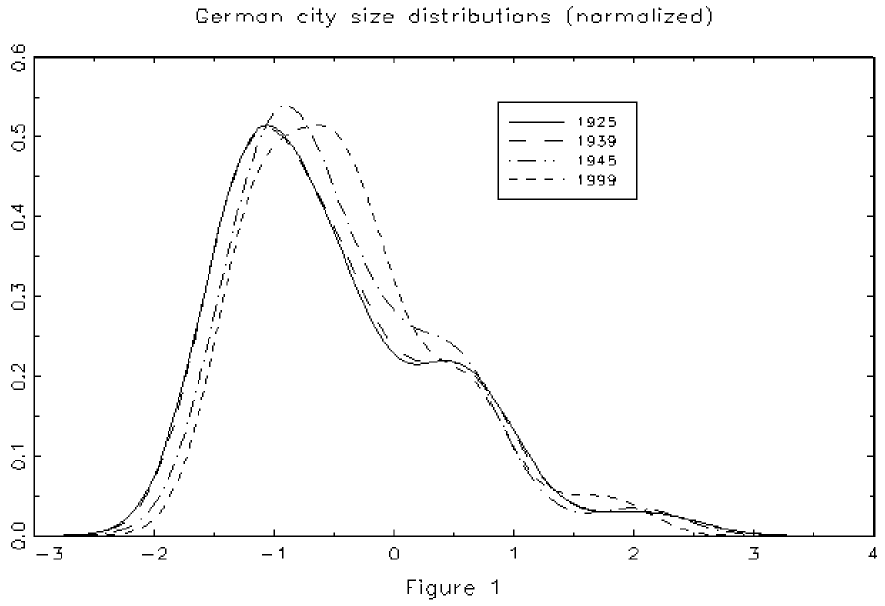
Year	total pop. (000)	% sample cities	mean city size	median city size
1925	39017	41.0	258329	124644
1935	41457	40.7	271881	136450
1945	46190	28.6	212859	107258
1955	52370	34.9	294494	156750
1965	59010	33.8	321229	176850
1975	61830	30.9	308143	180006
1985	60970	29.4	288691	173535
1999	66834	27.9	300295	177835
$\Delta_{'25-'99}$	71.3%	-32.1%	16.3%	42.7%

The impact of WWII also clearly shows up from Table 5.1. Whereas total West-German population increased by about 11% between 1935 and 1945, the average population of the cities in our sample decreased by over 20%, indicating that the urban population in particular suffered substantial losses during WWII.

5.3.1 Distribution characteristics

Before going into the analysis of the evolution of the German city size distribution and in order to fix ideas, Figure 5.1 shows the distribution[123] for both the beginning (1925) and the end (1999) of our sample period. Also included in the figure are the city size distributions right at the start (1939) and at the end (1945) of WWII. In order to control for the changes in mean city size (see Table 5.1), we normalized city sizes for each year by dividing each city's population by the mean city size in the same year.

Figure 5.1 German city size distributions (normalized)

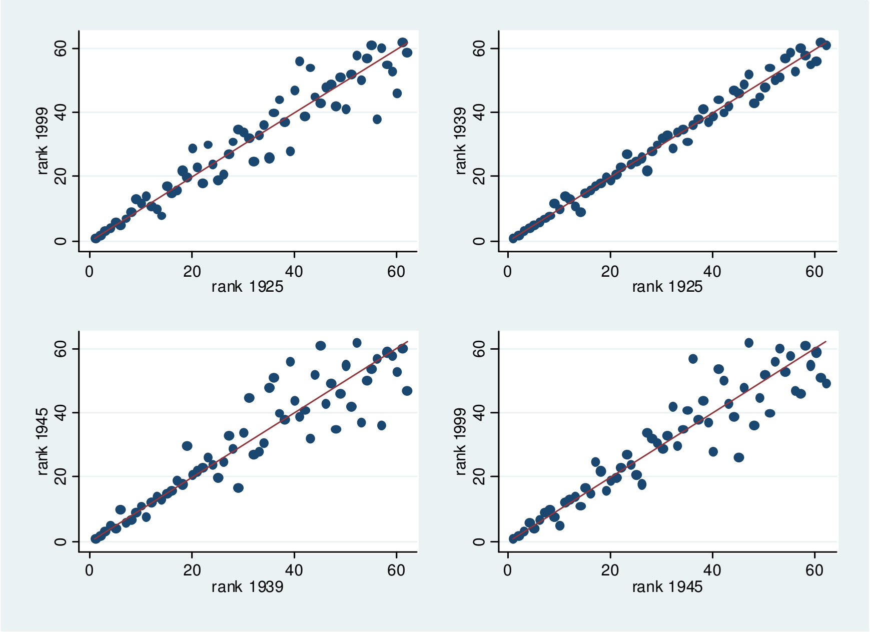


[123]The distributions are obtained by kernel estimation methods using a Gaussian kernel with the optimal bandwidth chosen using the method proposed in Silverman (1986).

These kernel estimates of the city size distribution already reveal several interesting facts about the evolution of the German urban system. The first observation is that in the pre-WWII period from 1925-1939 the distribution remained remarkably stable; the two distributions overlap almost exactly[124]. The second observation concerns the impact of WWII. The massive loss in urban population suffered during the war (see Table 5.1) induced a clear shift in the German city size distribution in a period of only six years. Comparing the 1945 distribution with the 1939 distribution, one can see that the main impact of the war is that the city size distribution loses mass in the lower tail and gains substantial mass in the middle. Keeping in mind that we are looking at the distribution of normalized city sizes this movement indicates that during the war the largest cities' population grew slower (or maybe more appropriate, considering the war destruction, declined more) than that of the smaller cities in our sample. Finally, the last interesting point to make about these kernel estimations is that in the period after the war the city size distribution does not revert to its pre-war state[125]; instead the movement that was initiated during the war seems to propagate itself with the distribution gaining even more mass in the middle and losing mass in the lower tail.

As a first glance at the dynamics within the distribution over our sample period, Figure 5.2 plots cities' initial against cities' final rank within the city size distribution for the pre-WWII, the WWII, the post-WWII and the total sample period. Table 5.2 complements this by providing the cities that made the largest rank movement up and the largest rank movement down within the city-size distribution respectively, as well as the absolute rank movement of the average city and the standard deviation of the average city's (absolute) rank movement within the city size distribution.

Figure 5.2 Rank changes



[124] Plotting the distributions for the years between 1925 and 1939 confirms this, but these are left out of the figure for sake of clarity.

[125] Again, plotting the distributions for the years between 1945 and 1999 confirms this pattern but are left out of the figure for sake of clarity.

Figure 5.2 and Table 5.2 clearly show that cities' intradistributional rank movements were considerably lower before WWII than during or after the war, conveying the same message of a major impact of the war on the urban landscape in Germany.

Table 5.2 Winners, losers and average absolute city rank movement

	pre-WWII	WWII	post-WWII	1925-1999
max (–)	-5 (Pforzheim)	-17 (Würzburg)	-21 (Flensburg)	-15 (Wannne-Eickel)
max (+)	5 (Stuttgart)	21 (Flensburg)	19 (Münster)	18 (Oldenburg)
mean	1.7	4.7	3.7	4.0
s.d.	1.6	5.5	4.4	3.9

Also the magnitude of the change in rank of the winner and the loser increases considerably during and after WWII. Over the whole sample period the average city moves four places up or down in the city size distribution. Interestingly, rank changes are more frequent (and larger) for smaller cities, which reflects the fact that the population differences between smaller cities are generally smaller so that a given population in- or decrease will more easily result in a rank change. Taken together, Table 5.2 and Figure 5.1 and 5.2 suggest that both the movement of the distribution as a whole as well as the relative position of cities within the distribution are of importance.

5.3.2 Distributional dynamics

To take a closer look at the (intra-)distributional dynamics suggested by Figure 5.1 and Table 5.2, we now turn to the estimation of the movement of the city size distribution over the sample period. In order to do so, we use Markov chain techniques following Dobkins and Ioannides (2000), Black and Henderson (2003) and Eaton and Eckstein (1997). These techniques quantify the dynamics of the city size distribution as a whole based on the intradistributional dynamics of the individual cities that make up this distribution (see also chapter 4, section 4.3.2, for more details and some references to the empirical economic growth literature where the use of Markov chains is also widespread). The use of Markov chain techniques requires the quantification of the distribution by discretizing it, i.e. each city is assigned to one of a predetermined number of groups based on its relative size. Letting f_t denote the vector of the resulting discretized distribution at period t and assuming that the distribution follows a homogenous, stationary, first order Markov process, the distributional dynamics can be characterized by the following Markov chain,

$$f_{t+x} = Mf_t \tag{5.1}$$

where M is the so-called x-period transition matrix that maps the distribution at period t into period $t+x$. Each element m_{ij} in the transition matrix represents the probability that a city makes a move within the discretized distribution from group i in period t to group j in period $t+x$. To discretize the city size distribution, we allocate each city to one of five groups based on its relative size. This requires the definition of the cut-off points that determine which city belongs to which group. Following Eaton and Eckstein (1997) and Quah (1993a), we choose cut-off points exogenously, i.e. at city sizes of 0.25, 0.5, 1 and 2 times the average city size, μ_t, for a given year t[126]. Table 5.3 shows the resulting discretized distributions for the same years as for which Figure 5.1 shows the kernel estimates of the entire empirical city size distribution.

Table 5.3 Discretized city size distributions

City size (*S*)	1925	1939	1945	1999
1) $S < 0.25\mu$	0.129	0.129	0.065	0.016
2) $0.25\mu > S < 0.5\mu$	0.387	0.355	0.435	0.371
3) $0.5\mu > S < \mu$	0.210	0.242	0.226	0.355
4) $\mu > S < 2\mu$	0.161	0.193	0.177	0.177
5) $2\mu > S$	0.113	0.081	0.097	0.081

Notes: The numbers in the Table indicate the share of cities that fall in a particular category in a particular year. For example in 1925, 13% of the cities fell in the smallest category.

Even though the distributions are substantially simplified by the discretization, the aforementioned pattern of stability before WWII and a shift towards the middle of the distribution during the war shows up in Table 5.3.

Having discretized the distribution, we can now turn to the estimation of the transition matrix, *M*. As we have yearly population data we choose to estimate the 1 year ($x=1$) transition matrix. Each transition probability, m_{ij}, in the transition matrix *M* is estimated by maximum likelihood along with its standard error, $\hat{\sigma}_{m_{ij}}$, i.e.

$$\hat{m}_{ij} = \frac{\sum_{t=1}^{T-1} n_{it,jt+1}}{\sum_{t=1}^{T-1} n_{it}}, \qquad with \quad \hat{\sigma}_{m_{ij}} = \sqrt{\frac{\hat{m}_{ij}(1-\hat{m}_{ij})}{N_i}} \tag{5.2}$$

where $n_{it,jt+1}$ denotes the number of cities moving from group i in year t to group j in year $t+1$, n_{it} the number of cities in group i in year t, and $N_i = \sum_{t=1}^{T-1} n_{it}$. Using (5.2) we estimate the 1-year transition matrix for the pre-WWII period, the 6-year transition matrix during WWII (1939-1945) and the 1-year transition matrix for the post-WWII period. Tables 5.4-5.6 show the corresponding estimates of these matrices.

The diagonal elements of the 1-year transition matrices before (Table 5.4) and after (Table 5.6) the war are close to one, which indicates that the city size distribution does not change dramatically over a period of one year. It is very interesting to note however that where before WWII all off-diagonal elements are not significantly different from zero, this changes after WWII when almost all off-diagonal elements are significantly different from zero. This significant off-diagonal movement indicates less stability of the distribution after the war, which complies with the visual inspection of Figure 5.1.

[126] Although quantitatively the results are sensitive to the choice of cut-off points this has no effect on the qualitative outcomes of our analysis.

Table 5.4 Pre WWII 1-year transition matrix

t / t+1	1	2	3	4	5
1	0.983 (0.012)	0.017 (0.012)	0	0	0
2	0.006 (0.004)	0.988 (0.006)	0.006 (0.004)	0	0
3	0	0 (0)	0.995 (0.005)	0.005 (0.005)	0
4	0	0	0.007 (0.007)	0.986 (0.009)	0.007 (0.007)
5	0	0	0	0.034 (0.019)	0.966 (0.019)

Notes: 1,2,...,5 correspond to the different groups of the discretized distribution as in Table 5.3. The 2nd largest eigenvalue is 0.996.

Table 5.5 WWII transition matrix

1939 / 1945	1	2	3	4	5
1	0.250 (0.153)	0.750 (0.153)	0	0	0
2	0.091 (0.061)	0.818 (0.082)	0.091 (0.061)	0	0
3	0	0.200 (0.103)	0.733 (0.114)	0.067 (0.064)	0
4	0	0	0.083 (0.080)	0.833 (0.108)	0.083 (0.080)
5	0	0	0	0	1 (0)

Notes: 1,2,...,5 correspond to the different groups of the discretized distribution as in Table 5.3. The 2nd largest eigenvalue is 0.992.

Table 5.6 Post WWII 1-year transition matrix

t / t+1	1	2	3	4	5
1	0.933 (0.024)	0.067 (0.024)	0	0	0
2	0.003 (0.001)	0.989 (0.003)	0.008 (0.002)	0	0
3	0	0.004 (0.002)	0.994 (0.002)	0.002 (0.001)	0
4	0	0	0.006 (0.003)	0.982 (0.006)	0.012 (0.005)
5	0	0	0	0.018 (0.007)	0.982 (0.007)

Notes: 1,2,...,5 correspond to the different groups of the discretized distribution as in Table 5.3. The 2nd largest eigenvalue is 0.996.

The magnitude and direction of this significant movement after the war can also be inferred when looking at the difference between the upper and lower off-diagonal entries of the transition matrix. This shows that most movement occurs from the smallest cities and, be it a lot smaller in magnitude, from the largest cities towards the middle of the distribution. This movement can be seen as evidence of a tendency of the distribution to gain mass in the middle indicating a city size distribution with more cities of medium size.

Next we turn to the impact of WWII on the German city size distribution. Table 5.1 already showed that the German urban population suffered a substantial loss during the war with a decrease of more than 10%. The WWII-transition matrix in Table 5.5 shows the effect of this loss of urban population on the distribution in our sample. The most striking result is that 75% of the cities in the smallest category in 1939 made the transition to the 2nd to smallest category during the war period, indicating that the smallest cities suffered substantially less than the average city during the war. Second in magnitude is the finding that 20% of the cities in the middle category moved one category down to the 2nd to smallest group, indicating that small-medium sized cities suffered quite substantial losses during the war. Not a single city in the highest category shifts to a lower category due to the destruction during the war. This however does not indicate that these cities were not hit very hard during the war, it merely reflects the fact that these cities were very large before the war and the substantial loss of population during the war was not large enough to make them shift to a lower category in the discretized distribution.

Besides the fact that the transition matrices themselves are of interest, one can also use them to do interesting thought experiments (see initially Ioannides and Overman, 2004, but also Black and Henderson, 2003 and Eaton and Eckstein, 1997). Using our estimated pre- and post-WWII transition matrices we ask ourselves the following two questions:

1. Assuming WWII would not have happened and the transition matrix remained as in the pre-WWII period, what would the city size distribution have looked like in 1945 and 1999?
2. Assuming that the estimated transition matrix remained as during the pre-WWII (post-WWII) period, would the city size distribution converge and if so what would it look like in the steady state?

To answer the first question, we use the estimated pre-WWII transition matrix from Table 5.4 and the observed distribution in 1925 from Table 5.3 to predict the distribution in 1945 and 1999, using the following formula,

$$\hat{f}_{1925+x} = \hat{M}^{x}_{pre} f_{1925} \tag{5.3}$$

where $\hat{f}_{1925+x}$ is the distribution x years from the observed distribution in year 1925 and $\hat{M}^{x}_{pre}$ is the pre-WWII transition matrix multiplied x times by itself, e.g. $\hat{M}^{2}_{pre} = \hat{M}_{pre} \times \hat{M}_{pre}$. Column 1 and column 3 of Table 5.7 below give the resulting predicted distributions for 1945 and 1999 (taking $x = 20$ or 74) respectively.

Comparing these two with the actual observed distributions for these years in Table 5.3, we see that for 1945 the main difference between the predicted -as if WWII did not happen- distribution and the actual distribution is found in the smallest two groups.

Table 5.7 Predicted city size distributions

city size (S)	1945	1999	limit	limit
transition matrix	M_{pre}	M_{pre}	M_{pre}	M_{post}
observed distribution	f_{1925}	f_{1925}	-	-
1	0.130	0.115	0	0.010
2	0.341	0.257	0	0.237
3	0.255	0.347	0.515	0.485
4	0.199	0.232	0.404	0.160
5	0.075	0.049	0.081	0.108

Notes: M_{pre}, M_{post} denote the estimated 1-year transition matrix for the pre- and post-WWII period transition matrix (see Table 5.4 and 5.6) respectively. 1,2,...,5 correspond to the different groups of the discretized distribution as in Table 5.3.

The smallest group is predicted much too large, about 200% as large as observed and the second smallest group much too small, about 25% that of which observed. Comparing the actual and predicted -as if WWII did not happen- distribution for 1999, one sees a continuation of this pattern with the smallest group being predicted 10 times as large as observed and the second smallest group predicted at about two thirds the actual size. This confirms the notion that during WWII the smallest cities in the distribution grew faster, or better suffered a lower loss of population, than the average German city, and a continuation of this pattern after the war.

The same conclusion can be drawn from our second thought experiment: what would happen to the city size distribution if it continued to evolve as estimated by either the 1-year pre-WWII or 1-year post-WWII transition matrix? If one is willing to accept this continuation assumption, these limiting distributions are shown in column 3 and 4 for the pre-WWII and the post-WWII case respectively[127]. The pre-WWII limiting distribution is not that informative because it assigns zero mass to the two smallest groups, which is due to the fact that the pre-WWII transition matrix gives zero probability to a city in one of the three highest categories to make the transition to one of the two smallest categories (see Table 5.4). As cities in the two smallest groups do eventually transfer to the higher categories, this will in the limit result in an emptying of these two categories. It does however suggest a movement towards a city size distribution characterized by medium-large cities. The limiting distribution based on the transition matrix after the war gives a completely different picture, namely that of a city size distribution characterized by small-medium sized cities, i.e. a more equal spreading of population over the urban landscape.

The overall impression from the above analysis of both the intradistributional and distributional dynamics of the German city size distribution is that WWII did have a substantial direct and lasting effect on the German urban landscape[128]. In the introduction of this chapter we stated that theories on urban growth could be distinguished according to their prediction regarding the stability of the city-size distribution to shocks. Based on the evidence

[127] These limiting distributions correspond to the (normalized) eigenvector of the respective transition matrix associated with the eigenvalue equal to one. The condition for the limiting distribution to exist is for the second largest eigenvalue to be smaller than one, which holds in our case (see the notes at the bottom of Tables 5.4-5.6).

[128] A likelihood ratio test for the time-homogeneity of the transition probabilities before and after WWII, also confirms this notion. (LR-statistic, distributed $\chi(8)$: 70.70 [p-value 0.000]).

presented in this section our conclusion is that the (evolution of the) German (i.e. West-German) city-size distribution during the period 1925-1999 has been sensitive to the WWII shock[129] both in terms of the city size distribution as a whole, as well as in terms of the relative position of cities within the distribution. So far our analysis has been largely descriptive however, in the remainder of the chapter we will therefore turn from describing the evolution of the city size distribution to providing empirical tests that allow us to distinguish more properly between the three competing theories mentioned in the introduction.

5.4 ZIPF'S LAW AND GIBRAT'S LAW OF PROPORTIONAL EFFECT

As already mentioned in the introduction, the notion of a power law distribution describing the upper tail of city size distribution goes at least as far back as 1913 when Auerbach noted this to be the case for Germany. The empirical literature has mainly focused on a special case of such a power law, namely that of city sizes in the upper tail of the distribution being distributed Pareto with coefficient $a = 1$[130]. This empirical regularity is better known as Zipf's law and has been found to hold approximately for many countries over several years (see e.g. Soo, 2005 and Nitsch, 2005)[131]. The studies that test for Zipf's Law mainly do so by means of a Zipf regression, that is regressing the log of cities' rank *r* within the distribution on the log of city sizes. If city sizes are indeed distributed according to a power law it can be easily shown (see e.g. Eeckhout, 2004) that the rank of a particular city *i* in this distribution is given by:

$$r_i = N\left(\frac{S_o}{S_i}\right)^a \tag{5.4}$$

where S_i is the size of city *i*, S_o is a (arbitrary) minimum city size, and *N* the number of cities above this truncation point. Testing for Zipf's law can now be done by a regression of ln(*r*) on ln(*S*) in order to estimate the so-called Zipf coefficient, i.e. rewriting (5.4) in logs gives:

$$\ln(r_i) = \alpha - a\ln S_i + \varepsilon_i \tag{5.5}$$

, where $\alpha = \ln(N) + a\ln(S_o)$ is a constant and ε_i a random error term. If it cannot be rejected that the estimated $\hat{a}$ equals 1 this constitutes evidence in favor of Zipf's Law. We estimated (5.5) using OLS for all years in our sample separately:

$$\ln(r_{it} - 0.5) = \alpha_t - a_t \ln(S_{it}) + \varepsilon_{it} \tag{5.6}$$

[129] This is consistent with the findings in Brakman et al (2004a) and Bosker et al (2007a), see also section 5.5.

[130] Formally a variable (in our case city size), *S*, adhering to a power law is distributed according to a Pareto distribution if the density function of this variable satisfies, $p(S) = \frac{aS_o^a}{S^{a+1}} \quad \forall S \geq S_o^a$

[131] Approximately is the key word here, see also Brakman, Garretsen and van Marrewijk (2001), ch. 7 or Gabaix and Ioannides (2004) for evidence that the Zipf coefficient is sometimes significantly different from 1 and/or changing over time.

, where we add –0.5 to each city's rank following the recommendation in Gabaix and Ibragimov (2007). This ensures that we get unbiased estimates of a. Figure 5.3 below shows the results for a and its corresponding 5% confidence interval ($\hat{a} \pm 2\hat{\sigma}_a$) corresponding to the estimated OLS standard error of $\hat{a}$, $\hat{\sigma}_a$, for each of the years in our sample[132].

Figure 5.3 The estimated Zipf-coefficient over time

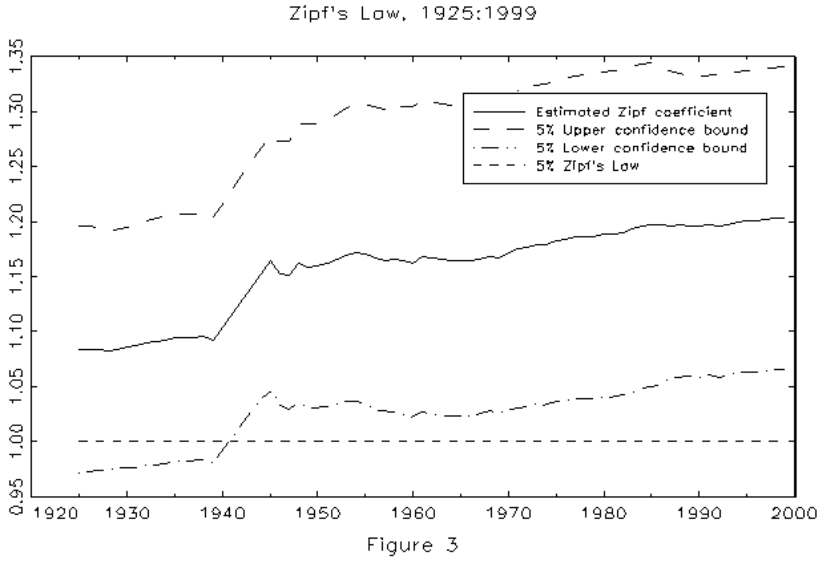


Figure 5.3 shows that during the pre-WWII period the point estimate of a is very close to 1. The impact of WWII also shows clearly when looking at the Zipf-regression results. The estimated Zipf coefficient, $\hat{a}$, increases from 1.09 in 1939 to about 1.16 right after the war, again confirming the notion of a more equal spread of urban population over the West-German cities in our sample due to the relative lower loss of urban population of smaller cities during WWII. In the post-WWII period the point estimate shows no return to 1, instead it steadily increases, finally reaching a point estimate of 1.20 in 1999.

These results are consistent with the idea that WWII has had a major impact, shocking the city size distribution from one adhering closely to Zipf's Law to one characterized by a more equal spread of urban population over the different cities. This immediate impact of the war is not reversed in the post-WWII period, instead the distribution moves to one characterized by an even more equal spread of urban population. This confirms our earlier findings in the previous section based on the Markov chain analysis. Note however that this post-WWII tendency to a more equal spreading of urban population in the post-WWII period is also found for other developed countries (see e.g. Soo, 2005 and Nitsch, 2005). The striking finding in case of Germany is that WWII has had a non-trivial contribution to this process.

[132] As was pointed out by Gabaix and Ioannides (2004), there are some pitfalls when estimating a Zipf regression by OLS. The standard errors of the estimated $\hat{a}$ are also typically underestimated leading to an overrejection of Zipf's Law. We keep to OLS noting that using the approximate standard errors as suggested by Gabaix and Ibragimov (2007) indeed does not result in a rejection of Zipf's Law. We also considered using the Hill estimator as in e.g. Dobkins and Ioannides (2000), however as noted in Embrechts et al. (1997) and Gabaix and Ioannides (2004) the small sample properties of this estimator are particularly bad and we therefore decided not to use it for our sample of 62 cities. Results not shown in the chapter are all available upon request.

5.4.1 Individual city size evolution and Gibrat's Law of proportional effect

Gabaix (1999) and also Eeckhout (2004) opened up an alternative way to empirically verify the relevance of Zipf's Law for a particular city size distribution. Here we follow Gabaix (1999) who showed that if individual city size growth adheres to Gibrat's Law of proportional effect, i.e. if city sizes grow randomly with the same expected growth rate and the same variance independent of city size, and is also subject to a lower reflective boundary (where smaller cities that hit the boundary from above are 'bounced back upward'), the city size distribution converges to one adhering to Zipf's law. It is the combination of a random walk process with a lower reflective boundary that gives a Pareto distribution with a coefficient of one, i.e. Zipf's Law. Without this lower reflective boundary, the random walk in city sizes would result in the city size distribution being log-normally distributed (see Eeckhout, 2004)[133]. As already briefly mentioned in the introduction, the fact that proportionate city growth gives rise to a lognormal city size distribution is also frequently referred to as Gibrat's Law (see e.g. Sutton, 1997 and Eeckhout, 2004). To avoid confusion, whenever we refer to Gibrat's Law from now on, we mean, as most authors in the modern city size distribution literature (see e.g. Gabaix, 1999, Gabaix and Ioannides, 2004 or Black and Henderson, 1999), Gibrat's Law of proportional effect. Gabaix's (1999) contribution, showing that Gibrat's Law combined with a lower reflective boundary leads to Zipf's Law in (the upper tail of) the overall city size distribution, provides a link with the results we showed in the previous section. In the next sections we look for empirical evidence on Gibrat's Law using both nonparametric and parametric techniques.

5.4.2 Nonparametric evidence on Gibrat's Law

Ioannides and Overman (2003) and Eeckhout (2004) resort to nonparametric multivariate kernel estimations to shed empirical light on the relevance of Gibrat's Law. Both papers find considerable evidence in favor of Gibrat's Law in case of the US urban system. Following this methodology, we plot the distribution of five-year city size growth rates conditional on initial city size (in logs) for the total, pre-WWII, post-WWII and WWII period along with the corresponding contour plots[134]. The results are shown in Figures 5.4a-5.4d below.

[133] For our present purposes we are not as such interested in the question whether the German city size distribution displays log-normality. Instead we want to find out if in the German case individual city growth is driven by a process of proportional growth. Also our dataset only covers the upper tail of the city size distribution, so that we are wary of making statements about the log-normality of the city size distribution as a whole. As pointed out by one of our referees, the normalized city size distributions shown in Figure 5.1 already suggest that city-sizes are neither log-normally nor Pareto distributed. Indeed, when we test for this using various standard tests for log- normality (skewness kurtosis, Shapiro Wilk or Shapiro Francia) or the 'Paretoness' of the distribution (Kolmogorov-Smirnov), we invariably reject the city size distribution to be log-normal or Pareto for the four years shown in Figure 5.1 (results available upon request).

[134] The stochastic kernels are estimated non-parametrically using a Gaussian kernel and with the bandwith chosen following Silverman (1986). The contour plots can be read in the same way as standard topographical height maps with the lines in the plots connecting points on the distribution of similar height.

Figure 5.4a Pre-WWII stochastic kernel estimate: log city size to 5 year city growth

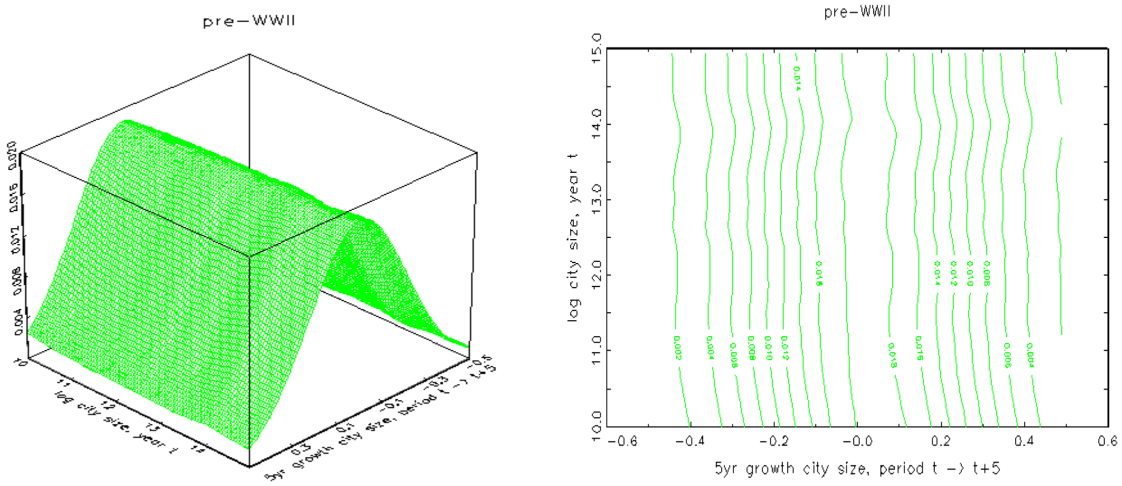


Figure 5.4b WWII stochastic kernel estimate: log city size 1939 to city growth WWII

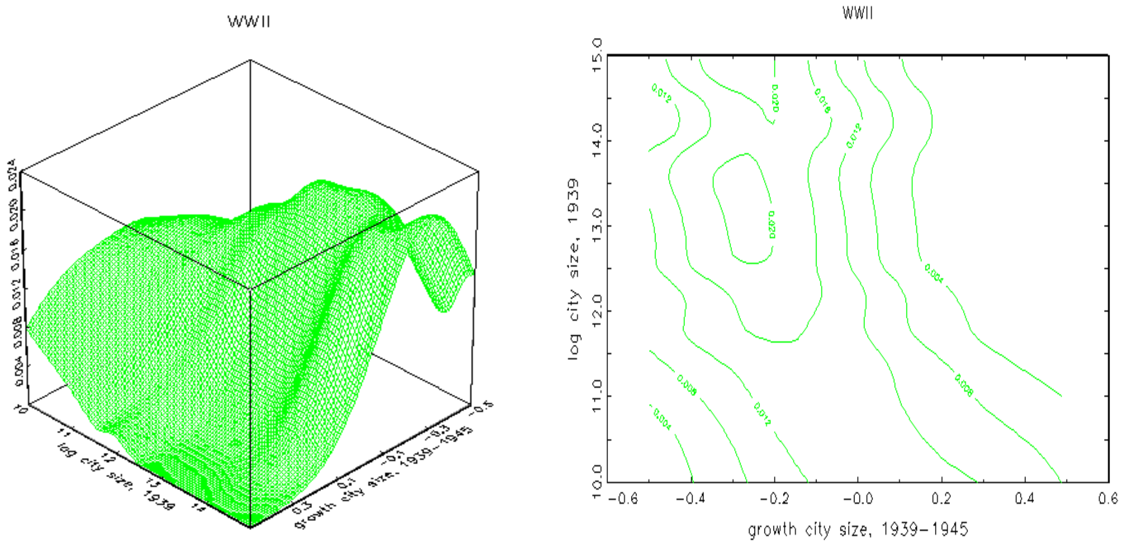


Figure 5.4c Post-WWII stochastic kernel estimate: log city size to 5 year city growth

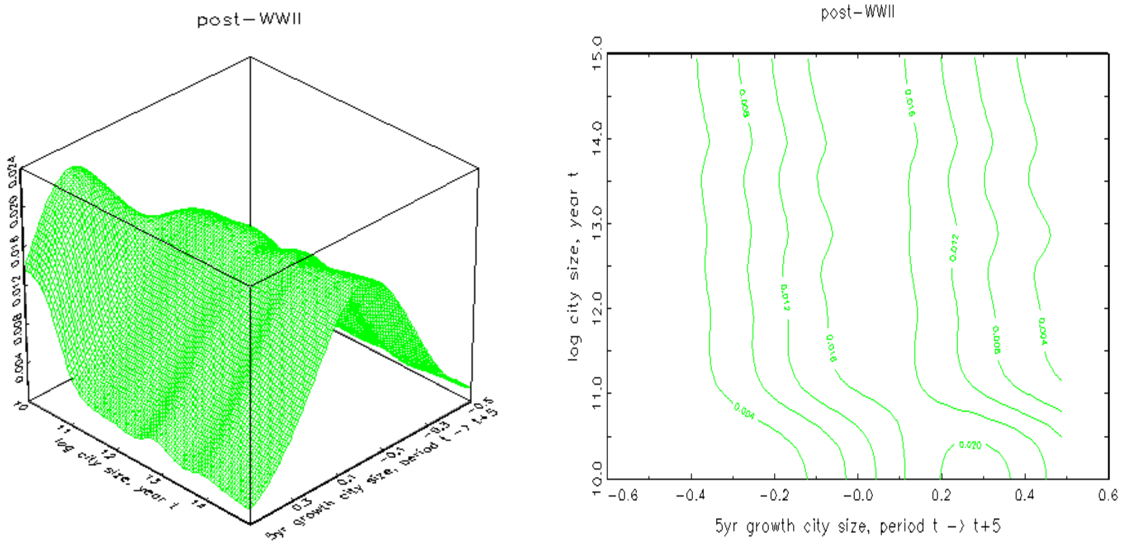
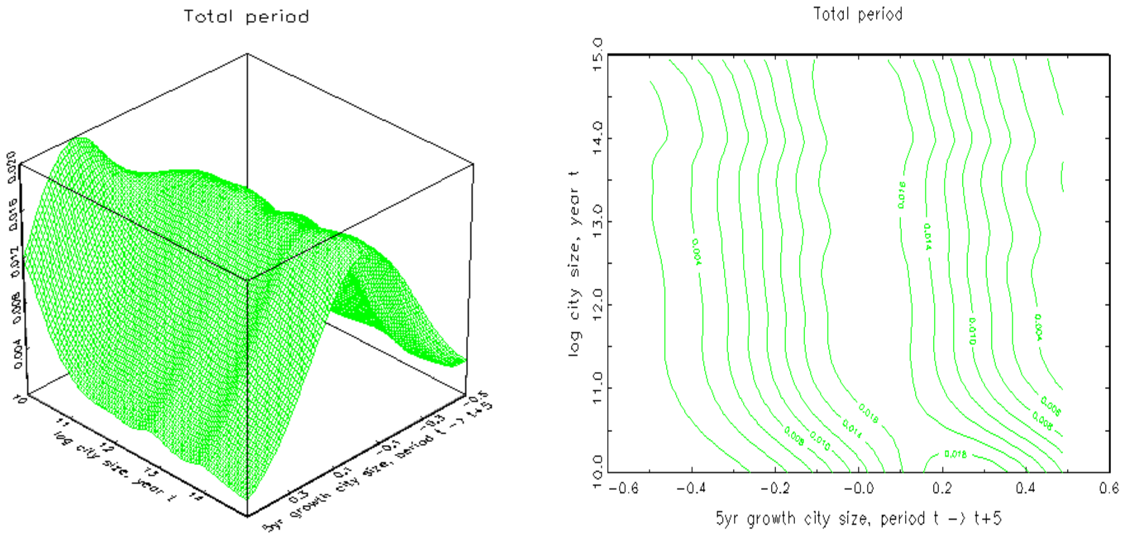


Figure 5.4d Stochastic kernel estimate 1925-1999: log city size to 5year city growth



By taking any point, say *S*, on the city size axis and taking the cross-section through the kernel estimate parallel to the city size growth rate axis, one obtains the distribution of five year growth rates conditional on the log of initial city size being S.

The plotted kernel in Figure 5.4a shows that before WWII, German city size growth behaved remarkably well according to Gibrat's Law: the estimated kernel and corresponding contour plot show no significant difference in the conditional growth rate distribution for cities of different size. This provides evidence that the growth rate of the cities in our sample in the pre-WWII period did not depend on the initial size of the city. It is interesting to relate these results to our estimated Zipf regression: the fact that before WWII individual city growth seems to adhere quite well to Gibrat's Law corresponds to the finding of a Zipf-coefficient very close to 1 during that period (see Figure 5.3).

After WWII, the evidence from the estimated stochastic kernel and contour plot in Figure 5.4c shows a completely different picture. For the largest cities in our sample city growth rates still seem largely independent of initial city size. This does not hold true for the smaller cities in our sample however. Figure 5.4c shows clear evidence that the smaller cities in our sample grow on average faster than the larger cities. This shift in the mean of the conditional distribution for smaller cities is most clearly seen from the contour plot, which makes a significant shift to the right for the smallest cities. The evidence for Gibrat's Law is thus weaker for the post-WWII period, reflecting itself in the higher Zipf-coefficient in Figure 5.3.

Before turning to the overall picture of individual German city growth over the whole sample period from 1925-1999, we turn to the immediate impact of WWII on German city sizes. As can be seen in Figure 5.4b, the largest cities suffered more heavily during the 6-year period of WWII. Smaller cities suffered less on average but the variance of city growth during the war also increases substantially for the smaller cities in the sample. This suggests that larger cities were not only damaged more heavily, the destruction of these larger cities was also less variable; or put differently, the larger cities were all hit quite similarly and more heavily compared to the smaller cities in our sample (evidence of the success of the Area Bombing-tactics of Allied Bomber Command aimed at destroying Germany's urban centers, see Brakman et al., 2004a).

Combining Figures 5.4a-5.4c, the estimated stochastic kernel for the whole sample period in Figure 5.4d does not provide very clear evidence that Gibrat's Law holds. The effect of the lesser destruction of smaller cities during WWII, and the higher growth rates of these same smaller cities after the war, dominates the equal growth rates before the war. As a result, the overall picture shows a similar (be it somewhat smaller) shift in the conditional growth rate distribution for the smaller cities in the sample as for the post-WWII case. This result of a size dependent mean growth rate, and a more or less initial size independent variance contrasts to empirical studies done using US city size data. Both Eeckhout (2004) and Ioannides and Overman (2003) find evidence for the USA of a more or less size independent mean growth rate and an increasing variance for smaller cities, which can be reconciled with the decreasing Zipf coefficient they find for the USA. The explanation of the higher Zipf coefficient found in our sample over the post-WWII period is a more equal spreading of urban population over the cities in Germany and not so much an increased variance of city growth for smaller cities.

5.5 UNIT ROOT TESTING AND PARAMETRIC EVIDENCE ON GIBRAT'S LAW

The nonparametric kernel estimates of the previous section remain largely based on pooled panel data evidence. In order to fully exploit the time series dimension of our data set, we now turn to a, more dynamic way of testing for Gibrat's Law. Given the many observations over time in our data set, we follow the suggestion made by Gabaix and Ioannides (2004) who state: *"Hence one can imagine that the next generation of city size evolution empirics could draw from the sophisticated econometric literature on unit roots"* (Gabaix and Ioannides, 2004, p. 20).

Clark and Stabler (1991) were one of the first to notice that Gibrat's Law can be tested using unit root tests. Following their exposition of the relationship between Gibrat's Law and unit root testing, assume that the size of city *i* at time *t*, can be related to the size of that same city at time *t-1* according to the following formula:

$$S_{it} = \gamma_{it} S_{it-1} \tag{5.7}$$

where γ_{it} denotes the growth rate of city *i* over the period *t-1* to *t*. Next assume that this growth rate can be decomposed into three components, a random component ε_{it}, a non-stochastic component relating the current growth rate to a (possibly time-varying) constant, past growth rates and initial city size:

$$\gamma_{it} = K_{it} S_{it-1}^{\delta_i} \prod_{j=1}^{p} \gamma_{it-j}^{\beta_{ij}} (1+\varepsilon_{it}) \tag{5.8}$$

where K_{it} is a possibly time-varying constant, and δ_i and β_{ij} are parameters measuring the relative importance of initial city size and past growth rates on current city growth respectively and ε_{it} is a random error term. Now Gibrat's Law would require $\delta_i = 0$, such that initial city size does not influence the growth of a particular city. In order to be able to test for this, substitute equation (5.8) into (5.7), take logs and subtract $\ln S_{it-1}$ from both sides of the equation to obtain the following estimatable equation:

$$\Delta \ln S_{it} = c_{it} + \rho_i \ln S_{it-1} + \sum_{j=1}^{p} \beta_{ij} \Delta \ln S_{i,t-j} + \varepsilon_{it} \tag{5.9}$$

where $c_{it} = K_{it}$, $\rho_i = (\delta_i - 1)$ and the following approximate equality is used: $\ln(1+\varepsilon_{it}) \approx \varepsilon_{it}$ for small values of ε_{it}. This shows immediately that testing for Gibrat's Law amounts to testing for a unit root in city sizes. If we find that ρ_i is not significantly different from zero, i.e. a unit root in city size, this constitutes evidence in favor of city *i*'s growth rate being independent of city *i*'s size. On the other hand an estimated ρ_i smaller than zero would indicate that the evolution of city *i* is a stationary process implying that city *i*'s growth rate declines with initial city size.

There are to date very few studies that actually perform unit root tests on individual city sizes. Notable exceptions are Clark and Stabler (1991), Black and Henderson (2003) and Sharma (2003). Clark and Stabler (1991) conclude in favor of the relevance of Gibrat's Law based on the evolution of the city sizes of the seven largest Canadian cities over the period 1975-1984. As the ADF unit root tests are infamous for their small sample properties, a sample period of only 10 years seems to put a substantial doubt on their results[135]. Black and Henderson (2003) find no evidence for Gibrat's Law using data on the US metropolitan areas when testing for a unit root in city sizes. As Clark and Stabler (1991) they also have only 10 observations over time (decade-by-decade city sizes over a total period of 90 years). They take explicit note of the hereby potentially induced small-sample problem and resort to a recently proposed panel unit root test, i.e. Levin-Lin-Chu (2002). Given their very large cross-section dimension, the use of this panel technique is likely to solve the problems associated with the small sample bias. However as noted by Gabaix and Ioannides (2004), their test does not correct for the potential autocorrelation in the residuals, which can severely bias the results regarding the unit root hypothesis and thus regarding the relevance of Gibrat's Law. Furthermore they do not control for cross-sectional dependence across the cities in the panel, which sheds further econometric doubt on the robustness of their results. Using Indian population data for the period 1901-1990, Sharma (2003) also performs unit root tests on the population of India's cities and finds evidence of non-stationarity of city growth. The latter paper does however not explicitly link its finding to Gibrat's Law, instead focusing more closely on the short-run dynamics of Indian city growth.

Given the features of our data set, i.e. annual population data for the period 1925-1999 (except for the 5 years during the WWII period, 1940-1944), we argue that we can use standard unit root tests on individual city sizes.[136] Given the decline of total urban population

[135] The authors note this problem and propose the estimation of a restricted SUR model to overcome this problem, however the distributional properties of this SUR estimator are not known and given the small number of cities this is unlikely to solve the potential small sample bias.

[136] We have to admit that the power properties of the unit root rest depend mostly on the time span of the data (here, 1925-1999) and not so much on the frequency of observation (see Campbell and Perron, 1991). In the German case, however, the large time span in combination with the annual data makes it possible to deal more adequately with the short-term dynamics (e.g. correcting for autocorrelation), which could be especially

over the sample period we decided to allow the constant term in the growth rate, K_{it}, to be possibly trend-wise changing over time[137], i.e. $K_{it} = K_i t$ resulting in the following equation that will be estimated for each city separately:

$$\Delta \ln S_{it} = c_i + \zeta_i t + \rho_i \ln S_{it-1} + \sum_{j=1}^{p} \beta_{ij} \Delta \ln S_{i,t-j} + \varepsilon_{it} \tag{5.10}$$

The first column in Table 5.8 shows the result of these city specific unit root tests. It shows the percentage of cities for which the null of a unit root is rejected at a 1%, 5% and 10% level.

Table 5.8 Results unit root tests on city sizes

	A: City specific tests	
alternative hypothesis:	trend stationary	trend stationary with break
significance level	%unit root rejected	%unit root rejected
1%	0%	59.1%
5%	0%	74.2%
10%	0%	75.8%
	B: Panel tests	
test	test-statistic (p-value)	
Levin-Lin-Chu	0.543 (0.706)	
Im-Pesaran-Shin	-2.061 (0.020)	

Notes: The null hypothesis is in all cases a unit root in city size. Following the suggestion in Ng and Perron (1995) we choose the optimal number of lagged growth rates to be included in the regression to control for autocorrelation using a 'general-to-specific procedure' based on the t-statistic. The maximum lag length to start off this procedure is set at 11. The panel test statistics are the t^*- and the $\bar{Z}-$ statistic in case of the Levin-Lin-Chu and Im-Pesaran-Shin test respectively. All panel statistics are calculated controlling for cross-sectional dependence by subtracting yearly averages as suggested by Im et al. (2003).

For completeness, it also gives the outcome of two different panel unit root tests. The first is the earlier mentioned Levin-Lin-Chu (2002) test, which tests the null of all series having a unit root versus the alternative of all series being stationary with the same autoregressive parameter. The second is the later developed Im-Pesaran-Shin (2003) test that tests the null of a unit root in all series versus the alternative of some of the series being stationary (with a potentially varying autoregressive parameter) and some of the series being nonstationary. Hereby the latter test is thus somewhat less restrictive under the alternative.

Both the individual unit root tests and the Levin-Lin-Chu panel unit root test do not reject the null hypothesis of a unit root. Even at a 10% level the null is never rejected for any of the cities in the sample. The Im-Pesaran-Shin test on the other hand rejects the unit root null at the 5% level (it does not reject it at the 1% level), hereby providing some weaker evidence in favor of Gibrat's Law. Overall, this would suggest overwhelming evidence in favor of Gibrat's Law with all cities seemingly growing independent of their size. There is,

important surrounding large shocks like WWII, the division of Germany in 1948 or the reunification with East-Germany in 1990.

[137] Black and Henderson (2003) also include a deterministic trend in their estimated equation.

however, one major caveat when drawing this conclusion from these standard unit root tests and that is (maybe not surprisingly) the WWII shock. As shown before, the large and sudden impact of WWII had a tremendous effect on the German urban landscape with large cities losing more population more systematically compared to the smaller cities in Germany. When performing a standard unit root test one implicitly assumes that the whole effect of this destruction during the war can be viewed as a one-time extreme realization from the distribution of the error term, i.e. ε_{it} in (5.10). This however seems somewhat unlikely, instead WWII can be argued to have had a more substantial impact changing the deterministic components of city size growth, i.e. the constant and/or trend in (5.10). If this would be the case, to ignore it when performing a standard unit root test results in an underrejection of the unit root null hypothesis (see Perron, 1989 and Perron, 1997). This would imply that Gibrat's Law is potentially overaccepted by standard unit root tests in the case of German city sizes.

To allow for the possibility of a change in the deterministic components of city size growth, we follow Perron (1997) and estimate the following equation:

$$\Delta \ln S_{it} = c_i + \theta_{1i} DU_t + \zeta_i t + \theta_{i2} DT_t + \theta_{i3} D(T_b)_t + \rho \ln S_{it-1} + \sum_{j=1}^{p} \beta_{ij} \Delta \ln S_{it-j} + \varepsilon_{it} \qquad (5.11)$$

where $DU_t = 1(t > T_b)$, $DT_t = 1(t > T_b)t$ and $D(T_b)_t = 1(t = T_b + 1)$ and T_b is the time at which the change occurs. The null hypothesis still remains that of a unit root, the alternative however changes from the series being stationary around a deterministic trend to the series being stationary around a deterministic trend that is allowed to change at time T_b. The exact timing of the break date is determined *endogenously* by the data (so that also T_b is actually city specific), see Perron (1997) for details and Hansen (2001) for a discussion. We choose this procedure over the option of exogenously setting the break date at WWII to allow for the possibility of other events that could have had a major impact on German city sizes during our sample period like, for example, the separation and subsequent reunion of West and East Germany (see Redding and Sturm, 2005 and Redding, Sturm and Wolf, 2007), which potentially may have left an even bigger mark on the evolution of some of the cities in our sample.

The results of these unit root tests when we allow for a one-time break are shown in the third column of Table 5.8. The impact of allowing for a one-time break in the deterministic components of city growth is quite striking. Instead of accepting the null of a unit root for all cities in our sample as the standard unit root tests did, now the unit root null is rejected for 74.2% of the cities in favour of these series being trend-stationary with a one-time shift in this trend. The date at which the break occurs is almost exclusively found at WWII, which shows that WWII's impact on city sizes is dominating that of other historical events affecting German cities during the second half of the 20th century[138]. The high rejection rate

[138] This is confirmed when doing the unit root tests on only the post-WWII period, the unit root hypothesis being accepted for about only 38% of the cities without allowing for a break in the series. When allowing for a break the division (1948) or reunification (1990) of West and East Germany did not show up as clearly as WWII in case of the whole sample period. Using a rather different analytical framework, Redding and Sturm (2005) do find evidence that the post-WWII division of Germany had a significant effect. Apart from the different framework, one reason that they find a stronger impact of this division on (west-) German cities is that their sample includes more (smaller) West-German cities located near the former border between West and East

of the unit root null hypothesis also implies that the relevance of Gibrat's Law, which seemed evident based on the results of the standard unit root tests, gets a substantial blow. Correcting for the impact of WWII on the evolution of individual city size, Gibrat's Law is found to hold only for about one quarter of the cities in our sample.

Together with the evidence obtained using nonparametric methods in the previous section, this dynamic evidence on Gibrat's Law sheds substantial doubt on the relevance of random city size growth in case of Germany, especially for the post-WWII period[139]. Instead the data seem to indicate that city growth does depend on initial city size, with smaller cities growing faster than larger ones[140]. This faster growth of smaller cities does not comply with urban economic theories exhibiting random city growth. It suggests that other theories of urban growth are perhaps more relevant to explain the post-WWII experience of the German urban landscape. As the post-WWII and pre-WWII period are so different with respect to the implications regarding the relevance of urban economic theories, WWII seems to have had a crucial (initializing) role in the changing of both the type and evolution of the German urban system.

Having found compelling evidence on the irrelevance of random city growth as an explanation of the evolution of the German urban system, the next section, building on earlier work by Davis and Weinstein (2002) and Brakman et al. (2004a), provides additional evidence based on the evolution of *relative* city sizes, i.e. the position of cities within the city size distribution, by which we try to distinguish between the two other competing theories of urban growth, namely increasing returns to scale and locational fundamentals.

5.6 UNIT ROOT TESTING AND WWII'S IMPACT ON RELATIVE GERMAN CITY SIZES

In previous work, Brakman et al. (2004a) and Bosker et al. (2007a), we already looked at the immediate impact of WWII on German relative city size. Drawing on the methodology developed by Davis and Weinstein (2002, 2004), these papers argue that the destruction during WWII was largely exogenous to the level of economic activity in cities and use the level of destruction during WWII as instruments for population growth during WWII when estimating the following equation:

$$\Delta \ln s_{ipost-WWII} = \alpha \Delta \ln s_{iWWII} + X_i \beta + \varepsilon_i \tag{5.12}$$

, where s_i denotes city *i*'s size relative to total German population, i.e. S_i/S_{tot}, X_i are other exogenous variables that can be included in the regression and ε_i is a random error term.

Germany. An interesting alternative method that could be useful in this case is the spatial interactions analysis as in Dobkins and Ioannides (2001).

[139] The rejection of Gibrat's Law could also be an explanation why we find mixed evidence on Zipf's Law, see Figure 5.3 plus discussion.

[140] The results of the unit root tests are much the same when looking at small or large cities separately. The rejection of the unit root null in case of the IPS test seems to be largely due to the smaller cities in the sample however. Also the average autoregressive parameter of the smallest cities is somewhat smaller than for the largest cities, giving further evidence for a difference in growth process for small and large cities. Results are available upon request.

Estimating this equation for Germany, Brakman et al. (2004a) find evidence that the average German city had in 1963 returned to a relative city size of about 60% of its pre-WWII level. Bosker et al. (2007a) extend this simple framework by estimating a threshold regression in the spirit of Hansen (2000) hereby allowing for the possibility of multiple equilibria and find evidence for the existence of two different equilibria, with the least destructed cities shifting to an equilibrium characterized by a larger relative city size.

The proposed framework, i.e. estimating (5.12)[141], can however be argued to be subject to some caveats. First, the estimation results are sensitive to the choice of period over which post-WWII growth is calculated. Second, the estimation results are only able to describe the impact of WWII on the average German city; they are unable to say something about the individual experience of a particular city (the experience of the average city can even be argued to be of secondary importance when the fit of the regression is far from perfect). Third, and most important, the estimation of (5.12) is merely a static cross-section regression. Concluding that relative city sizes are mean reverting or random over time is impossible on the basis of such a simple cross-section. Instead, as argued by Hohenberg (2004) the historical evolution of the urban structure must always be studied in terms of *fully dynamic models*. This is exactly what we purport to do here. Exploiting the long time dimension of our data set, we are able to look at the evolution of each individual city's relative size employing fully dynamic econometric estimation techniques.

Davis and Weinstein (2002) already mentioned the fact that the proper test for the persistence of shocks would be performing unit root tests on relative city shares. They refrain from doing this and instead resort to the static framework in (5.12) on the basis of the earlier mentioned low power of these unit root tests in small samples. Considering the extent of our data set, we think we are in a much better position to apply such unit root tests and can hereby provide much more dynamic evidence on the 'mean reversion' of relative city sizes. More specifically we estimate the following equation for each of the cities in our sample:

$$\Delta \ln s_{it} = \xi_i + \zeta_i \ln s_{it-1} + \sum_{j=1}^{p} \Delta \ln s_{it-j} + v_{it} \qquad (5.13)$$

where s_i is the share of a particular city i in total German population, the lagged values of city growth included in the regression control for potential autocorrelation and v_{it} is a random error component. If ζ_i is found to be significantly smaller than 0, city share is stationary around ξ_i and any shock will not have a lasting effect. If on the other hand ζ_i is found to be equal to 0 then all shocks are permanent and city i's share in total German population follows a random walk[142]. We estimate (5.13) applying Augmented Dickey Fuller tests to all of the

[141] The same holds for the extended version allowing for multiple equilibria.

[142] One may be concerned that city shares are by definition bounded, so how can they exhibit a unit root, which would imply an unbounded variance? Essentially, what we are looking at is whether the process, within the bounds, can be described by a random walk. Cavaliere (2005) shows that testing for a unit root in limited time series can still be done, but some adjustments have to be made to the test statistics. He also shows that, when the process is relatively far away from the bounds, so that the range constraints are rather loose, the standard unit root tests perform quite well. As this is the case in our sample (the maximum city share is about 0.07 and the minimum city share is about 0.001, both not very close to the bounds of 1 and 0 respectively), and considering that tests for 'bounded unit roots' that allow for a break in the deterministic component(s) are not readily available, we prefer using the standard unit root tests here. Note also that not taking account of the presence of

cities in our sample. Table 5.9 shows the results; it also includes the results of the earlier mentioned Levin-Lin-Chu and Im-Pesaran-Shin panel unit root tests.

The results of both the individual city share unit root tests and the panel unit root tests are at odds with the notion of city shares being stationary over time. The null hypothesis of a unit root in city share is not rejected for almost all cities in the sample. This would constitute considerable evidence against the locational fundamentals theory. However, as was the case in our earlier unit root tests on city sizes (see Table 5.8), this conclusion does not take the possible different effect of the WWII shock into account. Also in case of city shares, the major impact of WWII could have resulted in a shift in the deterministic component of a city's relative size, i.e. a deterministic shift in the mean ξ_i. Such a change in the deterministic component could for example be due to a change in locational fundamentals as a result of the destruction in WWII.

Table 5.9 Results unit root tests on city shares in total German population

	A: City specific tests	
alternative hypothesis:	trend stationary	trend stationary with break
significance level	%unit root rejected	%unit root rejected
1%	0%	22.6%
5%	1.6%	27.4%
10%	4.8%	35.5%
	B: Panel tests	
test	test-statistic (p-value)	
Levin-Lin-Chu	0.445 (0.672)	
Im-Pesaran-Shin	0.430 (0.666)	

Notes: The null hypothesis is in all cases a unit root in city share. Following the suggestion in Ng and Perron (1995) we choose the optimal number of lagged growth rates to be included in the regression to control for autocorrelation using a 'general-to-specific procedure' based on the t-statistic. The maximum lag length to start off this procedure is set at 11. The panel test statistics are the t*- and the $\bar{Z}-$statistic in case of the Levin-Lin-Chu and Im-Pesaran-Shin test respectively. All panel statistics are calculated without controlling for cross-sectional dependence.

To allow for this possibility we apply the following unit root test suggested by Perron and Vogelsang (1992) that allows for a one time break in the mean of the series ξ_i (endogenously determined by the data) and is based on the estimate of ζ_i in the following regression:

$$\bar{s}_{it} = \zeta_i \bar{s}_{it-1} + u_{it} \tag{5.14}$$

where u_{it} is the random error term and $\bar{s}_{it}$ are the residuals of a regression that projects s_{it} on the deterministic component, i.e. a mean that is allowed to shift at time T_b. More formally:

the bounds results in an overrejection of the unit root hypothesis (see Cavaliere, 2005). Given that we, when using the standard unit root tests, already find a unit root in virtually all series, using the bounded unit root test instead would probably not change our results.

$$s_{it} = \mu_i + \gamma_i DU_t + \eta_{it} \tag{5.15}$$

where $DU_t = 1$ if $t > T_b$ and 0 otherwise. Estimating ζ_i in this way controls for the possible one-time shift in the deterministic mean in the `first stage' of the procedure (5.15) and estimates the autoregressive parameter, $\zeta_i$ in the `second stage' (5.14). Perron (1990) called this the additive outlier (AO) model, which is appropriate to model a sudden one-time change, which is clearly the case when considering the destruction caused by the heavy bombardments during WWII. Perron and Vogelsang (1992) discuss the appropriate test statistics when testing for $\zeta_i = 1$.

The results of applying the AO-model to test for a unit root in German city shares under the null versus stationary city shares around a possibly shifting mean under the alternative are also shown in Table 5.9. As was the case for the unit root tests on city sizes, the effect of taking account of the possible special nature of the WWII-shock (i.e. having an impact on the deterministic components of city shares) is quite substantial. At a 5% confidence level the unit root null hypothesis is rejected in favor of a stationary city share with a one-time break for 27% of the cities in our sample. Even more striking is the fact that for all the cities that are stationary, the timing of the break is (endogenously) found to be WWII. As when allowing for a one-time break in the deterministic component(s) in the unit root tests on city sizes, the impact of WWII overshadows the effects of the other historic events (most noteworthy the separation from and unification with East Germany) that could have had their impact on the evolution of individual cities and the city size distribution as a whole.

The evidence provided in Table 5.9 constitutes evidence against theories that fall under the locational fundamentals category. The `standard' unit root tests reject stationarity of city shares for all cities in our sample except one (Hamm). When explicitly taking account of a possible shift in locational fundamentals during WWII, by allowing for a change in the deterministic mean around which a particular city is stationary, stationarity of city sizes is accepted for a much larger proportion of the cities in our sample. This does however not save the locational fundamentals theories as being relevant in the case of Germany. Still random shocks have a persistent effect on the relative city share of more than 70% of the cities in our sample. Furthermore, although random shocks are not persistent in case of the cities that are found to be stationary, the extreme shock of WWII *did* have a lasting effect on the city share of those cities by changing the deterministic mean around which the city share is stationary[143]. As the unit root tests only find the break in the deterministic mean but do not help us to understand why this break occurs (other than the date at which the break occurs), Appendix 5.B provides a tentative look into some characteristics that distinguish the 17 "break-stationary" cities (27% of our sample at the 5% level, see Table 5.9) from the other cities in the sample.

Overall, we think that this dynamic evidence on the effect of the large shock experienced during WWII constitutes considerable evidence against the relevance of

[143] This in fact is corroborated in Bosker, et al. (2007a).

locational fundamentals theories in explaining the evolution of the German urban landscape. The relevance of the other theories (random growth and increasing returns to scale) is somewhat harder to assess using the results of the unit root tests in this section.

5.7 CONCLUSIONS

Most of the empirical literature on city size distributions has focused on the USA. Other countries might experience a different evolution of their city size distribution, as this chapter shows to be the case for West Germany. Using a unique annual data set for 62 West-German cities that covers most of the 20th century, we look at the evolution of both the city size distribution as a whole and of each city separately. The West-German case is of particular interest as its urban system has been subject to some of history's largest (exogenous) urban shocks, most notably WWII and the German division and subsequent reunification. Our data set allows for the identification of these shocks and provides evidence on the effects of these 'quasi-natural experiments' on the city size distribution as a whole as well as on each individual city separately that we subsequently use to distinguish between three competing theories that explain urban growth.

Our first main finding is that (the evolution of) the German city size distribution is permanently affected by the World War II shock, more so than by any other shocks. Cities that have been hit relatively hard due to the substantial bombings and the subsequent allied invasion do not recover the loss in relative size. After the war, the city size distribution does not revert to its pre-WWII level, but shifts to one characterized by a more even distribution of population over the cities in the sample. Compared to the impact of WWII, the separation from and later reunion with East Germany has had much less impact on relative city sizes.

Our second main contribution is that we show that, once corrected for the heavy destruction during WWII, (panel) unit root tests that are used to test for the validity of proportional city growth reject Gibrat's Law of proportional effect for about 75% of all cities. This constitutes considerable evidence against urban economic theories exhibiting random growth, a finding that is further confirmed by additional non-parametric evidence.

Finally, evidence from (panel) unit root tests on relative city size also shed substantial doubt on the relevance of locational fundamental theories in explaining the development of the German urban system, even when allowing these fundamentals to change during WWII. This finding contrasts sharply with Davis and Weinstein (2002), who found that the average Japanese city completely recovered from the destruction suffered during WWII, and subsequently concluded that locational fundamentals were the most likely candidate to explain the observed evolution. Here we find, extending the evidence provided in Bosker et al. (2007a), that the results are most consistent with urban economic theories emphasizing increasing returns to scale. We think that the lesser constraint posed on urban development by Germany's first nature geography compared to that of Japan (see also Head and Mayer, 2004), the fact that the German cities suffered more heavy destruction for a longer period of time compared to their Japanese counterparts[144] and the impact of the unequally distributed

[144] Compare the fact that 87% of the 62 cities in our sample suffered a population loss during WWII to the 80% of the 300 cities <u>not</u> suffering population decline during WWII in Japan (see p.1278 in Davis and Weinstein,

loss of market potential across West German cities due to the split of the country in East and West (see Redding and Sturm, 2005), are the main reasons for this difference. The lesser constraint posed by physical geography on the development of (large) cities in Germany, combined with the more widespread destruction suffered during WWII, left a much more level playing field on which the forces stressed by urban economic theories exhibiting increasing returns to scale could ‘do their work’, changing both the overall shape of the city size distribution as well as the relative size and/or rank of the cities making up this distribution.

Overall, we find the evidence provided by the evolution of the German city size distribution most consistent with theories exhibiting increasing returns to scale. However, we have to note that we did not provide a proper test for theories exhibiting increasing returns to scale, concluding only in favor of them on the basis of the empirical evidence against the other two competing theories of urban development. Developing a proper test of urban growth theories exhibiting increasing returns is in our view a fruitful area of future research that could further substantiate our claim.

2002; compare also Figure 1 in Brakman et al., 2004a to Figure 1 in Davis and Weinstein, 2002). The Allied bombing campaign against German cities started in 1942, whereas Japanese cities could only be reached by US bombers during the last five months of the war in 1945.

APPENDIX 5.A DATA

Table 5.A1 West-German cities in our sample

Berlin West
Hamburg
München
Köln
Lübeck
Essen
Dortmund
Düsseldorf
Stuttgart
Bremen
Duisburg
Hannover
Nürnberg
Bochum
Wuppertal
Bielefeld
Braunschweig
Mönchengladbach-Rheydt
Münster
Augsburg
Frankfurt am Main
Krefeld
Ludwigshafen am Rhein
Oberhausen
Offenbach am Main
Hagen
Kassel
Freiburg im Breisgau
Hamm
Mainz
Herne
Mülheim an der Ruhr
Heidelberg
Würzburg
Recklinghausen
Remscheid
Regensburg
Bottrop
Aachen
Pforzheim
Ulm
Koblenz
Witten
Hildesheim
Fürth
Kaiserslautern
Trier
Wanne-Eickel
Bamberg
Gladbeck
Wattenscheid
Flensburg
Solingen
Osnabrück
Kiel
Oldenburg
Darmstadt
Mannheim
Gelsenkirchen
Bonn
Karlsruhe
Wiesbaden

These cities are included out of a total of 81 cities in the original data set (see Brakman et al. (2004a) for a detailed description of the original data and its sources) on the basis of the availability of **annual** city population data over the period 1925-1999. The cities that were left out of the original data set were dropped on the basis of failing to comply to one (or more) of the following criteria:

I *More than two consecutive years with no population data.*

II *Not able to correct for a so-called Gemeindereform, i.e. local government reorganization, that occurred in the early 1930s for several cities in the industrial Ruhr-area and for most of the sample cities during the 1970s.*

The first exclusion criterion results in about 16 cities to be excluded from the data set. If at most two observations are missing we construct the city population for the missing years by interpolating (such a correction is made only 6 times). The other 3 are excluded based on the second criteria. Most cities in our data set are affected by the Gemeindereform of the 1970s and most cities in the Ruhr-area also by the Gemeindereform in the 1930s. In order to have the same unit of analysis in terms of city boundaries we have decided to take the city boundaries at the time of WWII as a point of reference. For example if due to a local government reorganization an adjacent town becomes part of one of the cities in our sample we extend this city boundary redefinition to all pre-WWII years if this redefinition happened before WWII and we ignore it if it happened after WWII. The fact that for most of the cities in our sample the exact number of people that is added due to a local government reorganization is recorded in the Statistical Yearbooks allows us to correct for this quite accurately.

More formally in case of a pre-WWII Gemeindereform at time *T* in city *i*, we adapt the series as follows,

$$\hat{S}_{iT-k} = S_{iT} \frac{S_{iT-k}}{(S_{iT} - S_{inew})} \tag{5.16}$$

where S_{iT} is the population at time *T* including the newly added towns, S_{inew} is the number of people living in the newly added towns, $S_{iT\text{-}k}$ is the city population as reported in year *T-k*, i.e. before the city boundary redefinition, and $\hat{S}_{iT-k}$ is the newly calculated, as if the new city boundary was already in affect, city population at time *T-k*. If instead a city was subject to a post-WWII Gemeindereform this is incorporated as,

$$\hat{S}_{iT} = S_{iT} - S_{inew} \qquad \text{and} \qquad \hat{S}_{iT+k} = \hat{S}_{iT} \frac{S_{iT+k}}{S_{iT}} \tag{5.17}$$

where S_{iT} and S_{inew} are defined as above and $\hat{S}_{iT}$ ($\hat{S}_{iT+k}$) is the newly calculated, as if city boundary redefinition did not happen, city population at time *T* (*T+k*). Thus the crucial assumption made in case of a pre-WWII Gemeindereform is that the city's and its newly added town's population grew at the same rate before the reform. Similarly in case of a post-WWII Gemeindereform the crucial assumption is that after the Gemeindereform the city's and its newly added town's population grew at the same rate.

APPENDIX 5.B CITY CHARACTERISTICS AND BREAK STATIONARITY

The unit root tests in section 5.6 indicated that 17 cities experienced a one-time substantial impact of WWII on the share of their population in total German population. It turns out that 15 of these 17 cities experienced a **negative** impact of the WWII shock. To give some indication about possible differences between the break-stationary and the nonstationary cities, Table 5.B1 below shows the mean and standard deviation of several characteristics for 1) the 17 break stationary and 2) the 45 nonstationary cities. Also included are the results of a simple test of the equivalence of the means of the two groups given their variances; the final column contains this test when focussing only on the cities with a negative break in their city share. These statistics show some interesting results.

Cities that have experienced a break in the deterministic mean of their city's share in total German population have been hit significantly more severe during the war. The percentage of the housing stock destroyed is 15% higher and the amount of rubble per capita 11 m3 more than for cities with a random evolution of their city share. This shows that these cities have suffered substantially more as a result of the bombardments during the war. Another interesting, possibly related, finding is that after the war these cities have had a lesser inflow of refugees than the cities with a nonstationary city share, possibly reflecting the fact that these refugees did not go to the more heavily destructed cities.

Table 5.B1 City characteristics

	break stationary		nonstationary			neg.break
	mean	s.d.	mean	s.d.	t-test	t-test
War related						
% housing destroyed	50.88	17.56	35.26	16.77	**3.163**	**6.020**
m3 rubble per capita	20.46	6.56	9.81	7.00	**5.480**	**5.256**
reconstruction aid	0.01	0.01	0.01	0.01	0.734	0.887
% refugees 1960	0.15	0.04	0.17	0.04	**-2.043**	**-2.834**
Geography related						
Distance to München (km)	366.24	128.96	419.38	165.19	-0.175	-0.169
Distance to Hamburg (km)	380.82	98.82	329.62	151.59	-0.485	-0.862
Distance to Köln (km)	155.08	93.61	182.87	141.03	0.036	0.083
minimum Distance to East Germany (km)	178.83	65.48	182.96	64.03	-0.686	-1.111
City size related						
pre-WWII population	369250	640162	241725	289447	0.791	0.903
growth pre-WWII	-0.03	0.06	-0.02	0.08	-0.511	-1.055
growth WWII	-0.29	0.20	-0.12	0.18	**-3.027**	**-6.316**
growth post-WWII	0.13	0.20	-0.07	0.22	**3.408**	**3.633**

The city size related characteristics confirm the notion that in those cities with a shift in relative city share more war damage resulted in a larger decline of this relative city share during the war. Mean WWII-growth is significantly less for these cities (-29% vs. -12%). What is interesting is that although post-WWII growth is larger for these same cities, this higher growth does not compensate for the losses suffered during WWII, hereby resulting in a permanent impact on relative city share. Another interesting thing to notice is that pre-WWII city size (1939) does not differ significantly between the two groups of cities[145]. Finally the cities are also compared on the basis of several geography-related characteristics. However distance from the later East-German border or to one of Germany's economic centres (Hamburg, München, or Köln in the industrial Ruhr-area) does not differ significantly between the two groups.

[145] The final column of Table 5.B1 indicates that these results are even more profound when focussing only on the negative city share breaks.

Conclusions

What can be concluded about the empirical relevance of geographical economics on the basis of the results presented in this thesis? I would argue that the overall evidence points towards a positive answer to this question. Relative location, or second nature geography, matters non-trivially for the distribution of economic activity. This thesis shows that it is relevant at various levels of spatial aggregation and at different stages of economic development. Spatial interaction between regions matters; as a result, to ascribe differences in regions' economic development to region-specific characteristics only overlooks the influence that economic developments in neighboring regions can have on (the evolution of) a region's economic own prosperity.

So, space matters, but does it matter in the way specified by new economic geography theory (NEG)? More specifically, are, as emphasized by NEG theory, trade costs a major determinant of the observed spatial distribution of economic activity? Again, my answer would again be largely positive. Despite some of the difficulties involved in taking NEG theory to the data (and the empirical results back to theory) that were discussed primarily in chapters 1 and 2, the results in these two chapters and those in chapter 3 confirm the finding of earlier studies (among others Redding and Venables, 2004; Breinlich, 2006; Hanson, 2005; Brakman et al., 2004), and show that a region's real market access, as specified by NEG theory, significantly influences its economic performance. Regions that have an advantage in terms of their accessibility, will more easily sell their products to other regions, and hereby experience a level of economic prosperity that is higher than relatively more secluded regions. Chapter 1 moreover shows that NEG theory can go beyond its simple two-region, unidimensional second nature geography setup. The cost of losing analytical tractability when considering a multi-region setup with a more realistic depiction of regions' spatial interdependency, does in my view outweigh the disadvantage of being unable to clearly relate empirical results obtained using real-world data back to 'simple' two-region NEG theory. Chapter 1 shows that combining empirical results with careful simulations of the underlying multi-region NEG model can be very fruitful, especially from an empirical and a policy perspective.

Why then do I say that the answer is *largely* positive? Although I strongly believe that the agglomeration and dispersion mechanisms as specified by NEG theory are very important in determining the spatial distribution of economic activity, some caveats are worth mentioning. Chapters 2 and 3 show that, when doing empirical work, the relevance of geographical economics in explaining economic development can easily be overestimated and can also depend quite strongly on how one specifies the spatial linkages between regions.

Specifying regions' spatial interdependencies should in principle not be a problem as NEG quite clearly states that it is trade costs that determine these interdependencies. As discussed in chapter 2, accurate trade cost data are usually not readily available however, so it is up to the empirical researcher to define what he/she believes is a good approximation of the true trade costs between regions. Chapter 2 shows that the choice of approximation is not a trivial one. The empirical example in the chapter shows that the way trade costs are measured nontrivially affects the conclusion regarding the relevance of market access in determining cross-country per capita income levels. The evidence presented in chapter 2 does however not allow for a definitive answer regarding the preferred trade cost approximation. Some recommendations do follow from chapter 2. First, when using a trade cost function to approximate trade costs, the (implicit) assumptions regarding the impact of trade cost related variables can, and should, always be tested. Second, making use of additional information on the strength of regions' spatial interdependencies should always be used if available. The two-step approach, advocated by Redding and Venables (2004), does a much better job at identifying the relevant elements of the trade cost function compared to the direct estimation method introduced by Hanson (2005). The latter has a much harder time allowing for a more elaborate trade cost function[146]. Identifying the parameters on each of the trade cost related variables becomes very problematic given the non-linear estimation procedure on the one hand and the fact that these parameters are identified by functional form assumptions only (no extra information on the strength of spatial linkages is used). The use of implied trade costs, can provide an interesting alternative when using the direct estimation method. Overall, chapter 2 calls for future work to be much more careful in specifying how trade costs, arguably the most important NEG ingredient, are modeled. To this end the chapter offers several guidelines that can help in making an appropriate choice. A choice that will depend quite heavily on the level of spatial aggregation of the sample under consideration (the importance of particular trade cost components likely differs when looking at countries, regions, cities or neighborhoods). The trade cost approximation used should receive much more explicit attention from those doing empirical work in geographical economics.

Chapter 3 also contains an important message regarding the empirical relevance of geographical economics. Besides showing that second nature geography is an important determinant of Sub Saharan Africa's (SSA) dismal economic performance, hereby justifying the focus of many multi- and bilateral aid organizations on improving SSA infrastructure and further strengthening African economic integration, chapter 3 also shows that the relevance of geographical economics is easily overstated for an important reason. One can easily overstate the importance of regions' spatial dependence (or even wrongly conclude that it matters) when not adequately controlling for (unobserved) spatial heterogeneity. That is, when omitting a spatially clustered variable that also determines regions' economic prosperity and that is correlated with real market access[147], the effect of this variable can easily (but wrongly so) be ascribed to the included real market access term derived from NEG-theory. As a result the effect of second nature geography is overestimated. To avoid this issue, I strongly advocate for future studies trying to establish the relevance of geographical economics to

[146] Which is probably one of the reasons why authors using the direct estimation method usually allow only distance and sometimes also a border effect in their trade cost function. Note however also that when considering trade between regions, some trade cost related variables that are important when looking at trade between countries, may not be as relevant (e.g. language or cultural differences, tariffs or institutions) any more.

[147] An example would be education, see chapter 3.

make use of panel data instead of cross-section methods. By allowing for region-specific fixed effects the use of panel data controls for all (unobserved) time-invariant spatial heterogeneity. It does not solve the problems of (unobserved) time-varying (spatially clustered) explanatory variables that can be suspected to influence the result regarding the importance of second nature geography, but the ability to control for fixed spatial heterogeneity gives it a clear advantage over the use of cross-section techniques. In case of SSA, chapter 3 shows that controlling for unobserved fixed spatial heterogeneity significantly reduces the impact of real market access on per capita income levels.

Overall, the evidence provided in Part I of my thesis does in my view strengthen the notion that second nature geography, in the way specified by NEG, is very relevant in explaining the spatial distribution of per capita income levels across countries and/or regions. Market access does matter nontrivially for economic prosperity.

Such an outright positive answer cannot as easily be given when considering the evidence provided in Part II of my thesis. Particularly, claiming that one is really observing NEG at work is much harder on the basis of the evidence provided in these two chapters. The empirical evidence presented in chapters 4 and 5 shows that, when looking at the impact of shocks that affect either trade costs, the mobility of factors of production, or economic activity itself, it is quite difficult to undisputedly claim that one observes NEG at work and not other theories stressing non-pecuniary externalities or even first nature geography in explaining the distribution of economic activity. Chapter 4 shows that trade liberalization and increased labor mobility following the end of Apartheid in South Africa are likely to have resulted in increased economic agglomeration at the regional level and chapter 5 shows that an economic system does not necessarily revert to its ‘old equilibrium’ after having suffered a tremendous shock that literally destroys the spatial equilibrium at the time of the shock. Although consistent with some of NEG’s predictions, these results are also consistent with other theories that explain the distribution of economic activity by other mechanisms than those stressed by NEG.

The main reason why it is so hard to undisputedly establish NEG’s relevance when it comes to the effect of shocks lies partly in the fact that NEG theory does not predict anything about how a spatial economic system *adjusts* to such shocks. It only gives clear predictions about the long run equilibrium the economy will move to after having experienced these shocks. It does not specify in any way how or how quickly, the spatial economy (gradually) moves from one equilibrium to the other. Additional complications arise from the fact that NEG models are typically characterized by multiple equilibria, so that the effect of for example a reduction in trade costs does not result in any change in the spatial equilibrium in some cases whereas it can results in very sudden, very drastic changes in others. These highly non-linear, even non-continuous changes are very hard to model using standard (linear) econometric techniques. The Markov chain, and kernel estimation techniques used in chapters 4 and 5 do provide a natural way to model these non-linearities in the effect of shocks on the spatial distribution of economic activity. They do however have the disadvantage of being unable to clearly say anything about what determines the observed (evolution of the) spatial economic system under consideration. As a result one can only conclude that the results provided in chapter 4 and 5 are not inconsistent with some of NEG's predictions. Although I strongly believe that it is very likely that the mechanisms put forward by NEG are influencing the evolution of regional economic activity in South Africa and that of the urban system in Germany, the results in chapter 4 and 5 (and 4 in particular) only clearly show that space matters; to say that it is truly NEG at work would be merely speculative, though not implausible.

Combining the results of Part I and II, the overall conclusion I would like to draw regarding the empirical relevance of geographical economics is that NEG has rightfully given space its place in mainstream economics. Its main message is that relative location has the potential to nontrivially affect economic development, stressing the importance of taking explicit account of the spatial interaction between economic agents when doing either theoretical, policy related or empirical work. Research providing empirical evidence that shows that second nature geography matters as specified by NEG theory has grown tremendously over the last few years[148]. However, work remains to be done before one can clearly claim that also the *empirical* geographical economics literature has *"become of age"*. Below I discuss several issues that I find especially important avenues for future research.

First, I think empirical work on geographical economics should take the developments in the spatial econometrics literature, see e.g. Anselin (1988, 2003) and Anselin, Rey and Florax (2004), much more seriously (and vice versa). The spatial econometrics literature stresses the endogeneity problems that arise when one for example wants to estimate how a region's income level is affected by that of its neighbors (as is basically done when estimating NEG's wage equation). It proposes useful econometric methods that control for these endogeneity problems. These endogeneity problems have not gone unnoticed in empirical work in geographical economics, but they are typically not solved using the standard spatial econometric techniques. Instead, instrumentation, or replacing market access with either lagged values of real market access or market access at a higher spatial level of aggregation, is used to deal with the endogeneity problem. The reason for the unpopularity of using spatial econometrics, is likely due to the fact that most standard spatial econometric work specifies regions' interdependencies in a very ad hoc manner (as I in effect did in chapter 4), which can be argued to be too simplistic from a geographical economist's point of view, since NEG theory clearly points to the importance of trade costs in determining regions' spatial interactions. I think that both views have a lot to learn from each other. Mion (2004), Behrens, Ertur and Koch (2007), Fingleton (2006; 2007) and Ertur and Koch (2007) show that combining economic theory with spatial econometric theory can be very fruitful. Future work on the empirical relevance of geographical economics is likely to benefit substantially from incorporating insights (and estimation methods) from the spatial econometrics literature (see also Fingleton, 2003, p.205).

Second, future empirical work should focus much more on disentangling the exact ways in which the spatial interactions between regions matter. What we mainly know from

[148] See among others Breinlich (2006), Amiti and Javorcik (2008), Knaap (2006), Brakman, et al. (2006), Amiti and Cameron (2007), Head and Mayer (2004), Hanson, (2005), Brakman et al. (2004), Crozet (2004), and Redding and Venables (2004), and of course the results in this thesis.

the current body of empirical work on geographical economics is that space matters at different levels of spatial aggregation and at different levels of economic development. Opening the black box of why and which aspects of spatial interaction matter, is necessary however to further our understanding of why we observe the clustering of economic activity in some cases and dispersion in others. It will require providing clear empirical evidence on the agglomeration and dispersion mechanisms at work at different levels of spatial agglomeration, in different types of economic activity or at different stages during the economic development process. The recent availability of detailed datasets at the micro-level will greatly help in unraveling these different mechanisms (as already shown in Combes, Duranton and Gobillon, 2008; Combes, Duranton, Gobillon, Puga and Roux, 2007 and Amiti and Cameron, 2007). Using these micro-level data at the firm or household level will hopefully increase our understanding of the exact reasons why firms or people decide to cluster together or not.

Third, and somewhat related to the previous point, I think that when considering the impact of relative location on economic prosperity, we should broaden our view beyond standard economic reasons as to why developments in other regions can be of (tremendous) importance to economic development of a region itself. In particular, political or institutional developments in neighboring regions are very likely to influence what happens in a region itself. The political science literature (see e.g. Simmons and Elkins, 2004) shows that political or institutional developments tend to spillover to neighboring regions. Also civil conflict in neighboring countries has been shown to directly affect a country’s economic development (see e.g. Murdoch and Sandler, 2002); not only through a direct spillover of conflict but also because of refugee flows (Moore and Shellman, 2007), forgone growth opportunities as a result of unproductive investments needed to guard the country against conflict spillovers, decreased trade flows, etc. In a similar vein the concept of neighboring region could also be allowed to mean more than the purely geographical interpretation stressed by geographical economics; sharing common cultural, religious or institutional values are probably at least as important as being geographically close to each other (see Beck, Gleditsch and Beardsley, 2006). Besides being of interest themselves, omitting these (and other not-directly economic) reasons of spatial interaction could very well lead to an overstatement of the relevance of the second nature geography effects emphasized by NEG theory. A first attempt at this, see Bosker and Garretsen (2006), shows that looking at other spatial effects besides those stressed by economic geography, indeed seems to be important and may affect the conclusions drawn about the relevance of space as emphasized by NEG.

Finally, I think that it would be very interesting to look for the importance of geographical economics at different points in history. Was second nature geography as important when the Dutch or the British ruled the world’s oceans as it is now? Did it also influence the development of China, India, the Muslim Empire or Europe before the industrial revolution, and if so how? Or where transport costs simply too high and trade flows too little to have any effect on the spatial distribution of economic activity back then? Some recent papers have started to look at these issues (see e.g. Crafts and Venables, 2003; O’Rourke and Williamson, 2004 and 2008; Bosker, Brakman, De Jong , Garretsen, and Schramm, 2008 and

Bosker, Buringh and van Zanden, 2008). They all show that empirically establishing exactly how and why the nature and extent of regions' spatial interactions have evolved over the course of history is very interesting indeed, but also requires much creativity given the lack of accurate data that can be used to assess these types of questions.

Coming back to the question raised by the satellite picture of the world at night in Figure I.1 in the introduction of my thesis: does geographical economics provide an answer to the question why we observe such an uneven distribution of economic activity at almost every geographical scale? I think it does. It offers very useful, empirically verified, insights in addition to the earlier explanations offered by more traditional 'spatial and non-spatial' economic theories. Economic activity does not take place in isolation. On the contrary, I would argue that in today's globalized economy the spatial interdependencies between countries, regions or even cities are gaining in importance. The effects of this ongoing process can in my view not be studied without taking second nature geography, as stressed by geographical economics, into account. Geographical economics has put geography or space back at the heart of mainstream economic analysis, right where it belongs.

Nederlandse samenvatting

Sommige delen van de wereld zijn duidelijk meer welvarend dan andere. Er bestaan niet alleen grote verschillen in welvaart tussen landen; ook tussen verschillende regio's binnen hetzelfde land, tussen steden binnen dezelfde regio of zelfs tussen wijken in dezelfde stad doen zich vaak grote welvaartsverschillen voor. De geografische economie probeert door middel van theoretisch gefundeerde verklaringen, inzicht te verschaffen in het hoe en waarom van deze ruimtelijke welvaartsverschillen. Hierbij speelt de economische interactie tussen steden, regio's en/of landen een cruciale rol.

Dit proefschrift verifieert de empirische relevantie van dit type theorieën. Het identificeert, bediscussieert en biedt oplossingen voor enkele problemen die zich voordoen bij het empirisch toetsen van de inzichten uit de geografische economie. Vervolgens laat het zien dat de geografische economie een zeer nuttige bijdrage kan leveren aan het beter begrijpen van ruimtelijke welvaartverschillen op diverse geografische schaalniveaus. Dit gebeurt onder meer aan de hand van analyses van de grote verschillen in economisch succes tussen de landen in Sub-Sahara Afrika, van de toename in regionale inkomensongelijkheid in Zuid-Afrika sinds het einde van het Apartheid regime, en van de langetermijneffecten van de geallieerde bombardementen op de ontwikkeling van West-Duitse steden.

Hoofdstuk 1 introduceert de belangrijkste inzichten die voortgekomen zijn uit de geografisch economische theorie aan de hand van een veelgebruikt, veelzijdig model uit deze literatuur (Puga, 1999). Het bespreekt waarom de theorie, omwille van wiskundige elegantie en analytische oplosbaarheid, zich met name toelegt op het analyseren van modellen die slechts twee regio's of landen omvatten. Alhoewel theoretisch zeer goed verdedigbaar omwille van elegantie en oplosbaarheid, loopt elke empirische toepassing meteen tegen deze twee-regio aanname aan. Vrijwel elke empirische studie omhelst namelijk een analyse van de economische gang van zaken in meer dan twee regio's. Dit maakt het erg lastig om sommige resultaten van empirische studies direct terug te relateren aan de theorie, en daarnaast bemoeilijkt het het maken van beleidsaanbevelingen op basis van empirische resultaten die gebaseerd zijn op meer dan twee regio's.

Als oplossing voor dit probleem wordt vervolgens een combinatie van simulatie en empirie aangedragen. Alhoewel de meeste geografisch economische modellen niet analytisch oplosbaar zijn wanneer ze meer dan twee regio's omvatten, is het wel mogelijk om de uitkomsten van een model met meer dan twee regio's te verkrijgen op basis van computersimulaties. Aan de hand van een voorbeeld op basis van de regionale ontwikkeling van de regio's van de vijftien oude lidstaten van de Europese Unie, toont het hoofdstuk aan dat de combinatie van empirie en simulatie zeer vruchtbaar kan zijn en veel nuttiger is dan het relateren van empirische resultaten aan de standaard twee-regio modellen. Ook laat het zien dat de toenemende integratie tussen de landen in de Europese Unie in de nabije toekomst waarschijnlijk eerder tot meer dan tot minder agglomeratie van economische activiteit, en dus een toenemende ongelijkheid tussen regio's, zal leiden.

Hoofdstuk 2 adresseert een andere complicatie die zich voordoet bij het empirisch toetsen van de inzichten uit de geografische economie. In de geografisch economische theorie spelen handelskosten, alle kosten die gemoeid zijn met het verschepen van goederen tussen twee locaties, een cruciale rol. De hoogte ervan bepaalt in grote mate de ruimtelijke verdeling van economische activiteit over de verschillende regio's[149]. Het is dan ook niet verrassend dat handelskosten ook in elke empirische studie op het gebied van de geografische economie een prominente plaats innemen. Echter, accurate data over de omvang van deze handelskosten ontbreken veelal tussen regio's en zelfs tussen landen. Dit vloeit enerzijds voort uit het feit dat zelfs accurate transportkosten of handelstarieven data veelal niet voorhanden zijn, en is anderzijds het gevolg van het feit dat handelskosten ook samenhangen met minder goed meetbare dingen zoals taal- en cultuurovereenkomsten, verschillen in wetgeving, infrastructuur, of in de meer obscure gevallen, corruptie, wegblokkades of oorlogsdreiging. In de empirische geografisch economische literatuur wordt dit gebrek aan accurate handelskosten data ondervangen door gebruik te maken van data die wèl beschikbaar is, maar die slechts een benadering van de echte handelskosten bieden. Voorbeelden hiervan zijn de afstand tussen regio's, de kans dat twee willekeurige mensen uit twee regio's dezelfde taal beheersen, de kwantiteit en kwaliteit van de infrastructuur, lidmaatschap van een vrijhandelszone, et cetera. Maar in welke mate dragen elk van deze zogenaamde proxies bij aan de omvang van de handelskosten tussen regio's? Om dit te bepalen, wordt een handelskostenfunctie verondersteld, een formule die de relatieve belangrijkheid en onderlinge samenhang van elk van de proxies in het verklaren van handelskosten vastlegt. Verschillende handelskostenfuncties zijn totnogtoe gebruikt in de empirische literatuur, echter zonder dat de consequenties van de keuze van de gebruikte handelsfunctie voor de uiteindelijke resultaten voldoende belicht zijn.

Hoofdstuk 2 behandelt juist dit laatste punt. Het bediscussieert de verschillende manieren waarop in de empirische literatuur omgesprongen is met het modelleren van handelskosten en het specificeren van de handelskostenfunctie. Ook introduceert het hoofdstuk een mogelijk alternatieve empirische benadering van handelskosten waarbij de specificatie van de handelskostenfunctie niet nodig is. Tot slot wordt de impact van de keuze van handelskostenbenadering bekeken. Door gebruik te maken van steeds dezelfde dataset en slechts de handelskostenbenadering te variëren, wordt duidelijk aangetoond dat het nogal uit kan maken welke specifieke benadering wordt gekozen. Aan de hand van deze resultaten en de eigenschappen van de verschillende handelskostenbenaderingen worden tot slot enkele aanbevelingen gedaan, die de keuze van een geschikte handelskostenbenadering faciliteren.

[149] Zijn de handelskosten erg hoog, dan loont het niet om alle regio's vanuit één regio te voorzien van goederen en zal elke regio economische activiteit vertonen. Zijn de handelskosten lager, dan is de kans groter dat economische activiteit zich concentreert in één of een paar regio's omdat bedrijven hun producten goedkoop naar andere regio's kunnen transporteren en ondertussen agglomeratievoordelen ondervinden door in dezelfde regio te gaan produceren. Worden de handelskosten zo laag dat het eigenlijk niet meer uitmaakt waar een bedrijf zich vestigt, omdat het toch vrijwel kosteloos consumenten in elke regio van goederen kan voorzien, dan kan de economische activiteit zich wederom uitspreiden over alle regio's vanuit de geagglomereerde regio.

De focus van hoofdstuk 3 ligt in tegenstelling tot hoofdstuk 1 en 2 geheel bij de empirische toepassing van de inzichten uit de geografische economie. Aan de hand van de theorie wordt getoetst in hoeverre de geografische economie een bijdrage kan leveren aan het beter begrijpen van de grote verschillen in levensstandaard tussen de landen in Sub-Sahara Afrika. Gebruikmakend van de handelspatronen van deze landen wordt allereerst een inschatting gemaakt van de mate waarin ieder Sub-Sahara Afrikaans land toegang heeft tot de markten van de andere landen in Sub-Sahara Afrika en tot die in de rest van de wereld. Hierbij spelen de verschillende determinanten van handelskosten, die in hoofdstuk 2 ruim aan bod zijn gekomen, een grote rol. Vervolgens wordt getoetst of, zoals voorspeld door de geografisch economische theorie, die landen die een betere toegang hebben tot de markt in andere landen, het economisch beter doen dan landen die veel minder geïntegreerd zijn in de wereldeconomie. De resultaten laten zien dat, ook nadat gecontroleerd is voor de invloed van andere factoren die de armoede in Sub-Sahara Afrika kunnen verklaren[150], een goede toegang tot de markt van andere landen, en met name ook tot de markten van andere Sub-Sahara Afrikaanse landen, van groot belang is voor de economische ontwikkeling van het subcontinent. Door de expliciete specificatie van de handelskostenfunctie kunnen tevens beleidsaanbevelingen worden gedaan ten aanzien van het verbeteren van de integratie van de Sub-Sahara Afrikaanse landen onderling en van de integratie van deze landen in de wereldeconomie. Met name het voorkomen van burgeroorlogen, het verbeteren van de infrastructuur zowel binnen als tussen landen, en het stimuleren van de Afrikaanse samenwerking op het gebied van handel door het sluiten van vrijhandelsverdragen, dragen zeer positief bij aan het vergroten van de deelname van de Sub-Sahara Afrikaanse landen aan de wereldhandel en hiermee aan het bestrijden van de armoede en de hieruit voortvloeiende socio-economische problemen die veel van deze landen kenmerkt.

In hoofdstuk 4 staat een andere empirische toepassing centraal, namelijk de ontwikkeling van de regionale inkomensongelijkheid in Zuid-Afrika na het einde van het Apartheid regime. Sinds het einde van dit regime in 1994 groeit de Zuid-Afrikaanse economie gestaag, echter niet alle delen van het land profiteren in gelijke mate van deze groei, enkele regio worden zelfs gekenmerkt door een afname van economische activiteit. Dit hoofdstuk kijkt of de geografische economie een nuttige bijdrage kan leveren aan het beter begrijpen van de verschillen in economische ontwikkeling tussen de Zuid-Afrikaanse regio's. Hierbij staan twee kenmerken van het Apartheid regime centraal. De eerste is de internationale sancties die de Zuid-Afrikaanse deelname aan de wereldhandel ernstig belemmerde, en de tweede zijn de grote beperkingen die het overgrote deel van de bevolking werd opgelegd in waar ze mochten wonen en werken. Met het afschaffen van de Apartheid verdwenen zowel de internationale sancties en kan elke Zuid-Afrikaan, ongeacht afkomst, nu gaan en staan waar hij of zij maar wil. De geografische economie doet voorspellingen over de consequentie van deze twee veranderingen voor de ruimtelijke verdeling van economische activiteit. Het openen van een land voor producten uit het buitenland zal met name die regio's gunstig beïnvloeden die het makkelijkst kunnen profiteren van deze nieuwe handelsmogelijkheden. Regio's met een

[150] Zoals bijvoorbeeld het opleidingsniveau van de bevolking, het klimaat, de bevolkingsdichtheid en de aanwezigheid van natuurlijke hulpbronnen.

gunstige ligging, met goede im- en exportfaciliteiten, en met een adequate infrastructuur, profiteren meer dan afgelegen regio's. De toegenomen arbeidsmobiliteit zal vooral de agglomeratie van economische activiteit bevorderen. Mensen (en bedrijven) kunnen nu gaan en staan waar ze willen en zullen daarheen trekken waar ze de beste economische kansen zien. Aangezien dit veelal de gebieden zijn die al grote economische bedrijvigheid kennen, leidt dit vaak tot een verdere leegloop van gebieden die economisch al weinig te bieden hadden, met een toename in de regionale inkomensongelijkheid als resultaat.

Om te zien of deze voorspellingen vanuit de geografisch economische theorie hout snijden, brengt hoofdstuk 3 de ruimtelijke ontwikkeling van de inkomensongelijkheid tussen Zuid-Afrikaanse regio's nauwkeurig in kaart. In tegenstelling tot de eerste drie hoofdstukken gebruikt dit hoofdstuk hierbij geen empirische strategie die nauw aansluit op de theorie. Hiermee verliest het aan de ene kant de directe link met de theorie, aan de andere kant maakt dit het mogelijk om een breder scala aan (ruimtelijk) econometrische technieken op de data los te laten, die het mogelijk maken om een veel nauwkeuriger beschrijving te bieden van de ontwikkeling van de regionale inkomensverschillen. De resultaten laten zien dat de regionale ontwikkelingen in Zuid-Afrika veelal consistent zijn met de voorspellingen die de geografische economie doet. Niet alleen is de regionale inkomensongelijkheid in Zuid-Afrika toegenomen, ook zijn het veelal precies de verwachte regio's die het economisch veel beter doen dan de rest van het land. Juist in de regio's die veruit het grootste deel van Zuid-Afrika's internationale handel voor hun rekening nemen (Durban, Richard's Bay, East London, Port Elizabeth) en de regio's die al een economische voorsprong hadden toen de Apartheid werd afgeschaft (Johannesburg, Pretoria, en Kaapstad) groeit het inkomen per hoofd het snelst. Deze regio's fungeren als locale groeipolen die de economische activiteit in omliggende regio's als het ware opzuigen, precies zoals de geografisch economische theorie voorspelt. Uitzondering op deze regel zijn de regio's die flink gegroeid zijn vanwege de sterk in omvang toegenomen toeristensector en die regio's waar grote voorraden natuurlijke hulpbronnen te vinden zijn. De vooruitzichten voor succesvol actief beleid gericht op het verkleinen van de regionale inkomensverschillen lijken op basis van de resultaten van dit hoofdstuk uiterst somber. Investeringen in de infrastructuur, zowel op het gebied van transport als communicatie, en in het opleidingsniveau van de bevolking in achterblijvende regio's lijken de twee potentieel meest succesvolle strategieën te zijn.

Tot slot wordt in hoofdstuk 5 aan de hand van de ontwikkeling van de grootste West-Duitse steden in de twintigste eeuw, een van de meest saillante voorspellingen van de geografische economie onder de loep genomen. In tegenstelling tot andere dominante theorieën over de ontwikkeling van stedelijke systemen, leidt een grote schok in de onderlinge verdeling van economische activiteit tussen steden volgens de geografische economie niet altijd tot een herstel van het stedelijke systeem van voor de schok. Het feit dat de geografische economie door zulke zogenaamde multiple evenwichten wordt gekenmerkt, heeft de nodige aandacht gekregen in zowel de theoretische als empirische literatuur. In geval van Japan tonen Davis en Weinstein (2002) bijvoorbeeld aan dat zich na de grote verwoestingen tijdens de Tweede Wereldoorlog, vrijwel precies hetzelfde stedelijke systeem ontwikkelde als dat voor de oorlog bestond. Op basis hiervan trekken zij de relevantie van de geografische economie in twijfel.

Hoofdstuk 5 bouwt voort op eerdere studies (Brakman et al., 2004 en Bosker et al., 2007) en verifieert de relevantie van de multiple evenwicht hypothese voor het West-Duitse stedensysteem. De massale bombardementen van de West-Duitse steden tijdens de Tweede Wereldoorlog, verwoestten gemiddeld 40% van de woningvoorraad en leidde tot een gemiddelde bevolkingsafname van ongeveer 15%. Gebruikmakend van jaarlijkse stadspecifieke bevolkingsgegevens vanaf 1925 tot en met 1999 en met behulp van uiteenlopende econometrische technieken, laten de resultaten in dit hoofdstuk zien dat het West-Duitse stedensysteem niet hersteld is in haar hoedanigheid van voor de Tweede Wereldoorlog. De blijvende impact van de hevige verwoestingen in deze periode uit zich met name in een (blijvend) kleiner verschil in bevolking tussen de grote en kleinere West-Duitse steden dan voor de oorlog. In tegenstelling tot Japan kan de multiple evenwicht hypothese op basis van het empirisch bewijs in dit hoofdstuk niet verworpen worden, en het biedt hiermee een steun in de rug van de relevantie van de geografische economie. De belangrijkste verklaringen voor het verschil in conclusie tussen Japan en West-Duitsland met betrekking tot de langetermijn impact van de verwoestingen in de Tweede Wereldoorlog, zijn waarschijnlijk het verschil in fysieke geografie van de twee landen[151], de veel grotere en wijdverspreide verwoesting van de West-Duitse steden, en de additionele impact van de splitsing van Duitsland in Oost en West op het West-Duitse stedensysteem.

Het algehele beeld dat uit de vijf hoofdstukken van dit proefschrift naar voren komt, is dat de geografische economie een grote bijdrage kan leveren aan het beter begrijpen van de verschillen in welvaart tussen zowel landen, regio’s en steden. Zoals benadrukt wordt in de theorie, laat het empirisch bewijs in dit proefschrift zien dat de economische interactie tussen landen of regio’s een niet triviale impact heeft op hun economische ontwikkeling. De grote bijdrage van de geografische economie is dat het onderzoekers, die economische ontwikkeling voorheen louter relateerde aan land- of regiospecifieke factoren, wakker heeft geschud. Economische ontwikkeling vindt niet in isolatie plaats, ontwikkelingen in het ene land kunnen van grote invloed zijn op wat er in andere landen gebeurt. Met de toenemende globalisering wordt de onderlinge afhankelijk tussen landen, regio’s en zelfs steden alleen nog maar groter. Zonder de inzichten uit de geografische economie serieus te nemen, kunnen de effecten van dit proces niet adequaat bestudeerd worden.

[151] West-Duitsland is veel minder bergachtig dan Japan. In Duitsland is de omvang van een stad hierdoor veel minder beperkt dan in Japan.

References

Abreu, M., H. de Groot and R. Florax, 2005. Space and growth: a survey of empirical evidence and methods. *Région et Développement*, 21, p.12-43.

Aghion, P. and P. Howitt, 1998. *Endogenous growth theory*. The MIT Press, Cambridge, Massachusetts.

Amiti, M. and B.S. Javorcik, 2008. Trade costs and location of foreign firms in China. *Journal of Development Economics*, 85, p.129-149.

Amiti, M. and L. Cameron, 2007. Economic geography and wages. *Review of Economics and Statistics*, 89, p.15-29.

Anderson, G. and Y. Ge, 2005. The size distribution of Chinese cities. *Regional Science and Urban Economics*, 35, p.756-776.

Anselin, L., 1988. *Spatial Econometrics: Methods and Models,* Kluwer, Dordrecht.

Anselin, L., 2003. Spatial externalities, spatial multipliers and spatial econometrics. *International Regional Science Review*, 4, p.153-166.

Anselin, L., S.J. Rey and R.J.G.M. Florax, 2004. *Advances in Spatial Econometrics*, Springer-Verlag.

Amjadi, A. and A.J. Yeats, 1995. Have transport costs contributed to the relative decline of Sub-Saharan African exports? Some preliminary evidence. *World Bank Policy Research Paper*, no.1559, World Bank, Washington.

Ancharez, V., 2003. Determinants of trade policy reform in Sub-Saharan Africa. *Journal of African Economics*, 12, p.417-443.

Anderson, J.E. and E. van Wincoop, 2004. Trade Costs. *Journal of Economics Literature*, 42, p.691-751.

Audretsch, D.B. and M.P. Feldman, 1996. R&D spillovers and the geography of innovation and production. *American Economic Review*, 86, p.630-640.

Auerbach, F., 1913. Das Gesetz der Bevölkerungskonzentration, *Petermanns Geographische Mitteilungen*, 59, p.74-76.

Aziz, J. and C.K. Duenwald, 2001. China's provincial growth dynamics. *IMF working paper*, no. 01/3. IMF, Washington.

Baier, S.L. and J.H. Bergstrand, 2001. The growth of world trade: tariffs, transport costs, and income similarity. *Journal of International Economics*, 53, p.1-27.

Baldwin, R., P. Martin and G.I.P. Ottaviano, 2001. Global income divergence, trade and industrialization: the geography of growth take-offs. *Journal of Economic Growth*, 6, p.5-37.

Baldwin, R., R. Forslid, Ph. Martin, G.I.P. Ottaviano and F. Robert-Nicoud, 2003. *Economic Geography and Public Policy,* Princeton University Press, Princeton, New Jersey.

Bandyopadhyay, S., 2004. Twin peaks: distribution dynamics of economic growth across Indian states, in A. Shorrocks and R. van der Hoeven (eds.) *Growth, Inequality and Poverty,* Oxford University Press, Oxford, p.176-197.

Barro, R.J. and X. Sala-i-Martin, 1991. Convergence across states and regions. *Brookings Papers on Economic Activity*, 1, p.107-182.

Bartholomew, D.J., 1981. *Mathematical models in social science*, John Wiley, Chichester, UK.

Bartholomew, D.J., 1982. *Stochastic models for social processes,* John Wiley, Chichester, UK.

Beck, N., K.S. Gleditsch and K. Beardsley, 2006. Space is more than geography: using spatial econometrics in the study of political economy. *International Studies Quarterly*, 50, p.27-44.

Behrens, K., C. Ertur and W. Koch, 2007. 'Dual' gravity: using spatial econometrics to control for multilateral resistance. *CORE discussion paper*, no.2007/59, Louvain.

Behrens, K., A.R. Lamorgese, G.I.P. Ottaviano and T. Tabuchi, 2007. Changes in transport and non-transport costs: local vs global impacts in a spatial network. *Regional Science and Urban Economics*, 37, p.625-648.

Behrens, K. and J.F. Thisse, 2007. Regional economics: A new economic geography perspective. *Regional Science and Urban Economics*, 37, p.457-465.

Bernstein, A. and J. McCarthy, 2005. Thinking big in small-town South Africa. *Business day*, Aug. 5.

Bickenbach, F. and E. Bode, 2003. Evaluating the Markov property in studies of economic convergence. *International Regional Science Review*, 26, p.363-392.

Black, D. and V. Henderson, 2003. Urban evolution in the USA. *Journal of Economic Geography*, 3, p.343-372.

Blonigen, B.A., R.B. Davies, G.R. Davies and H.T. Naughton, 2007. FDI in space: spatial autoregressive in foreign direct investment. *European Economic Review*, 51, p.1303-1325.

Boarnet, M.G., 1998. Spillover and the locational effects of public infrastructure. *Journal of Regional Science*, 38, p.381-400.

Bosker, E.M., 2006. The spatial evolution of regional GDP disparities in the 'old' and the 'new' Europe. *Papers in Regional Science*, forthcoming.

Bosker, E.M., 2007a. Growth, agglomeration and convergence: a space-time analysis for European regions. *Spatial Economic Analysis,* 2, p.91-100.

Bosker, E.M., 2007b. Sub-Saharan Africa's manufacturing trade: trade costs, zeroes, and export orientation. *Working paper*, Utrecht University. Available online at http://maartenbosker.googlepages.com.

Bosker, E.M., 2007c. Black holes and the bell curve. *Working paper*, Utrecht University. Available online at http://maartenbosker.googlepages.com.

Bosker, E.M., S. Brakman, H. Garretsen and M. Schramm, 2007a. Looking for multiple equilibria when geography matters: German city growth and the WWII shock, *Journal of Urban Economics*, 61, p.152-169.

Bosker, E.M., S. Brakman, H. Garretsen and M. Schramm, 2007b. Adding geography to the new economic geography. *Working paper,* Utrecht University. **[chapter 1 of this thesis]**

Bosker, E.M., S. Brakman, H. Garretsen and M. Schramm, 2007c. A century of shocks: the evolution of the German city size distribution 1925-1999. *Regional Science and Urban Economics*, forthcoming. **[chapter 5 of this thesis]**

Bosker, E.M., S. Brakman, H. Garretsen, H. de Jong and M. Schramm, 2008. Ports, plagues and politics: explaining Italian city growth 1300-1861. *European Review of Economic History*, 12, p.97-131.

Bosker, E.M., E. Buringh and J.L. van Zanden, 2008. From Baghdad to London: The dynamics of urban growth in Europe and the Arab world 800-1800. *Working paper,* Utrecht University. Available online at http://maartenbosker.googlepages.com.

Bosker, E.M. and H. Garretsen, 2006. Economic development and the geography of institutions. *Working paper* Utrecht University. Available online at http://maartenbosker.googlepages.com.

Bosker, E.M. and H. Garretsen, 2007a. Economic geography and economic development in Sub-Saharan Africa, *Working paper,* Utrecht University. **[chapter 3 of this thesis]**

Bosker, E.M. and H. Garretsen, 2007b. Trade costs, market access and economic geography: why the empirical specification of trade costs matters. *Working paper,* Utrecht University. **[chapter 2 of this thesis]**

Bosker, E.M. and W.F. Krugell, 2008. Regional income evolution in South Africa after Apartheid. *Journal of Regional Science*, forthcoming. **[chapter 4 of this thesis]**

Brakman, S., H. Garretsen and Ch. van Marrewijk, 2001. *An Introduction to Geographical Economics,* Cambridge University Press, Cambridge, UK.

Brakman, S., H. Garretsen and Ch. van Marrewijk en M. van den Berg, 1999. The return of Zipf: towards a further understanding of the rank-size distribution. *Journal of Regional Science*, 39, p.183-213.

Brakman, S., H. Garretsen and M. Schramm, 2004a. The strategic bombing of German cities during World War II and its impact on city growth. *Journal of Economic Geography*, 4, p.201-218.

Brakman, S., H. Garretsen and M. Schramm, 2004b. The spatial distribution of wages: estimating the Helpman-Hanson model for Germany. *Journal of Regional Science,* 44, p.437-466.

Brakman, S., H. Garretsen and M. Schramm, 2006. Putting new economic geography to the test: Freeness of trade and agglomeration in the EU regions. *Regional Science and Urban Economics*, 36, p.613-635.

Breinlich H., 2006. The spatial income structure in the European Union – what role for Economic Geography? *Journal of Economic Geography*, 6, p.593-617.

Bröcker, J., 1998. How would an EU-membership of the Viségrad countries affect Europe's economic geography? *The Annals of Regional Science,* 32, p.91-114.

Bulli, S., 2001. Distribution dynamics and cross-country convergence: a new approach. *Scottish Journal of Political Economy*, 48, p.226-243.

Buys, P, U. Deichmann and D. Wheeler, 2006. Road network upgrading and overland trade expansion in Sub-Saharan Africa. *World Bank Policy Research Paper*, WPS4097, World Bank, Washington.

Cavaliere, G., 2005. Limited time series with a unit root. *Econometric Theory*, 21, p.907-945.

Cheshire P. and S. Magrini, 2000. Endogenous processes in European regional growth: convergence and policy. *Growth and Change*, 31, p.455-479.

Ciccone, A., 2002. Agglomeration effects in Europe. *European Economic Review*, 46, p.213-227.

Ciccone, A. and R.E. Hall, 1999. Productivity and the density of economic activity. *American Economic Review*, 86, p.54-70.

Clark, S.J. and J.C. Stabler, 1991. Gibrat's Law and the growth of Canadian cities. *Urban Studies*, 28, p.635-639.

Cliff A.D. and J.K.Ord, 1981. *Spatial processes: Models and applications*, Pion, London.

Coe, D.T. and A.W. Hoffmaister, 1999. North-South trade: is Africa unusual? *Journal of African Economics,* 8, p.229-256.

Coe, D.T., A. Subramanian, N.T. Tamirisa and R. Bhavnani, 2002. The missing globalization puzzle. *IMF Working Paper*, No. 02/171, IMF, Washington.

Collier, P., 2002. Primary commodity dependence and Africa's future. *Keynote Speech, World Bank's Annual Conference on Development Economics.*

Collier P. and J.W. Gunning, 1999. Explaining African economic performance. *Journal of Economic Literature,* 37, p.64-111.

Collier P. and A.J. Venables, 2007. Trade preferences and manufacturing export response: lessons from theory and policy. *World Development*, 30, p.1326-1345.

Combes, P-P. and H.G. Overman, 2004. The spatial distribution of economic activities in the EU, in V. Henderson and J-F. Thisse (eds.) *The Handbook of Regional and Urban Economics,* volume IV*,* North Holland, p.2845-2911.

Combes, P-P., G. Duranton and L. Gobillon, 2008. Spatial wage disparities: sorting matters! *Journal of Urban Economics*, forthcoming.

Combes, P-P., G. Duranton and L. Gobillon, D. Puga and S. Roux, 2007. The productivity advantages of large markets: distinguishing agglomeration from firm selection. *Paper presented at the North American Regional Science Conference*, November 9th, Savannah, Georgia.

Combes, P-P., G. Duranton and H. Overman, 2005. Agglomeration and the adjustment of the spatial economy. *Papers in Regional Science*, 84, p.311-349.

Córdoba, J.-C., 2008. On the distribution of city sizes. *Journal of Urban Economics*, 63, p.177-197.

Crafts, N.F.R. and A.J. Venables, 2003. Globalization in history: a geographical perspective, in M.D. Bordo, A.M. Taylor and J.G. Williamson (eds.), *Globalization in Historical Perspective*, The University of Chicago Press.

Crozet, M., 2004. Do migrants follow market potentials? An estimation of a new economic geography model. *Journal of Economic Geography*, 4, p.439-458.

Dall'Erba, S., 2005. Distribution of regional income and regional funds in Europe 1989-1999: an exploratory spatial data analysis. *The Annals of Regional Science,* 39. p.121-148.

Davis, D.R. and D.E. Weinstein, 2002. Bones, bombs and breakpoints: the geography of economic activity. *American Economic Review*, 92, p.1269-1289.

Davis, D.R. and D.E. Weinstein, 2004. A search for multiple equilibria in urban industrial structure. *NBER working paper*, no. 10252.

Démurger, S., 2001. Infrastructure development and economic growth: an explanation for regional disparities in China? *Journal of Comparative Economics*, 29, p.95-117.

Dickey, D.A. and W.A. Fuller, 1981. Likelihood ratio statistics for autoregressive time series with a unit root. *Econometrica*, 49, p.1057-72.

Dobkins, L.H. and Y.M. Ioannides, 2000. Dynamic evolution of the U.S. city size distribution, in: J.M. Huriot and J.F. Thisse (eds) *The Economics of Cities*. Cambridge University Press, Cambridge UK.

Dobkins L.H. and Y.M. Ioannides, 2001. Spatial interactions among US cities. *Regional Science and Urban Economics*, 31, p.701-731.

Eaton, J. and Z. Eckstein, 1997. Cities and growth: theory and evidence from France and Japan. *Regional Science and Urban Economics*, 27, p.443-474.

Eeckhout, J., 2004. Gibrat's Law for all cities. *American Economic Review*, 94, p.1429-51.

Embrechts, P., P. Kluppelberg and T. Mikosch, 1997. *Modelling extremal events for insurance and finance*, Springer, New York.

Ertur, C. and W. Koch, 2007. Growth, technological interdependence and spatial externalities: theory and evidence. *Journal of Applied Econometrics*, 22, p.1033-1062.

Esteban, J.M. and D. Ray, 1994. The measurement of polarization. *Econometrica*, 62, p.819-852.

Fingleton, B., 1997. Specification and testing of Markov chain models: an application to convergence in the European Union. *Oxford Bulletin of Economic and Statistics*, 59, p.385-403.

Fingleton, B., 1999. Estimates of time to economic convergence: an analysis of regions in the European Union. *International Regional Science Review*, 22, p.5-34.

Fingleton, B., 2003. Externalities, economic geography, and spatial econometrics: conceptual and modeling developments. *International Regional Science Review*, 26, p.197-207.

Fingleton, B., 2006. The new economic geography versus urban economics: an evaluation using local wage rates in Great Britain. *Oxford Economic Papers*, 58, p.501-530.

Fingleton, B., 2007. Competing models of global dynamics: evidence from panel models with spatially correlated error components. *Working paper*, Cambridge University.

Fingleton, B. and E. Lopez-Bazo, 2003. Explaining the distribution of manufacturing productivity in the EU regions, in: Fingleton, B. (ed), *European Regional Growth*, Springer-Verlag, Berlin, p.375-409.

Fingleton, B. and E. Lopez-Bazo., 2006. Empirical growth models with spatial effects. *Papers in Regional Science*, 85, p.177- 198.

Fingleton, B. and P. McCann, 2007. Sinking the iceberg? On the treatment of transport costs in new economic geography, in B. Fingleton (ed.), *New Directions in Economic Geography,* Edward Elgar, p.168-204.

Foroutan F. and L. Pritchett, 1993. Intra-Sub-Saharan African trade: is it too little? *Journal of African Economics,* 2, p.74-105.

Forslid, R., J.I. Haaland, K.H. Midelfart-Knarvik and O. Maestad, 2002a. Integration and transition: Scenarios for the location of production and trade in Europe. *The Economics of Transition*, 10, p.93–117.

Forslid, R., J.I. Haaland, K.H. Midelfart-Knarvik, 2002b. A U-shaped Europe? A simulation study of industrial location. *Journal of International Economics,* 57, p.273–297.

Fowler, C.S., 2007. Taking geographical economics out of equilibrium: implications for theory and policy. *Journal of Economic Geography*, 7, p.265-284.

Friedman, M., 1992. Do old fallacies ever die? *Journal of Economic Literature*, 30, p.45-66.

Fujita, M., P.R. Krugman and A.J. Venables, 1999a. *The spatial economy; Cities, Regions, and Interantional Trade*, MIT Press, Cambridge, MA.

Fujita M., P.R. Krugman and T. Mori, 1999b. On the evolution of hierarchical urban systems. *European Economic Review*, 43, p. 209-251.

Fujita, M. and P.R. Krugman, 2004. The new economic geography: Past, present and the future. *Papers in Regional Science*, 83, p.139-164.

Fujita, M. and J-F Thisse, 2002. *Economics of Agglomeration*, Cambridge University Press, Cambridge, UK.

Fujita, M. and T. Mori, 2005. Frontiers of the New Economic Geography. *Papers in Regional Science,* 84, p.377-407.

Gabaix, X, 1999. Zipf's Law for Cities: An Explanation. *Quarterly Journal of Economics*, 114, p.739-767.

Gabaix, X. and R. Ibragimov, 2007. Rank-1/2: a simple way to improve OLS estimation of tail exponents. *NBER technical working paper*, no.342.

Gabaix, X. and Y.M. Ioannides, 2004. The evolution of city size distributions, in: V. Henderson and J.F. Thisse (eds), *Handbook of Regional and Urban Economics*, volume IV, North-Holland.

Gallup, J.L., J.D. Sachs and A.D. Mellinger, 1999. Geography and economic development. *International Regional Science Review*, 22, p.179-232.

Gianetti, M., 2002. The effects of integration on regional disparities: convergence, divergence or both? *European Economic Review*, 46, p.539-567.

Glaeser, E.L., R. La Porta, F. Lopez-de--Silanes and A. Shleifer, 2004. Do institutions cause growth? *Journal of Economic Growth*, 9, p.271-303.

Hallack, J.C., 2006. Product quality and the direction of trade. *Journal of International Economics,* 68, p.238-265.

Hansen, B., 2000. Sample splitting and threshold regression. *Econometrica*, 69, p.575-603.

Hansen, B., 2001. The new econometrics of structural change: dating breaks in U.S. labor productivity. *Journal of Economic perspectives*, 15, p.117-128.

Hanson, G.H., 1998. Regional adjustment to trade liberalization. *Regional Science and Urban Economics*, 28, p.419-444.

Hanson, G.H., 2005. Market potential, increasing returns, and geographic concentration. *Journal of International Economics,* 67, p. 1-24.

Head, K. and Th. Mayer, 2004. The empirics of agglomeration and trade, in V. Henderson and J-F. Thisse (eds.) *The Handbook of Regional and Urban Economics,* volume IV,North Holland, p.2609-2665.

Head, K. and Th. Mayer, 2006. Regional wage and employment responses to market potential in the EU. *Regional Science and Urban Economics,* 36, p. 573-594.

Head, K. and J. Ries, 2001. Increasing returns versus national product differentiation as an explanation for the pattern of U.S.-Canada trade. *American Economic Review,* 91, p. 858-876.

Helliwell J. and A-C Verdier, 2001. Measuring internal trade distances: a new method applied to estimate provincial border effects in Canada. *Canadian Journal of Economics*, 34, p.1024-1041.

Helpman, E., 1998, The size of regions, in D. Pines, E. Sadka and I. Zilcha (eds.), *Topics in Public Economics*, Cambridge University Press, Cambridge, UK.

Helpman, E., M. Melitz and Y. Rubinstein, 2007. Estimating trade flows: trading partners and trading volumes. *NBER working paper*, no.12927.

Henderson, J.V., Z. Shalizi, A.J. Venables, 2001. Geography and development. *Journal of Economic Geography*, 1, p.81-105.

Hering, L. and S. Poncet, 2006. Market access impact on individual wages: evidence from China. *Working paper,* CEPII, Paris.

Hohenberg, P.M., 2004. The historical geography of European cities: an interpretive essay, in: J.V. Henderson and J.F. Thisse (eds), *Handbook of Regional and Urban Economics*, volume IV, North-Holland.

Hu, D., 2002. Trade, rural-urban migration, and regional income disparity in developing countries: a spatial general equilibrium model inspired by the case of China. *Regional Science and Urban Economics*, 32, p.311-338.

Hummels, D., 1999, Have international transportation costs declined?, *manuscript, University of Chicago*.

Hummels, D., 2001. Toward a geography of trade costs, *mimeo,* Purdue University.

Hummels, D., 2007. Transportation costs and international trade in the second era of globalization. *Journal of Economic Perspectives*, 21, p.131-154.

Im, K.S., M.H. Pesaran and Y. Shin, 2003. Testing for unit roots in heterogeneous panels. *Journal of Econometrics*, 115, p.53-74.

IMF, 2007. *Regional Economic Outlook: Sub-Saharan Africa,* Washington.

Ioannides, Y.M. and H.G. Overman, 2003. Zipf's Law for cities: an empirical examination. *Regional Science and Urban Economics*, 33, p.127-137.

Ioannides, Y.M. and H.G. Overman, 2004. Spatial evolution of the US urban system. *Journal of Economic Geography*, 4, p.131-156.

Johnson, P.A., 2000. A nonparametric analysis of income convergence across the US states. *Economics Letters*, 69,p. 219-223.

Knaap. T., 2006. Trade, location, and wages in the United States. *Regional Science and Urban Economics,* 36, p. 595-612.

Krugell, W.F., 2005. *The geographical economy of South Africa*. Unpublished PhD-thesis. North-West University, Potchefstroom, South Africa.

Krugell, W.F. and W.A. Naudé, 2005. The geographical economy of South Africa. *Journal of Development Perspectives*, 1, p.85-128.

Krugell, W.F., G. Koekemoer and J. Allison., 2005. Convergence or divergence of South African cities and towns? Evidence from kernel density estimates. *Paper presented at the Biennial Conference of the Economic Society of South Africa: Development Perspectives: Is Africa Different?* Durban, South Africa.

Krugman, P.R., 1991. Increasing returns and economic geography. *Journal of Political Economy*, 99, p.483-499.

Krugman, P., 1995. *Development, Geography and Economic Theory*, MIT Press, Cambridge, MA.

Krugman, P.R., 1998. What's new about the new economic geography? *Oxford Review of Economic Policy*, 14, p.7-17.

Krugman, P.R. and R. Livas Elizondo, 1996. Trade policy and third world metropolis. *Journal of Development Economics*, 49, p.137-150.

Krugman, P.R. and A.J. Venables, 1995. Globalization and the inequality of nations**,** *The Quarterly Journal of Economics*, 110, p.857-880.

LaFountain, C., 2005. Where do firms locate? Testing competing models of agglomeration. *Journal of Urban Economics*, 58, p.338-366.

Lall, S.V., 2007. Infrastructure and regional growth, growth dynamics and policy relevance for India. *The Annals of Regional Science*, 41, p.581-599.

Lall, S.V. and Z. Shalizi., 2003. Location and growth in the Brazilian northeast. *Journal of Regional Science*, 43, p.663-681.

Lee, K., H. Pesaran and R. Smith., 1998. Growth empirics: a panel data approach – a comment. *The Quarterly Journal of Economics*, 113, p.319-323.

Le Gallo, J., 2004. Space-time analysis of GDP disparities among European regions: a Markov chain approach. *International Regional Science Review*, 27, p.138-163.

Le Gallo, J. and S. Dall'Erba, 2006. Evaluating the temporal and spatial heterogeneity of the European convergence process, 1980-1999. *Journal of Regional Science*, 46, p.269-288.

Levin, A., C.-F. Lin and C.-S.J. Shu, 2002. Unit root tests in panel data: asymptotic and finite-sample properties. *Journal of Econometrics*, 108, p.1-24.

Limao, N. and A.J. Venables, 2001. Infrastructure, geographical disadvantage, transport costs, and trade. *The World Bank Economic Review*, 15, p.451-479.

Longo R. and K. Sekkat, 2004. Economic obstacles to expanding intra-African trade. *World Development,* 32, p.1309-1321.

López-Bazo, E., E. Vayá and M. Artís, 2004. Regional externalities and growth: evidence from European regions. *Journal of Regional Science*, 44, p.43-73.

Magrini, S., 1999. The evolution of income disparities among the regions of the European Union. *Regional Science and Urban Economics*, 29, p.257-281.

Mankiw, M., D. Romer and D. Weil, 1992. A contribution to the empirics of economic growth. *The Quarterly Journal of Economics*, 107, p.407-437.

Mansori, K.S., 2003. The geographic effects of trade liberalization with increasing returns in transportation. *Journal of Regional Science,* 43, p.249-268.

Martin, R., 1999. The new ‘geographical turn’ in economics: some critical reflections. *Cambridge Journal of Economics*, 23, p.63-91.

McCann, P., 2001. A proof of the relationship between optimal vehicle size, haulage length and the structure of distance-transport costs. *Transportation Research*, 35A, p.671 – 693.

McCann, P., 2005. Transport costs and new economic geography. *Journal of Economic Geography*, 5, p.305 - 318.

Melitz, M, 2003. The impact of trade on intra-industry reallocations and aggregate industry productivity. *Econometrica,* 71, p.1695-1725.

Mion, G., 2004. Spatial externalities and empirical analysis: the case of Italy. *Journal of Urban Economics*, 56, p.97-118.

Monfort, P. and R. Nicolini, 2000. Regional convergence and international integration. *Journal of Urban Economics*, 48, p.286-306.

Moore, W.H. and S.M. Shellman, 2007. Whither will they go? A global analysis of refugee flows, 1955-95. *International Studies Quarterly*, 51, p.811-834.

Mossi, M.B., P. Aroca, I.J. Fernández and C.R. Azzoni, 2003. Growth dynamics and space in Brazil. *International Regional Science Review*, 26, p.393-418.

Murdoch, J.C. and T. Sandler, 2002. Economic growth, civil wars, and spatial spillovers. *Journal of Conflict Resolution*, 46, p.91-110.

Naudé, W.A. and W.F. Krugell, 2003. An inquiry into cities and their role in sub-national economic growth in South Africa. *Journal of African Economies*, 12, p.476-499.

Naudé, W.A. and W.F. Krugell, 2006. Economic geography and growth in Africa: the determinants of sub-national growth in South Africa. *Papers in Regional Science,* 85, p.443-457.

Ndulu, B., L. Chakraborti, L. Lijane, V. Ramachandran and J. Wolgin, 2007. *Challenges to African Growth*, World Bank, Washington.

Neary, J.P., 2001. Of hypes and hyperbolas: introducing the new economic geography. *Journal of Economic Literature,* 39, p.536-561.

Nel, E., 2002. South Africa’s manufacturing economy: problems and performance, in A. Lemon and C.M. Rogerson (eds.), *Geography and economy in South Africa and its neighbours*. Aldershot: Ashgate. p.81-94.

Ng, S. and P. Perron, 1995. Unit root tests in ARMA models with data dependent methods for the selection of the truncation lag. *Journal of the American Statistical Association*, 90, p.268-81.

Nitsch, V., 2005. Zipf zipped. *Journal of Urban Economics*, 57, p.86–100.

O’Rourke, K.H. and J.G. Williamson, 2004. From Malthus to Ohlin: trade, industrialisation and distribution since 1500. *Journal of Economic Growth*, 10, p.5-34.

O’Rourke, K.H. and J.G. Williamson, 2008. Did Vasco da Gama matter to European markets? Testing Frederick Lane’s hypothesis fifty years later. *CEPR discussion paper*, no.5418, London.

Ottaviano, G.I.P. and J.F. Thisse, 2004. Agglomeration and economic geography, in V. Henderson and J-F. Thisse (eds.) *The Handbook of Regional and Urban Economics,* volume IV*,* North Holland, p. 2563-2608.

Ottaviano, G.I.P., T. Tabuchi and J.F. Thisse, 2002. Agglomeration and trade revisited. *International Economic Review*, 43, p.409-435.

Overman, H.G. and Y.M. Ioannides, 2001. Cross-sectional evolution of the U.S. city size distribution. *Journal of Urban Economics*, 49, p.543-566.

Overman, H.G., S.J. Redding and A.J. Venables, 2003. The economic geography of trade, production and income: a survey of empirics, in: K. Choi and J. Harrigan (eds), *Handbook of International Trade*, Basil Blackwell, Oxford. p.353-387

Perron, P., 1989. The Great Crash, the oil price shock, and the unit root hypothesis. *Econometrica*, 57, p.1361-1401.

Perron, P., 1997. Further evidence on breaking trend functions in macroeconomic variables. *Journal of Econometrics*, 80, p.355-385.

Perron, P. and T.J. Vogelsang, 1992. Testing for a unit root in time series with a changing mean: corrections and extensions. *Journal of Business & Economic Statistics*, 10, p.467-472.

Pflüger, M., 2004. A simple, analytically solvable, Chamberlinian agglomeration model. *Regional Science and Urban Economics*, 34, p.565-573.

Poirson, H, 2001. The impact of intersectoral labour reallocation on economic growth. *Journal of African Economics,* 10, p. 37-63.

Pons, J., E. Paluzie, J. Silvestre and D.A. Tirado, 2007. Testing the new economic geography: migrations and industrial agglomerations in Spain, *Journal of Regional Science*, 47, p.289-313.

Porter, M., 2003. The economic performance of regions. *Regional Studies*, 37, p.549-578.

Puga, D. and A.J. Venables, 1996. The spread of industry: spatial agglomeration in economic development. *Journal of the Japanese and International Economics,* 10, p.440-464.

Puga, D., 1999. The rise and fall of regional inequalities – spatial agglomeration in economic development. *European Economic Review*, 43, p.303-334.

Quah, D., 1993a. Empirical cross-section dynamics in economic growth. *European Economic Review,* 37, p.426-434.

Quah, D., 1993b. Galton's fallacy and tests of the convergence hypothesis. *Scandinavian Journal of Economics*, 95, p.427-443.

Quah, D., 1996a. Empirics for growth and convergence. *European Economic Review,* 40, p.1353-1375.

Quah, D., 1996b. Regional convergence clusters across Europe. *European Economic Review,* 40, p.951-958.

Quah, D., 1997. Empirics for growth and distribution: polarization, stratification and convergence clubs. *Journal of Economic Growth*, 2, p.27-59.

Ravallion, M., 1996. Issues in measuring and modeling poverty. *Economic Journal*, 106, p.1328-1343.

Redding, S. and P.K. Schott, 2003. Distance, skill deepening and development: will peripheral countries ever get rich? *Journal of Development Economics,* 72, p.515-541.

Redding, S. and D.M. Sturm, 2005. Costs of remoteness: evidence from German division and reunification. *CEPR discussion paper*, no. 5015, London.

Redding, S., D.M. Sturm and N. Wolf, 2007. History and industry location: evidence from German airports. *CEP discussion paper*, no.809, London.

Redding, S. and A.J. Venables, 2004. Economic geography and international inequality. *Journal of International Economics,* 62, p.53-82.

Rey, S.J., 2001. Spatial empirics for economic growth and convergence. *Geographical Analysis*, 33, p.195-214.

Rey, S.J. and M.V. Janikas, 2005. Regional convergence, inequality and space. *Journal of Economic Geography*, 5, p.155-176.

Rey, S.J. and B.D. Montouri, 1999. US regional income convergence: a spatial econometric perspective. *Regional Studies*, 33, p.143-156.

Robert-Nicoud, F., 2004. The structure of simple 'New Economic Geography' models. *CEPR Discussion Paper,* no. 4326. London.

Rodrik, D., A. Subramanian and F. Trebbi, 2004. Institutions rule: the primacy of institutions and integration in economic development. *Journal of Economic Growth*, 9, p.131-165.

Rogerson, C.M., 1991. Beyond racial Fordism: restructuring industry in the "new" South Africa. *Tijdschrift voor economische en sociale geografie*, 82, p.355-366.

Rosenthal, S. and W.C. Strange, 2001. The determinants of agglomeration. *Journal of Urban Economics*, 50, p.191-229.

Rossi-Hansberg, E. and M.L.J. Wright, 2007. Urban structure and economic growth. *Review of Economic Studies*, 74, p.597-624.

Sachs, J.D. and A.M. Warner, 2001. Natural resources and economic development. *European Economic Review*, 45, p.827-838.

Sahn, D.E. and C. Stifel, 2003. Urban-rural inequality in living standards in Africa. *Journal of African Economics,* 12, p.564-597.

Santos Silva, J.M.C. and S. Tenreyro, 2006. The log of gravity. *The Review of Economics and Statistics*, 88, p.641-658.

Sharma, S., 2003. Persistence and stability in city growth. *Journal of Urban Economics*, 53, p.300-320.

Shorrocks, A.F., 1978. The measurement of mobility. *Econometrica*, 46, p.1013-1024.

Silverman, B.W., 1986. *Density estimation for statistics and data analysis*. Chapman and Hall, London.

Simmons, B.A. and Z. Elkins, 2004. The globalization of liberalization: policy diffusion in the international political economy. *American Political Science Review*, 98, p.171-189.

Simon, H., 1955. On a class of skew distribution functions. *Biometrika*, 44, p.425-440.

Solow, R.M., 1956. A contribution to the theory of economic growth. *The Quarterly Journal of Economics,* 70, p.65-94.

Soo, K.T., 2005. Zipf's Law for cities: a cross-country investigation. *Regional Science* and Urban Economics, 35, p.239-263.

Stelder, D., 2005. Where Do Cities Form? A Geographical Agglomeration Model for Europe*, Journal of Regional Science,* 45, p.657-679.

Subramanian, A. and N.T. Tamirisa, 2003. Is Africa integrated in the global economy? *IMF Staff Papers,* 50, IMF, Washington.

Sutton, J, 1997. Gibrat's legacy. *Journal of Economic Literature*, 35, p. 40-59.

Temple, J. and L. Wössmann, 2006. Dualism and cross-country growth regressions. *Journal of Economic Growth,* 11, p.187-228.

Tianlun, J., J.D. Sachs and A.M. Warner., 1996. Trends in regional inequality in China. *China Economic Review*, 7, p.1-21.

Venables, A.J., 1996. Equilibrium locations of vertically linked industries. *International Economic Review,* 37, p.341-359.

Wooldridge, J.M., 2003. *Introductory Econometrics-A Modern Approach,* Thomson, USA .

World Bank, 2007. *Accelerating Development Outcomes in Africa-Progress and Change in the Africa Action Plan*, Washington.

Ying, L.G., 2000. Measuring the spillover effects: some Chinese evidence. *Papers in Regional Science*, 79, p.75-89.

Zipf, G.K., 1949. *Human behavior and the principle of least effort*, Addison-Wesley, Cambridge, MA.