Essays in International Trade and Economic Geography

Jan David Bakker

University College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Trinity 2019

To Anne and Karl, because I owe it all to them.

To Lu, because she is the reason I get up in the morning.

Acknowledgements

I am grateful to my supervisors Ferdinand Rauch and Tony Venables for their guidance and support throughout this work. I thank the members of the International Trade Research Group at the University of Oxford for their support and many insipiring conversations over the last three years.

My gratitude goes to all the colleagues at the Centre for Economic Performance, who have made it a second home as well as a stimulating academic environment.

I am grateful to Maria Balgova, Greg Curello, Nikhil Datta, Elsa Leromain, Josh de Lyon, Niclas Moneke, Wolfgang Ridinger and Ben Waltmann for their insightful comments, conversations and companionship. It has been a great privilege to work with and learn from Kalina Manova and Steve Pischke. Leander Heldring, Chris Roth, Daniela Scur and Johannes Wohlfart led the way and I am grateful for the insights they shared.

For financial support I am indebted to Kalina Manova, Ferdinand Rauch, the Economic and Social Research Council, the Department of Economics at the University of Oxford and University College Oxford.

I am particularly grateful to Lu Liu, without her this thesis would not have seen the light of day.

Abstract

This thesis studies how the spatial distribution of economic activity matters for the transmission of, and is in turn affected by different shocks.

In Chapter 2, I show that the agglomeration-congestion cost trade-off studied by urban economists alters both the distributional effects of, and gains from trade. Correspondingly, I demonstrate that openness to trade affects the spatial distribution of economic activity across cities of different densities. First, I show that the export intensity of firms and sectors is higher in denser locations. I propose an open economy economic geography model that rationalizes these stylized facts. When testing the underlying mechanisms proposed by the model, I find that the higher export intensity in denser places is driven by differences in productivity across firms, differences in factor intensities across sectors and trade-specific gains from agglomeration that lead to lower variable export cost in denser places. In line with the model predictions, I further find that a decrease in trade cost, proxied by exogenous changes in export market access, leads to a reallocation of economic activity to denser places. In currently on-going work I am exploring whether ignoring the increase in aggregate congestion costs from the rise in spatial concentration leads to an overestimation of the gains from trade.

In Chapter 3 I study how an exogenous population shock affects the spatial distribution of population. To this end I exploit the end of movement restrictions for black South Africans at the end of Apartheid to generate an exogenous population shock. Until 1991, black South Africans were severely restricted in their location choices and many were forced to live in homelands. Following the abolition of apartheid, they were free to migrate. Given the general gravity structure of migration, a town closer to the homelands is expected to receive a larger inflow of migrants than towns further away. Using this exogenous variation, I find that on average there is no endogenous reallocation of population following this exogenous immigration. When separately looking at rural and urban places, I find that there is a displacement of incumbents in the former while there are additional inflows in urban areas.

Chapter 4 provides evidence for the importance of trading opportunities for settlement location during the Iron Age. It finds that regions in the Mediterranean that are better connected across the open sea have more economic activity proxied by the number of archaeological sites. This correlation becomes particularly strong around 750 BC when the Phoenicians started to regularly cross the open sea. Analysis at the global level corroborates the finding that more connected locations have a higher population density in 1 AD. Overall, I find that trade and migration shocks have important effects on the long-run distribution of population and economic activity in space.

Contents

1 Introduction 1

2 International Trade and the Comparative Advantage of Cities 7

2.1 Introduction 10

2.2 Related literature 14

2.3 Data and stylized facts 16

2.4 Theory 18

2.5 Regression analysis 31

2.6 Evidence on mechanisms 35

2.7 Additional evidence from Chinese import competition in the US . . 42

2.8 Conclusion 49

Appendices 53

2.A Additional specifications 53

2.B Theory 66

3 Internal Migration and the Growth of Cities in Post-Apartheid South Africa 75

3.1 Introduction 78

3.2 Historical background 80

3.3 Related literature 86

3.4 Empirical analysis 93

3.5 Results 99

3.6 Conclusion 113

Appendices 115

3.A Additional specifications 115

4 Trading Opportunities and Settlement Formation in the Iron Age121

4.1 Introduction 124

4.2 Brief history of ancient seafaring in the Mediterranean 129

4.3 Data and key variables 136

4.4 Specification and results 143

4.5 Conclusion 156

Appendices **157**

4.A Additional specifications . . . 157

4.B Coding of Whitehouse sites . . . 163

5 Conclusion **171**

Bibliography . . . 173

One day I will find the right words, and they will be simple.

— Jack Kerouac

1 Introduction

The spatial distribution of economic activity within countries is highly uneven in space. Locations with very high densities of employment like Manhattan co-exist with vast sparsely populated areas such as large parts of the American Midwest. Cities, from London and New York to Lagos and Shanghai, are at the heart of today's integrated world economy. More than half of the global population lives in cities and produces more than 80% of global GDP. And the importance of cities is growing, the United Nations predict that by 2050 the urban population is going to double in size and almost 70% of the world's population is projected to live in cities. Concurrently, academic work on the spatial distribution of economic activity and economic geography has flourished and furthered our understanding on the determinants and implications of the spatial economy.

Over the last couple of decades, we have observed significant changes in the distribution of economic activity. Advanced economies have been subject to "regional divergence" with less skill-intensive place suffering from lower wage growth (Giannone, 2017). In developing countries, we have witnessed an unparalleled speed of urbanization. Both trends have created important policy questions. What drives urbanization in developing countries and how can it be affected by policies? And similarly, what are the underlying drivers of regional divergence in advanced economies, should policy try to influence these forces, and if so, how? These questions also matter for political economy reasons, as the support for populist and protectionist policies in advanced economies (e.g. Brexit in the UK, Trump in the US and the Front National in France) tends to be strongest in more rural and less educated regions. These developments have already manifested in a renewed interest in policy initiatives that promote industrial policy with a regional focus. The UK government promotes the "Northern Powerhouse" initiative to stimulate economic activity in Northern England to effectively relocate economic activity from London. The German government just proposed their "Plan for Germany" that aims to create "equivalent living conditions across cities and the countryside". An effective implementation of such place-based policies requires a deep understanding of the underlying drivers of the spatial distribution of economic activity.

This thesis aims to contribute to our understanding of what drives the development of the urban system and how it reacts to different shocks, in particular how it is affected by increased (international) economic integration. It adds to a growing literature in economics that treats the "local labour market" as the unit of analysis and allows for systematic differences such as heterogeneous technology across locations within a country, rather than focusing at the country level as the unit of analysis. This approach is able to take into account the persistent regional differences across locations and allows for heterogeneous effects of aggregate shocks across locations. In particular, I examine how the response to an aggregate shock such as a fall in trade cost varies across smaller towns and larger cities. This work aims to deepen our understanding of recent developments in the urban systems across developed and developing countries.

Chapter 2 explores whether increased global economic integration might have contributed to the regional divergence observed in advanced economies and whether accounting for the spatial economy matters for the distributional impact of and aggregate gains from trade. Chapter 3 adds to our understanding of urbanization dynamics in developing countries and to what extend they can be altered by policy. Chapter 4 looks further back into history to provide evidence to what extent trading opportunities mattered for the location of settlements and economic activity during the Iron Age.

Chapter 2 studies the interaction between cities and international trade. It provides evidence that the agglomeration-congestion cost trade-off studied by urban economists matters for the distributional effects of and the welfare gains from trade. Similarly, the openness to trade affects the spatial distribution of economic activity across cities of different employment densities. It presents evidence that the local economy in denser cities is more export intensive and that these differences in export intensity are driven by differences across firms within industries and differences at the industry level. To rationalize these stylized facts, I develop an open economy economic geography model with firm and sector heterogeneity. The model suggests three potential mechanisms for the differences in export intensity: Differences in productivity across firms within a sector, differences in factor intensity across sectors, and differences in the cost of exporting across locations due to trade cost-specific gains from agglomeration. Using French firm-level data, I find that differences in skill intensity can account for the difference in export intensity between sectors located in more and less dense locations. I find that differences in firm productivity matter for the differences in export intensity within sectors across locations, but can only account for part of the variation. Using the model structure, I find that variable trade cost decrease with city density, providing evidence of trade cost-specific gains from agglomeration that outweigh the congestion costs from higher density. I do not find evidence of gains from agglomeration that decrease the fixed costs of exporting, which tend to increase with city density. The model also generates predictions for the effects of a decrease in trade cost on the distribution of economic activity across city densities. In line with the model prediction I find that both within and across sector reallocation of economic activity, following an increase in exogenous export market access, lead to a spatial reallocation of economic activity towards denser places. This suggests that the increased integration of the global economy since 1990 has relocated economic activity to denser places and increased the spatial concentration in advanced economies. In the presence of spatial frictions, this implies differential welfare effects from globalization across locations in the short-run and might make it rational for voters in less dense locations to support protectionist policies. In currently on-going work I am also exploring to what extent this spatial reallocation matters for our measurement of the gains from trade. Part of the gains from trade are driven by a reallocation from less to more productive firms within industries. If this reallocation implies a relocation of economic activity to denser places it also implies that the aggregate congestion costs in the economy are going to increase. Standard models of international trade with heterogeneous firms (e.g. Melitz (2003)) ignore this increase in the cost of congestion, which could imply that they overestimate the gains from trade. In on-going work I explore whether the welfare formula derived by Arkolakis et al. (2012) that applies to the Melitz (2003) model might have to be adjusted by this increase in congestion cost once the spatial dimension of reallocation is taken into account.

Chapter 3 studies the effect of (internal) migration on cities and the urban system. The existing literature studies how cities react to shocks, and shows that cities tend to recover quickly from negative shocks, and that the urban system exhibits some path dependence (e.g. Davis and Weinstein (2002), Bleakley and Lin (2012) and Michaels and Rauch (2018)). Building on this literature, this chapter uses the end of apartheid in South Africa to study how the urban system in a developing country reacts to an exogenous positive population shock. If an exogenous population shock for some cities over time dissipates across the network this is evidence that the urban system behaves like an optimal network of relative city sizes. If there is no further relocation of population then this is evidence of path dependence. If the initial shock leads to further inflows this is evidence of multiple equilibria within the urban system. Empirically, in the case of South Africa, there is no reallocation of population following the initial shock in line with a path dependent urban system. However, there are differences in rural and urban areas. In rural areas there is evidence of displacement of incumbents, while in urban areas there is evidence of further agglomeration, i.e. additional population inflows following the initial shock.

This result has important policy implications as it suggests that temporary shocks have a permanent effect on the distribution of population in space, hence temporary policies are also likely to have permanent effects. It provides support that policies that induce migration will induce urbanization.

Chapter 4 revisits the interaction between trading opportunities and city formation in a different context. It provides evidence that trading opportunities – defined broadly as the ability to exchange goods, people, and ideas – mattered already for settlement location during the Iron Age. This chapter builds on a large dataset on archaeological sites around the Mediterranean starting in 3000BC. It shows that the connectedness of a location by sea within the Mediterranean, a proxy for the ability to trade via sea, affects the probability of the presence of a historic site, which proxies for economic activity. This correlation starts out weak and noisy and becomes strongest around 750BC when the Phoenicians arrive at the Mediterranean and start systematically crossing the open sea. It also provides evidence for the positive correlation between the sea connectedness and population density in 1 AD at the global level. Hence this chapter provides causal evidence on the effect of trading opportunities, broadly defined, on the growth and location of economic activity.

Overall, I find that trading opportunities and migration shocks have important effects on the spatial distribution of economic activity and that accounting for the explicit geography of countries and the insights from urban economics matter for our understanding of the effects of international economic integration.

2
International Trade and the Comparative Advantage of Cities

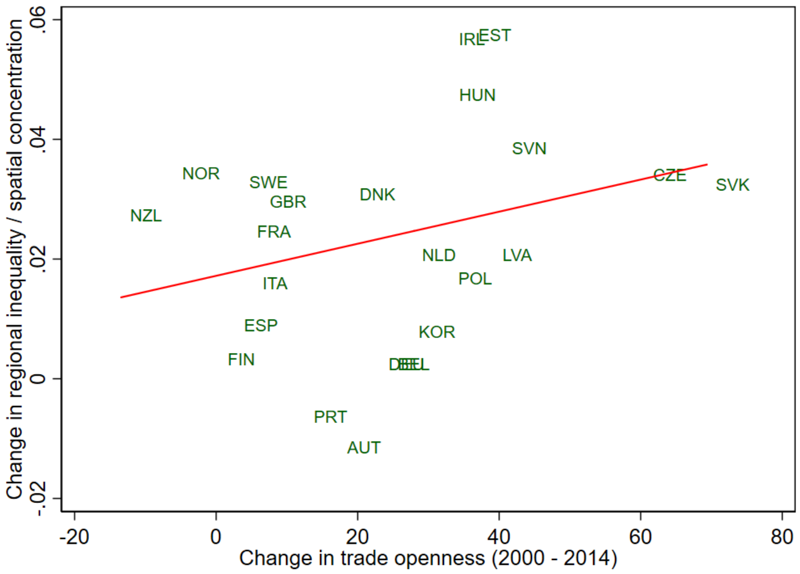
Abstract

This paper shows that openness to trade affects the spatial distribution of economic activity across cities of different densities. It suggests that the agglomeration-congestion cost trade-off studied by urban economists alters both distributional effects of, and gains from trade. I first present three stylized facts documenting that export intensity is higher in more densely populated areas. Guided by these facts, I build an open economy economic geography model with heterogeneous firms and sectors. The model yields two comparative statics with respect to a reduction in trade costs: Firms in larger cities are more productive and will expand due to selection into exporting; and sectors located in larger cities are less low-skill intensive and will expand in capital and skill-abundant advanced economies due to comparative advantage. I test these model predictions using exogenous changes in export market access for French firms and the rise in Chinese import competition in the US and find strong empirical support for both channels. I further find support for the underlying mechanisms suggested by the model: Including skill-intensity accounts well for the reduced-form correlation between city density and export intensity across sectors; and firm productivity accounts for part of the correlation between density and export intensity within sectors. Using the model structure, I show that the residual correlation can be explained by a negative correlation between the variable cost of exporting and city density within sectors, which provides evidence for trade-specific gains from agglomeration. These findings have potential implications for policy and welfare. Increasing openness to trade reallocates employment and economic activity to denser places, which has distributional implications across space. This reallocation also affects the average congestion cost in the economy with potential implications for the aggregate welfare gains from trade.

2.1 Introduction

The distributional effects of globalisation have recently come into renewed public focus. While the effects of international trade on inequality across heterogeneous workers have been studied extensively (Helpman, 2016), much less is known about the effect on heterogeneous regions.[1] Are metropolitan areas like New York City differently affected by trade than countryside towns like Grand Rapids, Michigan, and if so, what are the underlying economic mechanisms? The positive cross-country correlation between changes in openness to trade and regional inequality[2] presented in Figure 2.1 suggests a differential effect. Across countries, an increase in openness to trade is associated with an increase in the concentration of economic activity in bigger cities.

Figure 2.1: Trade openness and regional inequality across countries



Note: Change in trade openness and change in regional inequality between 2000 and 2014 for 26 advanced economies. Change in openness is defined as the change in (exports + imports)/GDP. Change in regional inequality is defined as the change in the regional Gini coefficient, which measures the degree of inequality in the distribution of economic activity in space. *Source*: OECD Regions and Cities database.

[1] The study by Brülhart et al. (fortchoming) is a notable exception.

[2] The extent to which economic activity is distributed unequally in space.

Starting from this cross-country correlation this paper proceeds in four steps to provide evidence linking globalisation to a reallocation of economic activity to larger cities, and a structural interpretation of the mechanisms at play. First, I present three related stylized facts documenting that export participation is higher in more densely populated areas.[3] Second, I develop an open economy economic geography model that rationalizes the cross-country correlation as well as the documented stylized facts and proposes two channels through which changes in trade openness affect regional inequality. Third, I employ exogenous changes in export market access to test the channels proposed by the model using French micro-data. I provide additional evidence from the rise in Chinese import competition in the US studied extensively by Autor et al. (2013) and others. Fourth, I provide evidence on the economic mechanisms that underly the differential export intensity, namely differences in sectoral factor intensity, firm productivity, and variable and fixed trade cost across cities of different densities.

I document in the cross-section of French commuting zones that export intensity is higher in more dense areas (stylized fact 1), and decompose it across two dimensions. Export intensity increases with location density within four-digit industries (stylized fact 2). I find that average sectoral export intensity also increases with employment density (stylized fact 3). This suggests a role for both firm and sector heterogeneity in understanding the heterogeneous trade intensity and its implications across locations.

To rationalize these stylized facts, I propose a firm-level and an industry-level channel, both of which capture important features emphasized in the international trade and urban economics literatures. The firm-level channel builds on recent research in urban economics by Combes et al. (2012) and Gaubert (2018) who provide evidence that, within narrowly defined industries, firms in larger cities are more productive. Research in international trade has shown that opening up to trade leads to a reallocation of market share from less to more productive firms within industries (Pavcnik, 2002, Melitz, 2003). Jointly, these two findings suggest

[3] Throughout the text I will use the terms city size and density interchangeably.

that a given reduction in trade cost translates into a heterogeneous local labour demand shock across different city sizes. Smaller cities host less productive firms that are affected more negatively by a given trade-cost reduction and therefore the city faces a more negative labour demand shock in this sector. This will reallocate employment from smaller to larger cities, such that a decrease in trade costs increases the spatial concentration of economic activity in the long-run and has heterogeneous employment and welfare effects across locations in the short-run.

The industry-level channel builds on recent work by Davis and Dingel (2015) and Gaubert (2018), who provide evidence of systematic spatial sorting of heterogeneous sectors. They find that more skill and more capital-intensive sectors are over-proportionally located in larger cities. Theories of endowment-driven comparative advantage in international trade emphasize trade-induced across-industry reallocation to capital and skill-intensive industries in countries that are abundant in these factors, i.e. in advanced economies. Combining these stylized facts suggests that a reduction in trade costs has a differential effect on the sectors that are located in smaller relative to those in larger cities. Smaller cities host sectors that are more exposed to import competition while larger cities host those that are more exposed to an export opportunity shock from trade opening. Therefore, employment and economic activity will reallocate from those sectors located in smaller cities to those located in larger cities, such that a reduction in trade costs increases the spatial concentration of economic activity in the long-run and implies heterogeneous employment and welfare effects across locations in the short-run.

I formalize this intuition by integrating the multi-sector spatial general equilibrium model from Gaubert (2018) with the international trade model by Bernard et al. (2007) to open a rich economic geography to international trade. The spatial equilibrium of the model features spatial sorting of more productive firms and more capital-intensive sectors into larger cities. In the open economy equilibrium with asymmetric countries, trade occurs both across industries driven by comparative advantage, and within industries driven by firm heterogeneity and love-for-variety utility functions. I study different versions of the model to highlight the effect of the firm-based and the industry-based channel separately. In line with the stylized facts outlined above, in the open economy equilibrium the export intensity is higher in denser places. In line with the second stylized fact the export intensity is higher in denser places without any across-industry heterogeneity. Similarly, the export intensity is higher in denser places in a version of the model that only features across-industry heterogeneity in terms of factor intensities, in line with the third stylized fact. The model also yields predictions for a reduction in trade cost or a switch from the closed to the open economy equilibrium. In a version of the model with symmetric countries and therefore only within-industry trade, the city size distribution in the open economy is more concentrated than in the closed economy in line with the firm-based channel outlined above. In a version of the model that only features two sectors that vary in their factor intensity and homogeneous firms, the city size distribution of the country that is more capital abundant is more concentrated in the open than in the closed economy as suggested by the industry-level channel.

I validate the model predictions empirically using exogenous changes in market access (following Redding and Venables (2004) and Hering and Poncet (2010)) and French micro-data, as well as the rise in Chinese import competition in the United States following Autor et al. (2013) and Acemoglu et al. (2016). In the empirical analysis I rely heavily on the model structure that implies that city size is a sufficient statistic for both the distribution of firms across different cities within a sector as well as the sectoral composition. I find strong support for the model predictions using the regressions implied by the model structure. Consistent with the firm-level mechanism, I show that conditional on the size of the aggregate trade shock the firms located in larger cities increase their revenue by more from a market access shock in France, and employment decreases by less from an import competition shock in the US. Consistent with the industry-level mechanism, I find that the industries located in larger cities respond more to an export opportunity shock and less to an import competition shock.

I also explore the underlying economic mechanisms that link employment density and export intensity in the model. Sectoral skill intensity accounts well for the correlation of density and sectoral export intensity. Firm productivity reduces but not fully account for the reduced-form correlation between firm export intensity and density. Using the model structure, I show that the remaining correlation is driven by variable trade costs that decrease with density, while fixed costs of exporting increase with density.

The remainder of this paper is organized as follows. Section 2.2 discusses the related literature and the contribution of this paper. Section 2.3 introduces the data and the stylized facts. In section 2.4, I describe the model that rationalizes the stylized facts and underlies the empirical analysis presented in section 2.5. Section 2.6 provides evidence on the underlying micro-mechanisms suggested by the model. In section 2.7, I provide additional evidence from the rise of Chinese import competition in the US. Section 2.8 concludes.

2.2 Related literature

This paper studies the interaction between openness to trade and the agglomeration-congestion cost trade-off, and its implication for the gains from trade and the spatial distribution of economic activity. Most closely related to this paper is recent work by Brülhart et al. (fortchoming) that studies the heterogeneous effects of trade on different town sizes in Austria after the fall of the Iron Curtain. They find that larger towns tend to have larger wage and smaller employment responses than smaller towns and argue that this is driven by heterogeneity in the labour supply elasticity across different city sizes. While the focus on the heterogeneity across different city sizes is similar, the papers are very complementary. Brülhart et al. (fortchoming) focus on the heterogeneity in the labour supply elasticity across cities while I focus on heterogeneity on the demand side. They explicitly do not consider the endogenous sorting of sectors across city sizes and do not allow for variation in the intensity of the trade shock, such that they do not explore the two mechanisms highlighted in this paper. While the empirical analysis in this paper allows for more heterogeneity in the effect of trade they instead use a more structural approach in order to address the welfare implications.

There is a small literature that looks at how international trade affects the economic geography within a country going back to Krugman and Elizondo (1996). Recent papers include Fajgelbaum and Redding (2014), who study how an increase in openness leads to higher population densities in areas with higher access to world markets and Coşar and Fajgelbaum (2016) who document that Chinese coastal cities specialize in traded goods relative to more remote locations. This literature focuses on the importance of intra-national trade costs and looks at settings such as Argentina in the late 19th century and China where intra-national trade costs are an important transmission mechanism for the effects of external integration. This paper complements the previous literature and adds to it in three ways. Firstly, it suggests a different mechanism through which international trade affects the economic geography based on the agglomeration-congestion cost trade-off and spatial sorting. Secondly, in my empirical application I look at the economic geography of an advanced economy whose spatial distribution is governed by different forces and arguably more stable than the one of an industrialising country. Thirdly, in contrast to the previous literature that focuses more on long-term macroeconomic development issues I study the effect on labour demand and thereby link trade to the emerging literature on regional divergence (Giannone, 2017).

In my empirical analysis, I build on the large literature that studies the effects of trade shocks, especially the rise in Chinese import competition, on employment and other variables in local labour markets (Kovak (2013), Autor et al. (2013)) and on the industry level (Acemoglu et al., 2016). In contrast to the previous literature I do not treat each commuting zone as an independent small open economy but rather model the economic geography of the country explicitly. This allows me to formalize and empirically highlight the heterogeneity of the effect of import competition across different commuting zones. I also let the model guide the endogenous spatial distribution of industries rather than treating them as exogenous or pre-determined.

This approach allows me to study the effects on the overall spatial distribution of economic activity, rather than only outcomes on the commuting zone level.

Methodologically, I build on recent empirical and theoretical advances that analyse spatial sorting of heterogeneous firms and sectors in economic geography and urban economics such as Combes et al. (2012), Davis and Dingel (2015) and Gaubert (2018). I contribute to this literature by studying the importance of spatial sorting in the open economy and how it matters for the effects of changes in trade openness. The only paper that jointly models spatial sorting and international trade is contemporaneous work by Garcia et al. (2018).

The paper also adds to the large literature on the distributional effects of trade (see Helpman (2016) for a recent survey), but rather than focusing on heterogeneous effects by skill or gender it focuses on heterogeneity across less and more populated regions. The results could also be relevant for the literature in political economy that tries to understand the regional distribution of the support for populist parties and protectionist policies.

2.3 Data and stylized facts

2.3.1 Data

In this section I present three related stylized facts documenting differential export intensity across employment densities of French commuting zones. The underlying firm-level data comes from two datasets provided by the French national statistical institut (Institut national de la statistique et des études économique, INSEE). The Unified Corporate Statistics System (FICUS) contains all French firms with revenues over 730,00 Euros and reports information on employment, capital, value added, production, and four-digit industry classification collected for tax purposes. It is matched with establishment-level employer-employee data, which indicate the geographical location of each establishment of a given firm year. As is standard in the literature, I use commuting zones (Zones d'emploi) to measure employment density and only focus on metropolitan France. I restrict the sample to manufacturing firms that are only located in one commuting zone allowing a clear spatial assignment.

I additionally complement this data with trade variables derived from the BACI data set (Gaulier and Zignago, 2010) and the gravity dataset provided by Head and Mayer (2014).

2.3.2 Stylized facts

Whether a reduction in international trade cost leads to an expansion or a contraction of economic activity in a region depends on the ability of the firms and sectors in that region to access the foreign market. If this ability is distributed unequally in space then the employment effects of trade are likely to be distributed unequally as well.

Figure 2.2 plots the partial correlation of the share of export sales in total sales of firm (i) in sector (j) with the employment density of the commuting zone (c) it is located in for the cross-section of firms in 1995. The regression is weighted by firm sales to get a measure on the commuting zone level and contains a vector of geographic controls, to account for intra-national trade costs:

$$\left(\frac{export\ sales}{total\ sales}\right)_{icj} = \beta log(emp\ dens_c) + \gamma X_c + \varepsilon_{icj} \tag{2.1}$$

where the vector of geographic controls (X_c) contains decile dummies for distance to Western and the Spanish border, as well as to the Atlantic and the Mediterranean coast. The positive partial correlation indicates that firms in denser places are more export intensive, suggesting that the firms that are able to expand their activity and grow from trade are over proportionally located in denser cities. Figure 2.2b plots the same partial correlation now including a four-digit sector fixed effect. The correlation becomes weaker both in terms of magnitude and significance but remains significant at the 10% level, indicating that within-sector heterogeneity across cities contributes to the overall positive correlation (Table 2.A.2 in the appendix reports the corresponding regression coefficients). Figure 2.2c provides evidence on the importance of across sector heterogeneity for the overall correlation between export intensity and density. Instead of aggregating firm-specific export intensity it plots the average (i.e. national) export intensity of sectors located in different commuting zones:

$$\left(\frac{export\ sales}{total\ sales}\right)_c = \beta log(emp\ dens_c) + \gamma X_c + \varepsilon_c \tag{2.2}$$
$$\text{where : } \left(\frac{export\ sales}{total\ sales}\right)_c = \sum_j \frac{sales_{cj}}{sales_c}\ export\ sales_j$$

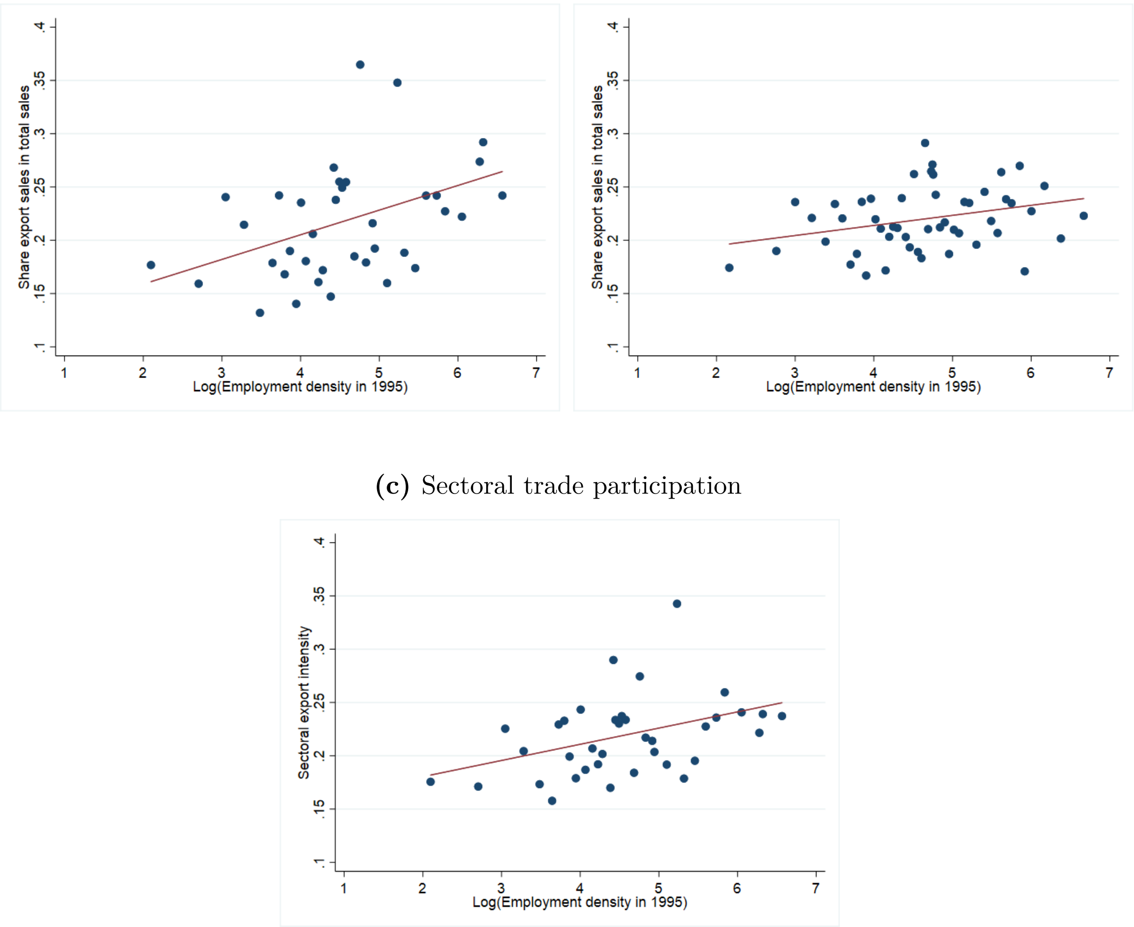
where X_c contains the same geographic controls as in equation 2.1 and β is the coefficient of interest. The positive correlation indicates that sectors located in denser cities are more export intensive.

International economic integration creates unequal revenue and employment growth across firms and sectors driven by their ability to exploit export opportunities. The documented stylized facts suggest that the export intensity and hence the ability of firms and sector to increase revenue and employment from increased international economic integration seems to vary systematically across more and less densely populated areas. Overall, export intensity is higher in denser places (figure 2.2a), and this holds true both within sectors (figure 2.2b) and across sectors (figure 2.2c). These stylized facts guide the development of a model that features both within and across sector mechanisms to explain the unequal effects of trade across space.

2.4 Theory

In this section I develop an open economy multi-sector economic geography model with heterogeneous firms. To this end I integrate the economic geography model developed by Gaubert (2018) with the international trade model from Bernard et al. (2007). Combining a rich economic geography with an international trade model allows me to capture how firm and sector heterogeneity an increase in openness to trade can have heterogeneous effects across regions of different employment density or population size. There are two countries, Home and Foreign ($k = H, F$), where Foreign can either be thought of the rest of the world or a specific country. In the empirical application I will think of Home as France or the United States and Foreign as the Rest of the World or China. I do not introduce any heterogeneity in

Figure 2.2: Trade participation across city densities



(a) Aggregate trade participation

(b) Firm trade participation

(c) Sectoral trade participation

The underlying regressions contain decile dummies for distance to the Mediterranean and Atlantic coast, and the Spanish and the Western border. They are run on the firm level but weighted by sales value on the 1995 cross-section of firms. Standard errors are clustered at the sector and commuting zone level for regressions displayed in figures 2.2a and 2.2b. The corresponding regression coefficients and unweighted specifications can be found in tables 2.A.2 and 2.A.3 in the appendix. Tables 2.A.5, 2.A.4 and 2.A.6 provide additional robustness on the main correlation in figure 2.2a.

terms of the economic geography of the two countries and therefore will suppress the country superscripts to ease readability when describing the spatial equilibrium.

2.4.1 Model setup

Preferences

There is a mass of N identical workers that supply one unit of labour inelastically, consume $h(L_c)$ units of housing and $c(L_c)$ units of the tradable consumption index, where L_c denotes the size of the city a given worker decides to locate in. Workers' preferences are given by:

$$U = \left(\frac{c}{\eta}\right)^{\eta}\left(\frac{h}{1-\eta}\right)^{1-\eta}$$

$$c = \prod_{j=1}^{S} c_j^{\xi_j}$$

$$c_j = \left[\int c_j(i)^{\frac{\sigma_j - 1}{\sigma_j}} di\right]^{\frac{\sigma_j}{\sigma_j - 1}}$$

where $\sum_{j=1}^{S} \xi_j = 1$. Workers maximize their utility subject to the budget constraint $Pc(L_c) + p_H h(L_c) = w(L_c)$, where P is the CES price index of the tradable consumption bundle (c), p_H is the price of housing and the income is given by the wage $w(L_c)$ earned from supplying one unit of labour.

Housing and cities

There is a large number of ex-ante identical potential city sites in each country with an immobile amount of land normalized to one ($\gamma = 1$), that is owned by absentee landowners. There are no trade costs between cities within a country.[4] Housing is immobile and produced according to the following production function:

$$h^S = \gamma^b \left(\frac{\ell}{1-b}\right)^{1-b} \tag{2.3}$$

Given the structure on housing demand and supply the equilibrium in the housing market implies that the amount of housing consumed in equilibrium is given by:

$$h(L_c) = (1-\eta)(1-b)L_c^{-b} \tag{2.4}$$

The amount of housing consumed is smaller in larger cities since the increase in housing production is constrained by the fixed amount of land. If we impose spatial

[4] This assumption is not crucial for any of the results but eases tractability.

equilibrium, i.e. that utility is equalized across space ($V(p_H, P, w) = \bar{U}$) we can derive the equilibrium wage as a function of city size:

$$w(L_c) = \bar{w}((1-\eta)L_c)^{b\frac{1-\eta}{\eta}} \tag{2.5}$$

where $\bar{w} = \bar{U}^{\frac{1}{\eta}}P$ is taken as numeraire. The wage increases with city size. This acts as a congestion cost that counterbalances the gains in productivity from agglomeration.

Production

The economy consists of a number of tradable sectors indexed by $j = 1, .., S$. Each sector is populated by a mass of firms that differ in their exogenously given raw efficiency (z). Firms compete according to monopolistic competition and each firm produces a unique variety (i) using the following production technology:

$$y_j(z, L_c) = \psi(z, L_c)k^{\alpha_j}\ell^{1-\alpha_j} \tag{2.6}$$

where the Hicks-neutral productivity shifter ψ depends on the raw efficiency draw of the firm (z) and the city size the firm locates in (L_c). Sectors are heterogeneous with respect to the factor share (α_j) of labour (ℓ) and capital (k), which they hire from absentee capitalists.

Firm entry and location choice Firm entry closely follows the setup in Melitz (2003). Firms initially pay a sunk market entry cost (f_{E_j}) and draw their raw efficiency z from cumulative distribution function $F_j(z)$. After the realization they decide whether to start producing or to exit immediately. If they decide to produce they choose which city size (L_c) to locate in and whether to only produce for the domestic market, paying per period fixed cost f_{P_j}, or to also export paying per period fixed cost f_{X_j}. Firms die with an exogenous probability δ. In order to match the stylized fact that more productive firms are located in larger cities I follow Gaubert (2018) in assuming that there is a complementarity between raw efficiency (z) and city size (L_c) such that ex-ante more productive firms increase their productivity by more from locating in a larger city. Intuitively, more able entrepreneurs are better able to benefit from the agglomeration externalities in denser places, such as access to technology or access to finance. Formally the assumption is that $\psi(z, L_c)$ is strictly log-supermodular in city size (L_c) and firm raw efficiency (z), and is twice differentiable:

$$\frac{\partial^2 log\psi(z, L_c)}{\partial L_c \partial z} > 0$$

In order to ensure a unique solution for the location problem of the firm the additional regularity condition that the elasticity of productivity with respect to city size is decreasing has to be imposed.

Firm problem Firm profits can be decomposed into profits from domestic and exporting activity $\pi = \pi^d + \pi^x$. Conditional on entry the firm maximises both domestic and exporting profits such that the firm problem is given by:

$$\begin{aligned}\max_{k,\ell,p_j^d,p_j^x,L_c,n} \pi_j =& (1 + T(L_c))(p_j^d \psi_j(z_i, L_c) k^{\alpha_j} \ell^{1-\alpha_j} - w_H(L_c)\ell - \rho_H k - \bar{c}_j^H f_{P_j}) \\ &+ n(1 + T(L_c))(p_j^x \tau_j^{-1} \psi_j(z_i, L_c) k^{\alpha_j} \ell^{1-\alpha_j} - w_H(L_c)\ell - \rho_H k - \bar{c}_j^H f_{X_j})\end{aligned}$$

where $\bar{c}_j^H = \rho^{-\alpha_j} \bar{w}^{1-\alpha_j}$ denotes the non-city size specific marginal costs of firms in sector j. Firms choose optimal factor inputs capital (k) and labour (ℓ), whether to export or not (n), optimal prices for the home market (p_j^d) and the foreign market (p_j^x) (if applicable), and in which city size (L_c) to locate in. $T(L_c)$ is a subsidy proportional to profits paid by city developers to attract firms. Given CES demands and monopolistic competition firms set prices at a constant mark-up over marginal cost. The profit function of a firm that locates in city size L_c is given by:

$$\begin{aligned}\max_{L_c} \pi_j =& \tilde{\kappa}_{1j} \rho_H^{-\tilde{\alpha}_j} (1 + T_j(L_c)) \left(\frac{\psi(z, L_c)}{w_H(L_c)^{1-\alpha_j}} \right)^{\sigma_j - 1} R_j^H P_j^{H^{\sigma_j - 1}} - (1 + T_j(L_c)) \bar{c}_j^H f_{P_j} \qquad (2.7) \\ &+ n(1 + T_j(L_c)) \left[\tilde{\kappa}_{1j} \rho_H^{-\alpha(\sigma_j - 1)} \left(\frac{\psi(z, L_c)}{w_H(L_c)^{1-\alpha_j}} \right)^{\sigma_j - 1} \tau_j^{1-\sigma_j} R_j^F P_j^{F^{\sigma_j - 1}} - \bar{c}_j^H f_{X_j} \right]\end{aligned}$$

where $\tilde{\kappa}_{1j} = \frac{((1-\alpha_j)^{1-\alpha_j} \alpha_j^{\alpha_j} (\sigma_j - 1))^{\sigma_j - 1}}{\sigma_j^{\sigma_j}}$.

City developers

To avoid a coordination failure there is one city-developer per potential site that maximizes profits and opens a city of given size if there is a demand for this city size. City-developers earn income through fully taxing the income of land-owners. They pay a subsidy proportional to profits ($T(L_c)$) in order to attract firms and compete according to perfect competition. They solve the following problem:

$$\max_{\{T_j(L_c)\}_{j\in 1,\dots,S}} \Pi_{L_c} = b(1-\eta)w(L_c)L_c - \sum_{j=1}^{S}\int_z T_j(L_c)\frac{\pi_j(z,L_c)}{1+T_j^i(L_c)}\mathbb{1}_j(z,L_c)f_j(z)dz \tag{2.8}$$

where $\pi_H(L_c) = b(1-\eta)L_c w(L_c)$ is the profit earned by the fully taxing landowners and $\mathbb{1}_j(z,L_c)$ is equal to 1 if firm z chooses to locate in this city and 0 otherwise.

2.4.2 Definition of the spatial equilibrium

The construction of the spatial equilibrium is qualitatively equivalent to the equilibrium in Gaubert (2018). The spatial equilibrium is given by:

(i) *workers maximize utility given prices*

(ii) *utility is equalised across all inhabited cities*

(iii) *firms maximize profits given factor prices and the aggregate price index*

(iv) *landowners maximize profits given prices*

(v) *city developers maximize profits given the wage schedule and the firm problem*

(vi) *National capital and international goods market clear, and the housing and the labour market in each city clear*

(vii) *capital is optimally allocated, and*

(viii) *firms and city developers earn zero profits.*

Since the introduction of international trade does not alter the structure of the equilibrium the existence and uniqueness proof in Gaubert (2018) still applies.

2.4.3 Constructing the spatial equilibrium

Subsidy

As the city developer problems is not affected by international trade it solves the same problem as in Gaubert (2018) such that the same lemma applies:

Lemma 1 ((Lemma 2 in Gaubert (2018))) *In equilibrium, city developers offer and firms take-up a constant subsidy to firms' profit* $T_j^* = \frac{b(1-\eta)(1-\alpha_j)(\sigma_j-1)}{1-(1-\eta)(1-b)}$ *for firms in sector* j*, irrespective of city size* L_c *or firm type* z*.*

Proof. The proof can be found in appendix C in Gaubert (2018).

Matching function

Whenever there is demand for a given city size, it is profitable for a city developer to open a city of that size. Workers are by the definition of the spatial equilibrium indifferent across locating in different city sizes. Firms are not indifferent across different city sizes as their profits vary with city size. The demand for cities is therefore determined by firms' location decisions. Given the subsidy derived above the variable profit of firms that only serve the domestic market and those that serve both the domestic and the foreign market are given by:

$$\max_{L_c} \pi_j^d = \tilde{\kappa}_{1j}\rho_H^{-\alpha(\sigma_j-1)}(1+T_j^*)\left(\frac{\psi(z,L_c)}{w_H(L)^{1-\alpha_j}}\right)^{\sigma_j-1} R_j^H(P_j^H)^{\sigma_j-1} \tag{2.9}$$

$$\max_{L_c} \pi_j^{d,x} = \tilde{\kappa}_{1j}\rho_H^{-\alpha(\sigma_j-1)}(1+T_j^*)\left(\frac{\psi(z,L_c)}{w_H(L)^{1-\alpha_j}}\right)^{\sigma_j-1} \left[R_j^H(P_j^H)^{\sigma_j-1} + \tau_j^{1-\sigma_j} R_j^F P_j^{F^{\sigma_j-1}}\right]$$

Note that the resulting first-order conditions only depend on the trade-off between gains from agglomeration ($\psi(z, L_c)$) and congestion costs ($w_H(L_c)$) and is independent of all other general equilibrium quantities. A crucial implication of this separability is that the optimal location decision is the same for exporters and non-exporters. The resulting first order condition that determines the optimal city size to locate in is given by:

$$\frac{\psi_{L_c}(z,L_c)L_c}{\psi(z,L_c)} = (1-\alpha_j)b\frac{1-\eta}{\eta}$$

where $\psi_{L_c}(z, L_c) = \partial\psi(z, L_c)/\partial L_c$.

This first-order condition accounts for firm and sector heterogeneity and generates spatial sorting across both dimensions. More capital-intensive sectors experience a lower congestion cost which enters scaled by the labour intensity of production $(1-\alpha_j)$. Since the productivity of more efficient firms grows faster with city size, they will sort into the larger cities. It implicitly defines the "matching function" $(L^*_{cj}(z))$ which defines the optimal city size as a function of z and therefore matches firms of different productivities to different city sizes for each sector:

$$L^*_{cj}(z) = \underset{L_c \in \mathcal{L}_c}{\operatorname{argmax}}\, \pi^*_j(z, L_c)$$

As the matching function is unaffected by trade it has the same properties as in in Gaubert (2018). Most importantly, the matching function $L^*_c(z)$ is increasing in z such that there is positive assortative matching between firm raw efficiency z and city size L_c and the set of city sizes in equilibrium $(\mathcal{L})$ is efficient (see Gaubert (2018) for a more detailed discussion).

General equilibrium

The general equilibrium has been determined up to the following set of variables: The productivity cut-offs of entry to the home market (z_j^{kd}) and the export market (z_j^{kd}), where $k \in \{H, F\}$, $m \in \{H, F\}$ and $k \neq m$ denote Home and Foreign and $j = 1, ..., S$ indexes industries, and the sector specific price level (P_j^k); overall expenditure on tradable goods (R^k); the rental rate of capital (ρ_k); and the wage (w_k), where the wage in Home is already pinned down by choosing $\bar{w}$ as the numeraire.

The free entry condition (equation 2.10) for each sector $j = 1, ..., S$ and country $k \in \{H, F\}$ is given by:

$$\begin{aligned}\Big(f_{E_j}+&(1-F(z_j^{kd}))f_{P_j} + (1-F(z_j^{kx}))f_{X_j}\Big)\bar{c}_j^k \\ &= \tilde{\kappa_{1j}}\rho_k^{-\tilde{\alpha}_j}\Big[R_j^k(P_j^k)^{\sigma_j-1}S(z_j^{kd}) + \tau_j^{1-\sigma_j}R_j^m(P_j^m)^{\sigma_j-1}S_j(z_j^{kx})\Big]\end{aligned} \tag{2.10}$$

where f_{E_j} is the units of the final good paid as sunk cost of entry, and z_j^{kd} and z_j^{kx} are the raw efficiency cut-offs for entering the domestic and the export market, respectively.

The zero profit cut-off condition for entering the domestic market (equation 2.11) and the export market (equation 2.12) in each sector j and country $k \in \{H, F\}$ are given by:

$$\bar{c}_j^k f_{P_j} = \tilde{\kappa_{1j}} \rho_k^{-\tilde{\alpha}_j} R_j^k (P_j^H)^{\sigma_j - 1} C_j(z_j^{kd}) \tag{2.11}$$

$$\bar{c}_j^k f_{X_j} = \tilde{\kappa_{1j}} \rho_k^{-\tilde{\alpha}_j} R_j^m (P_j^m)^{\sigma_j - 1} \tau_j^{1-\sigma_j} C_j(z_j^{kx}) \tag{2.12}$$

where $\tilde{\alpha}_j = \alpha_j(\sigma - 1)$.

The goods market clearing condition (equation 2.13) and the equilibrium price index (equation 2.14) for each sector j and country $k \in \{H, F\}$ are given by:

$$R_j^k = \tilde{\kappa_{1j}} \rho_k^{-\tilde{\alpha}_j} M_j^k \Big[R_j^k (P_j^k)^{\sigma_j - 1} S_j(z_j^{kd}) + R_j^m (P_j^m)^{\sigma_j - 1} \tau_j^{1-\sigma_j} S_j(z_j^{kx}) \Big] \tag{2.13}$$

$$1 = \tilde{\kappa_{1j}} \sigma_j \Big[M_j^k S(z_j^{kd}) + \tau_j^{1-\sigma_j} M_j^m S(z_j^{mx}) \Big] (P_j^k)^{\sigma_j - 1} \tag{2.14}$$

The factor market clearing conditions for capital (equation 2.15) and labour (equation 2.16) for each country $k \in \{H, F\}$ is given by:

$$\begin{aligned} \bar{K}_k =& \sum_{j=1}^{S} \tilde{\kappa_{1j}} \rho_k^{-\tilde{\alpha}_j} \frac{(\sigma_j - 1)(\alpha_j)}{\rho_k} M_j^k \\ & \times (R_j^k (P_j^k)^{\sigma_j - 1} S_j(z_j^{kd}) + \tau_j^{1-\sigma_j} R_j^m (P_j^m)^{\sigma_j - 1} S_j(z_j^{kx})) \end{aligned} \tag{2.15}$$

$$\begin{aligned} \bar{N}_k =& (1-b)(1-\eta)\bar{N}_k + \sum_{j=1}^{S} \tilde{\kappa_{1j}} \rho_k^{-\tilde{\alpha}_j} (\sigma_j - 1)(1 - \alpha_j) M_j^k \\ & \times (R_j^k (P_j^k)^{\sigma_j - 1} E_j(z_j^{kd}) + \tau_j^{1-\sigma_j} R_j^m (P_j^m)^{\sigma_j - 1} E_j(z_j^{kx})) \end{aligned} \tag{2.16}$$

where $S(z_j^A), C(z_j^A)$ and $E(z_j^A)$ are normalized values of sectoral sales and employment that are fully determined by the matching function $L_{cj}^*(z)$ for each sector:

$$E_j(z_j^A) = \int_{z_j^A} \mathbb{1}_A(z) \frac{\psi(z, L_{cj}^*(z))^{(\sigma_j - 1)}}{\Big[(1-\eta) L_{cj}^*(z)\Big]^{\frac{b(1-\eta)(1+(1-\alpha_j)(\sigma_j-1))}{\eta}}} f_j(z) dz$$

$$S_j(z_j^A) = \int_{z_j^A} \mathbb{1}_A(z) \left(\frac{\psi(z, L_{cj}^*(z))}{\Big[(1-\eta) L_{cj}^*(z)\Big]^{\frac{b(1-\eta)(1-\alpha_j)}{\eta}}} \right)^{\sigma_j - 1} f_j(z) dz$$

$$C_j(z_j^A) = \left(\frac{\psi(z_j^A, L_{cj}^*(z_j^A))}{\Big((1-\eta) L_{cj}^*(z_j^A)\Big)^{\frac{b(1-\eta)(1-alpha_j)}{\eta}}} \right)^{\sigma_j - 1}$$

where $A = d, x$ distinguishes between the domestic market and the export market and $\mathbb{1}_A(z)$ is equal to one if a firm with raw efficiency level z serves market A. Note that the sector-specific expenditure $R_j^k = \xi_j^k R^k$ is fully determined by R^k.

City size distribution

The equilibrium city size distribution is jointly determined by the matching function as determined by the firm problem and the city developers problem. Given the labour market clearing condition, the population living in a city of size L_c or smaller must equal the labour demand of all firms located in these city sizes and employment in construction:

$$\int_{L_{min}}^{L_c} u f_{L_c}(u) du = \sum_{j=1}^{S} M_j \int_{z_j^*(L_{min})}^{z_j^*(L_c)} \ell_j(z, L_{cj}^*(z)) f(z_j) dz_j + (1-\eta)(1-b) \int_{L_{min}}^{L_c} u f_{L_c}(u) du$$

where $L_{min} = inf(\mathcal{L})$ is the smallest city size in equilibrium. Differentiating this yields the city size density function:

$$f_{L_c}(L_c) = \kappa_4 \frac{\sum_{j=1}^{S} M_j \mathbb{1}_j(L) \ell_j(z_j^*(L_c)) f_j(z_j^*(L_c)) \frac{dz_j^*(L_c)}{dL_c}}{L_c}$$

where $\kappa_4 = \frac{1}{1-(1-\eta)(1-b)}$ and $\mathbb{1}_j(L_c)$ indicates whether firms of sector j are located in city size L_c or not.

2.4.4 Equilibrium properties

I use this model to study the interaction between the agglomeration-congestion cost-trade off and its implied spatial distribution of economic activity, and openness to international trade. The model was motivated by three related stylized facts: The export intensity is higher in denser locations, it is higher in denser location within narrowly defined industries, and it is higher in denser location across industries. To align the model more closely with the presented stylized facts, to simplify the analysis and to closely identify the channels linking trade openness and the spatial distribution of economic activity, I study the effects of within- and across-industry trade separately in different versions of the model.

Within-industry trade

The version of the model that features symmetric countries and no sector heterogeneity can rationalize stylized facts 1 and 2:

Proposition 1 *The export intensity of firms increases with city density.*

Proof. The proof can be found in appendix 2.B.1.

Intuitively, firms in larger cities are more productive and more productive firms are more export intensive, due to selection into exporting, so that in the open economy equilibrium firms in denser locations are more export intensive.

The same version of the model also has a prediction for changes in the spatial distribution of economic activity when moving from the closed to the open economy equilibrium:

Proposition 2 *If both countries are symmetric, the city size distribution in the open economy first-order stochastically dominates the city size distribution in the closed economy.*

Proof. The proof can be found in appendix 2.B.2.

In the symmetric country case trade only occurs within industries such that it does not induce any across-industry reallocations. Across firms within an industry trade induces a reallocation of market share and employment from less to more productive firms as in the standard Melitz model. Note that given the log-supermodularity of productivity and optimal firm behaviour the real productivity (productivity net of congestion cost) increases with city size. Hence, the reallocation from less to more productive firms implies a reallocation from small to larger cities for each sector j. The less productive firms that exit and shrink are located in smaller cities and the more productive firms that expand employment are located in larger cities. This spatial reallocation leads to a higher spatial concentration of sectoral employment in larger cities, in fact the spatial distribution of employment in sector

j in the open economy first-order stochastically dominates the distribution of employment in the closed economy. Since this holds for all sectors the overall city size distribution shifts to the right.

Across-industry trade

The version of the model that only features sector heterogeneity in terms of factor intensities and homogenous firms can rationalize stylized facts 1 and 3:

Proposition 3 *In a two-sector version of the model where factor intensity is the only heterogeneity across sectors and with no heterogeneity in raw-efficiencies, if the other country is relatively labour-abundant, then the sectoral export intensity is higher in denser locations.*

Proof. The proof can be found in appendix 2.B.3.

Since capital is abundant, the price of capital will be lower in equilibrium such that it is cheaper to produce varieties in that sector in the home country. Therefore, the price index in the capital-intensive sector relative to the labour-intensive sector is higher in Foreign, and hence Home will export a larger share of each variety in the capital-intensive sector, which makes it more export intensive. Given the spatial sorting behaviour of heterogeneous sectors, more capital-intensive sectors are located in larger cities such that sectors in larger cities are more export intensive.

The same version of the model also generates a prediction for changes in the spatial distribution of economic activity when moving from the closed to the open economy equilibrium:

Proposition 4 *In a two-sector version of the model where factor intensity is the only heterogeneity across sectors and with no heterogeneity in raw-efficiencies, if the other country is relatively labour-abundant, then the city size distribution in the open economy first-order stochastically dominates the city size distribution in the closed economy.*

Proof. The proof can be found in appendix 2.B.4.

Opening up to trade implies a fall in the relative price of capital from cost minimization and factor market clearing. This leads to a rise in the share of both factors employed in the capital-intensive industry. Since factor endowments remain unchanged employment in the capital-intensive sector increases while employment in the labour-intensive sector decreases. In spatial equilibrium more capital-intensive sectors are located in larger cities, as they are less affected by the congestion cost which is scaled by the labour intensity of production. In this version of the model the distribution of employment across city size in the capital-intensive sector first-order stochastically dominates the distribution in the labour-intensive sector. Hence, the reallocation of employment to the capital-intensive sector implies a reallocation of employment to the larger cities such that the distribution of population in the open economy first-order stochastically dominates the distribution in the closed economy. Therefore endowment-driven across-industry trade leads to spatial concentration in countries that have a comparative advantage in capital-intensive industries.

2.4.5 Comparative statics

Moving from autarky to a costly trade equilibrium is a very drastic change in trade openness and rarely observed in the data. Changes in trade openness τ_j provide a more realistic testing ground for the predictions of the model. In the within-industry version of the model a reduction in trade costs leads to differential effects on firm sales for firms located in smaller and larger cities. In particular, firms below the export raw efficiency cut-off (z^x), located in smaller cities (smaller than $L_{cj}(z^x_{ij})$), will loose revenue relative to exporting firms located in larger cities:

$$\begin{aligned}
\frac{\partial log(r_{cj}(z))}{\partial(\tau_j^{-1})} &\leq 0 \qquad \text{if} \qquad z_{icj}(L) < z^x_j \iff L_c < L_{cj}(z^x_{ij}) \\
\frac{\partial log(r_{cj}(z))}{\partial(\tau_j^{-1})} &> 0 \qquad \text{if} \qquad z_{icj}(L) < z^x_j \iff L_c > L_{cj}(z^x_{ij}) \qquad (2.17)
\end{aligned}$$

where $r_{cj}(z))$ denotes the revenue of a firm of raw efficiency z in sector j optimally located in a city of size c and $\partial(\tau_j^{-1})$ denotes a decrease in trade costs. We get a similar comparative static for the across-sector version of the model, where the exporting sector, located in larger cities, is going to expand sales following a decrease in trade costs ($\partial(\tau^{-1})$), such that sales originating from sectors located in larger cities will increase (note that $r_{cj} = r_c$ since cities completely specialize in single industries in this version of the model):

$$\frac{\partial log(r_c)}{\partial(\tau^{-1})} \begin{cases} < 0 & \text{if} \quad \alpha_j < \alpha^C \iff L_{cj} < L_c^C \\ > 0 & \text{if} \quad \alpha_j > \alpha^C \iff L_{cj} > L_c^C \end{cases} \tag{2.18}$$

where α_C denotes the cut-off in terms of capital intensity when sectors become a net exporting sector and L_c^C denotes the city size where the homogeneous firms in that sector decide to locate in.

2.5 Regression analysis

2.5.1 Estimation

The model can rationalize the stylized facts presented in section 2.3 and, in line with the cross-country evidence from figure 2.1, predicts heterogeneous effects of a reduction in trade costs across different densities. In this section, I test these model comparative statics using exogenous changes in market access following Redding and Venables (2004) using the BACI database and the gravity dataset provided by Head and Mayer (2010). I calculate changes in market access exogenous to French firms following Hering and Poncet (2010). I first estimate a standard gravity equation separately for each of the 114 sectors using all countries except France for the period 1995 - 2015.

$$\begin{aligned} log(x_{odt}) =& \gamma_{ot} + \delta_{dt} + \alpha_1 log(dist_{od}) + \alpha_2 \mathbb{1}[contig_{od}] + \alpha_3 \mathbb{1}[lang_{od}] + \alpha_4 \mathbb{1}[col_{od}] \\ & + \alpha_5 \mathbb{1}[EU_{od}] + \alpha_6 \mathbb{1}[FTA_{od}] + \varepsilon_{odt} \end{aligned} \tag{2.19}$$

where o and d indicate origin and destination country. γ_{ot} and δ_{dt} are time-varying importer and exporter fixed effects. $dist_{od}$ is the population weighted distance between origin and destination. $\mathbb{1}[contig_{od}]$, $\mathbb{1}[lang_{od}]$, $\mathbb{1}[col_{od}]$, $\mathbb{1}[EU_{od}]$ and $\mathbb{1}[FTA_{od}]$ are a set of dummies indicating whether the origin and destination country are on the same landmass, share a language, were in a colonial relationship, are both members of the EU and have an FTA, respectively. Based on the estimates from these regressions I define the market access of a French sector j at time t (MA_{FRjt}) as:

$$\begin{aligned} MA_{FRjt} = \sum_d dist_{FRd}^{\hat{\alpha}_{j1}} exp(\hat{\delta}_{djt}) exp(\hat{\alpha}_{j2}\mathbb{1}[contig_{FRd}] + \hat{\alpha}_{j3}\mathbb{1}[lang_{FRd}] \\ + \hat{\alpha}_{j4}\mathbb{1}[col_{FRd}] + \hat{\alpha}_{j5}\mathbb{1}[EU_{FRdt}] + \hat{\alpha}_{j6}\mathbb{1}[FTA_{FRdt}]) \qquad (2.20) \end{aligned}$$

Equipped with these measures of exogenous export opportunities I test the comparative statics of the model. The model equation 2.17 predicts that the effect of reduction in trade cost should be more positive for firms located in denser cities which, assuming that this interaction between city density and trade costs is well approximated using a linear term, can be mapped into the following regression framework:

$$\begin{aligned} \Delta log(r_{ijt}) = \beta_{f0} + \beta_{f1}\Delta log(MA_{jt}) + \beta_{f2} log(dens_{c95}) \\ + \beta_{f3}\left[\Delta log(MA_{jt}) \times log(dens_{c95})\right] + X_c'\gamma_f + \delta_j + \varepsilon_{ijt} \qquad (2.21) \end{aligned}$$

where the model predicts $\beta_{f1} > 0$ and $\beta_{f3} > 0$. Since this prediction holds both on the firm as well as on the city level I run both an unweighted regression and one that is weighted by initial firm sales, which tests the prediction in monetary terms on the city level. Note that the model does not provide any guidance whether employment size or density is the correct measure, as they are isomorphic. I follow the previous literature (e.g. Combes et al. (2012)) and use employment density in the regressions rather than population size. In line with the specifications for the stylized facts I include a vector of geographic characteristics (X_c) consisting of decile dummies for distance to the Atlantic and Mediterranean coast, and the Western and the Spanish border, since geography is and important determinant of trade activity, while not explicitly modelled.

The sector-level channel (equation 2.18) can be mapped into a regression framework in a similar fashion yielding:

$$\begin{aligned} \Delta log(r_{ct}) = \beta_{s0} + \beta_{s1}\Delta log(MA_{ct}) + \beta_{s2} log(dens_{c95}) \\ + \beta_{s3}\left[\Delta log(MA_{ct}) \times log(dens_{c95})\right] + X_c'\gamma_s + \varepsilon_{ct} \qquad (2.22) \end{aligned}$$

where $\Delta log(r_c)$ and $\Delta log(MA_{ct})$ are the average change in revenue and market access for sectors located in c and X_c contains the same set of geographical controls as above. The model predicts that $\beta_{s1} \geq 0$ and $\beta_{s3} > 3$. More export opportunities for the average manufacturing sector in city c should increase revenues and more so in denser cities, where the export-intensive industries are located. The model that generates this prediction abstracts from firm heterogeneity so to be consistent with the model the change in the average revenue has to be defined abstracting from firm sorting and solely rely on industry-level variation in the change of sales, I therefore define the change in revenue of the average sector located in c as follows.

$$\Delta log(r_c) = \sum_j \frac{r_{cj}}{r_c} \Delta log(r_j)$$

Note that this measure does not calculate the actual average change in sectoral sales in location c which had to be based on $\Delta log(r_{cj})$ rather than $\Delta log(r_j)$. By using the national rather than the local change in revenue I isolate the differences in sectoral response (i.e. the differences across sectors) to the market access shock from the differences across firms within the sector across locations. Analogously average market access is defined as:

$$\Delta log(MA_c) = \sum_j \frac{r_{cj}}{r_c} \Delta log(MA_j)$$

2.5.2 Results

The main results for the firm-level channel (equation 2.21) are displayed in table 2.1 using a long difference from 1995 to 2015.[5] The main specifications of interest, including four-digit sector fixed effects, are displayed in columns 2 and 4. The results are in line with the predictions of the model across weighted and unweighted specifications. An increase in export opportunities increases firm sales and does significantly more so for firms located in denser cities. The decrease in the coefficient of interest once we include sector fixed effects already foreshadows the importance of across sector differences.

[5] Tables 2.A.7 and 2.A.8 in the appendix present results for alternative sets of controls and stacked ten year differences, reducing firm attrition.

Table 2.1: Firm-level mechanism

	Δ_{20} log(sales)			
	Unweighted		Weighted by initial sales	
Δ_{20} log(MA_j)	0.010		0.068	
	(0.0397)		(0.0421)	
Δ_{20} log(MA_j)	0.043^a	0.022^b	0.058^a	0.030^a
× log(dens emp_{c95})	(0.0161)	(0.0094)	(0.0157)	(0.0111)
log(emp dens_{c95})	-0.017^c	-0.007	-0.032^b	-0.020
	(0.0085)	(0.0064)	(0.0134)	(0.0130)
Sector FE	No	Yes	No	Yes
Controls	Yes	Yes	Yes	Yes
Observations	23355	23355	23355	23355
Pseudo R^2	0.02	0.07	0.04	0.13

Twenty-year difference from 1995 to 2015. Standard errors clustered at the sector (four digit) level and the czone level in parenthesis. Variables are winsorized at the 3rd and 97th percentile for each sector. Controls are dummies for distance deciles to the Western and the Spanish border, and the Atlantic and Mediterranean coast. Statistical significance levels are indicated by ^{c} for $p < 0.10$, ^{b} for $p < 0.05$, ^{a} for $p < 0.01$.

The main results for the sector-level channel (equation 2.22) are presented in table 2.2. In line with the predictions of the model I find that an increase in average market access increases sales of the average sector across commuting zones. This positive association of market access with sales is stronger in denser places, indicating that the industries located in denser places are more able to take advantage of the export opportunities.

The model predictions of both the firm- and the sector-level channel find support in the data, indicating that international economic integration re-allocates economic activity from less to more dense locations. Next, I will test whether the mechanisms that underlie these model predictions find support in the data. In the baseline version of the model the firm-level reallocation is driven by productivity differences and the across-sector heterogeneity is driven by differences in input

Table 2.2: Sector-level mechanism

	Δ_{20} log(sales)			
	Unweighted		CZs weighted by initial sales	
Δ_{20} log(MA$_c$)	0.61^a	0.84^a	0.76^a	0.92^a
	(0.137)	(0.152)	(0.172)	(0.169)
Δ_{20} log(MA$_c$)		0.29^a		0.45^a
× log(dens emp$_{c95}$)		(0.084)		(0.104)
log(emp dens)	-0.02^a	-0.02^b	-0.04^a	-0.02^b
	(0.008)	(0.007)	(0.012)	(0.010)
Controls	Yes	Yes	Yes	Yes
Observations	352	352	352	352
Pseudo R^2	0.36	0.39	0.44	0.49

Twenty-year difference from 1995 to 2015. Robust standard errors in parenthesis. Variables are winsorized at the 3rd and 97th percentile. Controls are dummies for distance deciles to the Western and the Spanish border, and the Atlantic and Mediterranean coast. Statistical significance levels are indicated by ^{c} for $p < 0.10$, ^{b} for $p < 0.05$, ^{a} for $p < 0.01$.

intensity across sectors. Extending the baseline version of the model, the differences in export intensity could also be driven by heterogenous variable (τ_{cj}) or fixed trade costs ($f_{X,cj}$) across locations.

2.6 Evidence on mechanisms

In this section I test the mechanisms that generate the differential export intensity across locations in the model. First, I will provide evidence on the importance of spatial sorting (of firms and sectors) and the traditional benefits of agglomeration in terms of productivity as suggested by the baseline model. Second, I will allow the variable and fixed cost of exporting to potentially vary across locations and test whether there are trade-specific gains from agglomeration that outweigh the higher price level in denser cities such that they manifest in lower variable of fixed trade costs within industries across commuting zones.

2.6.1 Spatial sorting of heterogeneous sectors and differences in export intensity across industries

The model suggests that the spatial sorting of heterogeneous sectors is driven by differences in factor intensity across sectors, which in turn drives the heterogeneity of export intensity of sectors across location documented in figure 2.2c. To formally test this hypothesis I include a measure of skill intensity in the regression underlying figure 2.2c:[6]

$$\left(\frac{export\ sales}{total\ sales}\right)_c = \beta_1 log(emp\ dens_c) + \beta_2 log\left[\left(\frac{high\ skill}{low\ skill}\right)_c\right] + \gamma X_c + \varepsilon_c$$

where X_c is the vector of geographical controls used throughout all specifications (it includes dummies for distance deciles to the Western border, the Spanish border, the Mediterranean coast, and the Atlantic coast). If sector sorting due to factor intensity drives the reduced-form correlation between employment density and export intensity we would expect $\beta_2 > 0$ and $\beta_1 = 0$. The results are displayed in table 2.3. Column 1 reproduces the reduced-form correlation from figure 2.2b. Column 2 only includes skill intensity documenting that skill-intensive industries are more export intensive, corroborating the assumption that they are France's comparative advantage sectors. When including both average skill intensity and employment density I find that the effect of skill intensity remains highly significant while the coefficient on employment density shrinks by two-thirds and becomes insignificant (column 3), in line with the model mechanism.

2.6.2 Agglomeration, spatial sorting of heterogeneous firms, and differences in export intensity within industries

The model suggests that differences in productivity, driven by spatial sorting and classical agglomeration forces, drive differences in export intensity of firms across locations within industries. To test this formally I include a measure of TFP

[6] So far I have been agnostic about whether capital refers to human or physical capital. In this setup it is more convenient to use human rather than physical capital as it is available in the same units as low-skilled labour.

Table 2.3: MICRO-MECHANISM FOR SECTOR-LEVEL CHANNEL

	Sectoral export intensity		
$log(empdens_{c95})$	0.015^{a} (0.0059)		0.005 (0.0065)
$log(\frac{skill\ emp}{emp}_{c95})$		0.096^{a} (0.0290)	0.075^{a} (0.0355)
Controls	Yes	Yes	Yes
Observations	352	352	352
Pseudo R^2	0.63	0.64	0.64

Cross-section in 1995. Robust standard errors in parenthesis. Controls are dummies for distance deciles to the Western and the Spanish border, and the Atlantic and Mediterranean coast. Statistical significance levels are indicated by ^{c} for $p < 0.10$, ^{b} for $p < 0.05$, ^{a} for $p < 0.01$.

estimated using Olley and Pakes (1996) with the correction suggested by Ackerberg et al. (2015) into the reduced-form specification from figure 2.2b:

$$\left(\frac{export\ sales}{total\ sales}\right)_{icj} = \beta_1 log(emp\ dens_c) + \beta_2 log\left(\psi_{icj}\right) + \gamma X_c + \delta_j + \varepsilon_c$$

where X_c is the vector of geographical controls used throughout all specifications and δ_j is a four digit sector fixed effect. Theory predicts that $\beta_2 > 0$ and $\beta_1 = 0$, if the differences in export intensity are fully driven by differences in productivity. The results are displayed in table 2.4. Column 1 reproduces the reduced-form correlation between firm export intensity and employment density. Column 2 regresses export intensity on firm TFP, corroborating that more productive firms are more export intensive. Column 3 displays results for the full specification regressing export intensity on both employment density and firm TFP. Including TFP slightly dampens the effect of employment density on export intensity and the coefficient on employment density becomes marginally insignificant. However, the magnitude of the coefficient barely moves and is still much closer to the reduced-form correlation than to 0 indicating that employment density might affect export intensity through other channels that are not captured by productivity.[7]

[7] Another potential explanation, which I have not yet explored further, is that the result is driven by high measurement error in TFP.

Table 2.4: Mechanism for firm-level channel

	Share of export sales		
$log(emp\ dens_c)$	0.009^c (0.0049)		0.008 (0.0053)
$log(\psi_{icj})$		0.208^a (0.0373)	0.208^a (0.0372)
$log(w_{cj})$		0.087 (0.0745)	0.027 (0.0859)
Controls	Yes	Yes	Yes
Ind FE	Yes	Yes	Yes
Observations	118774	118774	118774
Pseudo R^2	0.44	0.47	0.47

Cross-section in 1995. Standard errors clustered at the sector (four digit) level and the czone level in parenthesis. Controls are dummies for distance deciles to the Western and the Spanish border, and the Atlantic and Mediterranean coast. Statistical significance levels are indicated by ^{c} for $p < 0.10$, ^{b} for $p < 0.05$, ^{a} for $p < 0.01$.

2.6.3 Differences in variable and fixed trade costs across commuting zones

Productivity differences due to spatial sorting and agglomeration are just one reason why export intensity could increase with employment density. The model suggests two other sources that could link employment density to export intensity, namely that the variable cost of exporting (τ_{cj}) or the fixed cost of exporting (f_{xcj}) could systematically vary with employment density. Intuitively, denser locations might reduce the variable trade cost because of better infrastructure and it might be easier to meet foreign buyers for firms located in denser places reducing variable or fixed cost of exporting. At the same time given the higher cost of space in cities fixed and variable cost of exporting could increase with city size. I first use the model structure to infer the unobserved variable and fixed trade costs and then test for trade-specific gains from agglomeration by correlating the estimated parameters with employment density.

Variable trade cost

From the model we can write the revenue from exporting conditional on exporting as:

$$r_{icj}^{X}=\tilde{\kappa}_{1j}\rho^{-\tilde{\alpha_j}}\left(\frac{\psi_{icj}}{w_{cj}^{1-\alpha_j}}\right)^{\sigma_j-1}\tau_{icj}^{1-\sigma_j}R_j^F(P_j^F)^{\sigma_j-1}$$

where we observe data for r_{icj}^{X}, ψ_{icj} and w_{cj}. From this equation we can identify relative firm-specific trade costs, by normalizing by a reference firm within the same sector which eliminates the remaining unobserved sector-specific variables $\tilde{\kappa}_{1j}$, $\rho^{-\tilde{\alpha_j}}$, R_j^F, $(P_j^F)^{\sigma_j-1}$, which yields:

$$\tilde{\tau}_{icj}=\frac{\tau_{icj}}{\bar{\tau}_j}=\left(\frac{r_{icj}^{X}}{\bar{r}_j^{X}}\right)^{\frac{1}{1-\sigma_j}}\left(\frac{\psi_{icj}}{\bar{\psi}_j}\right)\left(\frac{w_c}{\bar{w}}\right)^{-(1-\alpha_j)}$$

where $\tilde{\tau}_{icj}$ denotes the relative variable trade cost faced by firm i and a bar indicates a variable corresponding to the value of the sector-specific reference firm. As reference firm I choose the firm with the smallest export revenue to get an empirical counterpart to the cut-off firm. I assume that the elasticity of substitution is constant across sectors and equal to 4, and set the capital intensity (α_j) equal to the capital intensity from the TFP estimation.

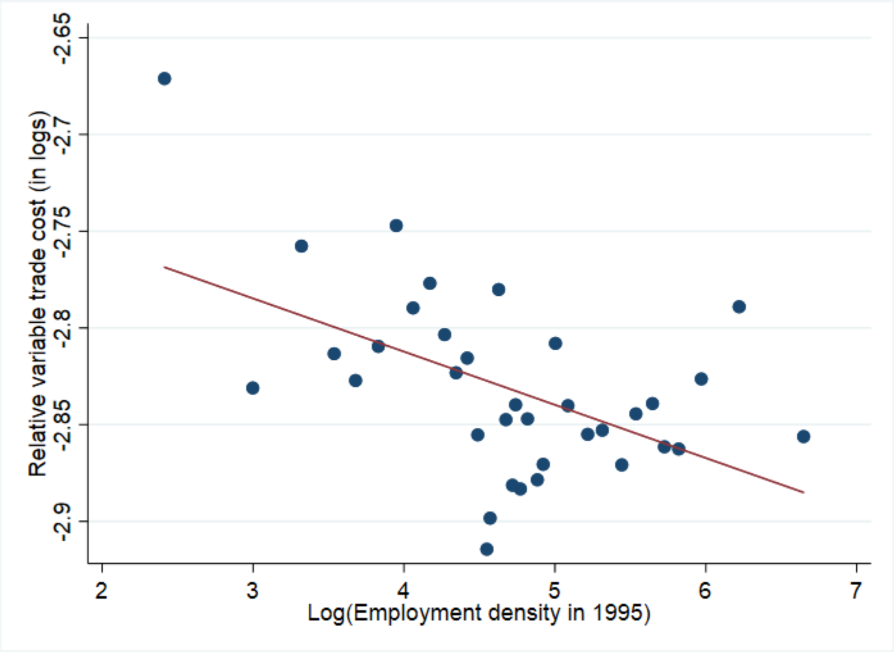
Figure 2.3 plots the partial correlation of the derived relative variable trade cost with employment density.[8] The regression contains the standard controls for geography used throughout and is weighted by export revenue in order to get the variable trade cost in monetary terms. The negative and statistically significant correlation indicates that variable trade costs are decreasing in employment density. Hence, besides the productivity advantages of dense places there also seems to be a positive effect that makes it less costly to export.

Fixed trade cost

From the model we know that a firm will export if exporting generates positive profits, i.e. $r_{icj}^{X}>f_{x,icj}$, where $f_{x,icj}$ is unobservable for all firms while r_{icj}^{X} is unobservable for firms that only serve the domestic market. Normalizing both

[8] Table 2.A.12 in the appendix contains the corresponding regression and additional specifications.

Figure 2.3: Variable trade cost across city densities



The corresponding regression coefficients and alternative specifications can be found in table 2.A.12.

revenue and fixed cost by the cut-off firm, for which $\bar{r}_j^X = \bar{f}_{x,j}$, we can derive values for $\tilde{r}_{icj}^X$ and $\tilde{f}_{x,icj}$, where $\tilde{r}_{icj}^X > \tilde{f}_{x,icj}$ still holds:

$$\tilde{r}_{icj}^X = \frac{r_{icj}^X}{\bar{r}_{cj}^X} = \left(\frac{\psi_{icj}}{\bar{\psi}_j}\right)^{\sigma_j - 1} \left(\frac{w_c}{\bar{w}}\right)^{(1-\sigma_j)(1-\alpha_j)} \left(\frac{\tau_{icj}}{\bar{\tau}_j}\right)^{1-\sigma_j}$$
$$\tilde{f}_{x,icj} = \frac{f_{x,icj}}{\bar{f}_{x,j}}$$

where the relative fixed cost $(\tilde{f}_{x,icj})$ remains unobservable but the relative revenue from exporting $(\tilde{r}_{icj}^X)$ is now a function of observables and parameters, that can be calculated.

Suppose the relative fixed cost of firm i in sector j located in commuting zone c consists of the following two parts:

$$ln(\tilde{f}_{x,icj}) = ln(\tilde{f}_{x,cj}) + \tilde{\varepsilon}_{icj}$$

where $\tilde{f}_{x,icj}$ denotes the firm-specific relative fixed trade cost and $\tilde{f}_{x,cj}$ the city-sector-specific relative fixed trade cost and $\tilde{\varepsilon}_{icj}$ the idiosyncratic component of the firm's cost which is assumed to be distributed normally: $\tilde{\varepsilon}_{icj} \sim \mathcal{N}(0, \sigma^2)$.

This setup, which follows the specification used by Allen (2014), allows me to use a maximum likelihood approach to identify the relative city-sector-specific fixed trade cost exploiting variation across firms within a city-industry pair. For each firm I observe whether it exports ($\mathbb{1}^X_{icj} = 1$) or does not export ($\mathbb{1}^X_{icj} = 0$). The ML estimator of the city-sector-specific fixed cost of exporting maximizes the following log-likelihood function:

$$\begin{aligned} l_{cj}(f_x) = \sum_{i=1}^{N} \mathbb{1}\left[\mathbb{1}^X_{icj} = 1\right] ln\left(\Phi\left(\sigma^{-1}ln(\tilde{r}^X_{icj}) - \sigma^{-1}ln(\tilde{f}_{x,cj})\right)\right) \\ +\mathbb{1}\left[\mathbb{1}^X_{icj} = 0\right] ln\left(1 - \Phi\left(\sigma^{-1}ln(\tilde{r}^X_{icj}) - \sigma^{-1}ln(\tilde{f}_{x,cj})\right)\right) \end{aligned}$$

where Φ is the cumulative distribution function of the standard normal distribution. Note that this likelihood function is identical to a probit regression of an exporter dummy on revenue from exporting and city-sector fixed effects:

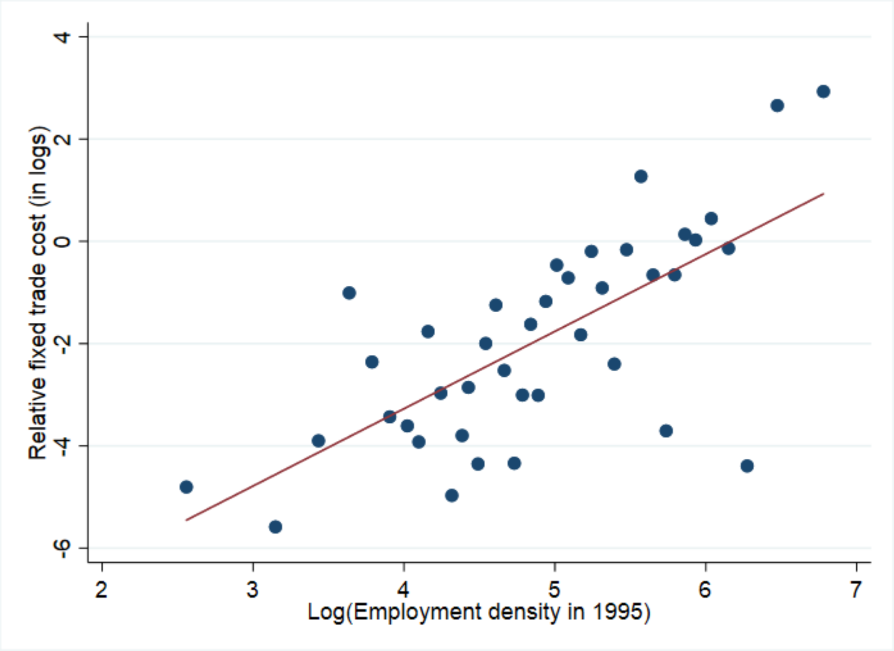
$$\mathbb{1}\left[\mathbb{1}^X_{icj} = 1\right] = \beta\tilde{r}^X_{icj} + \gamma_{cj} + \varepsilon_{icj}$$

where $\tilde{f}_{x,cj} = exp(-\frac{\hat{\gamma}_{cj}}{\hat{\beta}})$. Similar to the case of variable trade cost discussed above we can identify relative sector-city specific fixed trade cost up to a normalization.

Figure 2.4 plots the regression of these relative fixed trade cost on employment density conditional on the standard set of geographic controls and sector fixed effects. The statistically significant positive correlation with density suggests that any agglomeration benefits specific to fixed trade costs are outweighed by the generally higher price level in denser places.

Overall, while spatial sorting and productivity gains from agglomeration matter for the link between employment density and export intensity, there is strong evidence for variation of trade costs across different densities. Variable trade costs are decreasing in density indicating that there are agglomeration gains from density specific to the variable cost of exporting outweigh the cost of density in terms of higher prices. For fixed cost of exporting I find the opposite where any agglomeration gains from density are not strong enough to outweigh the increased cost from higher

Figure 2.4: Fixed trade cost across city densities



The corresponding regression coefficients and alternative specifications can be found in table 2.A.13.

local prices such that the fixed cost of exporting increases with city density.

2.7 Additional evidence from Chinese import competition in the US

2.7.1 Data

To complement the analysis in section 2.5 and to provide some further evidence on the predictions of the model I test them on data from the United States using the increase in import competition from China between 1991 and 2007 as an exogenous shock. In my empirical strategy, as well as the data and definitions used, I closely follow the previous literature (Autor et al., 2013, Acemoglu et al., 2016). Throughout the paper I present results estimated on the stacked sub-periods 1991 to 1999 and 1999 to 2007.

Trade data I use the data on sectoral trade flows that were used and provided by Acemoglu et al. (2016) and Feenstra et al. (2017). They provide trade flows for 392 manufacturing and industries at the 4-digit SIC code level. The trade data was originally downloaded from Comtrade and subsequently transformed into real 2007 dollars.[9]

Employment data I obtain data on local industry composition in 1991, 1999 and 2007 from the County Business Patterns (CBP). The CBP provides information on employment, payroll and firm-size distribution by county and industry. In order to avoid disclosure some establishments are not identified at the most disaggregated level and sometimes employment is only reported as an interval rather than a number. I use the algorithm developed by Autor et al. (2013) to impute employment by county and 4-digit SIC code. I then aggregate this data to the commuting zone level using cross-walks provided by David Dorn.[10] The detailed procedure of the algorithm is outlined in the online appendix in Autor et al. (2013). This gives a panel of observations at the industry-commuting zone level for 722 commuting zones and 392 industries for two periods.

While the main regressions are run on the industry-commuting zone level, for some robustness checks that require wage data not available on the industry-commuting zone level I use data at the commuting zone level provided by Autor et al. (2013). This dataset consists of commuting zone-specific import competition shocks, and changes in wages and employment for the periods 1990 to 2000 and 2000 to 2007.

2.7.2 Estimation

There are three main differences between the specifications run on the US data relative to the previous analysis. Firstly, I study the effect of an import competition shock rather than export opportunity shock complementing the earlier analysis. Secondly, given data constraints and the customs of the literature on the China

[9] A more detailed discussion on the preparation of the trade data can be found in Acemoglu et al. (2016).

[10] These cross-walks can be found at www.ddorn.net/data.htm

Shock I rely on different variables. Firstly, I use employment rather than sales as outcome and weighting variable following the earlier literature on the China Shock. Secondly, given the data availability I use population size rather than density for the interaction term, which is also in line with the theory. Thirdly, since I only have regional rather than firm-level data I estimate the firm-level mechanism only on the regional and not on the firm level.

Within-industry trade and firm heterogeneity

To test the model predictions I estimate an empirical counterpart to equation (2.17). Analogous to equation 2.21 I impose a linear interaction between city size and the trade shock and estimate the following equation:

$$\Delta L_{cjt} = \beta_0 + \beta_1 \Delta Imp_{jt} + \beta_2 L_{c90} + \beta_3 \left[\Delta Imp_{jt} \times L_{c90}\right] + \gamma_j + \delta_r + \alpha_t + \epsilon_{cjt} \tag{2.23}$$

where ΔL_{cjt} is the log change in employment in commuting zone c in sector j in period t multiplied by 100. ΔImp_{jt} denotes the change in imports from China in sector j and L_{ct} denotes the population in commuting zone c at the beginning of period t. γ_j are sectoral fixed effects, α_t time fixed effects and δ_j are fixed effects for the eight Census regions. The regressions are weighted by initial employment in each industry-commuting zone cell and standard errors are clustered at the three digit SIC level. The intuition outlined above predicts that $\beta_1 < 0$ and $\beta_3 > 0$. I estimate these equations using a 2SLS approach instrumenting endogenous trade flows from China to the US ($\Delta Imp_{jt}^{US,Ch}$) with trade flows from China to other advanced economies ($\Delta Imp_{jt}^{Ot,Ch}$) as in Acemoglu et al. (2016). The variables are defined as follows:

$$\Delta Imp_{jt}^{US,Ch} = \frac{\Delta M_{jt}^{US,Ch}}{Y_{j91} + M_{j91} - E_{j91}}$$
$$\Delta Imp_{jt}^{Ot,Ch} = \frac{\Delta M_{jt}^{Ot,Ch}}{Y_{j88} + M_{j88} - E_{j88}}$$

Import flows (ΔM_{jt}) are normalized by apparent consumption (production (Y) plus imports (M) minus exports (E)) at the beginning of the period, and before the period for the instrument, to avoid introducing any endogeneity through anticipation effects.

Results The main results are presented in Table 2.5.[11] The first column corroborates that the aggregate effect of an import competition shock is still negative when splitting industries into industry-commuting zone cells. Including the interaction term in column 2 yields an estimate of 1.23 which is statistically significant at the 1% level. The resulting coefficients remain highly statistically significant and the point estimate is 0.94 when controlling for regional and sectoral trends. So a one percentage point rise in industry import penetration reduces industry level employment by around three percentage points in a commuting zone with a population of a log point above the mean, while it reduces it by four percentage points in a mean-sized commuting zone.

While this evidence is in line with the predictions of the model that an import competition shock translates into a more negative labour demand shock in less populated commuting zones because of the spatial sorting of heterogeneous firms, it is also consistent with other mechanisms. The most apparent alternative explanation is based on variation in the labour supply elasticity across different city sizes as identified by Brülhart et al. (fortchoming) for border towns in Austria. The empirical pattern of relative changes in employment could be generated from a uniform labour demand shock across city sizes if the labour supply elasticity was decreasing with city size. While the demand and the supply-driven explanations have identical implications for changes in employment, they have different implications for wages. A supply-driven model suggests that the effect of an import competition shock on wages would be less negative in smaller cities and more negative in larger cities. The demand driven mechanism in my model on the other hand predicts that the effect on wages should also be smaller in bigger cities or equal across city sizes depending on the elasticity of labour supply, which is constant across city sizes.

I use these differentiating predictions on changes in the wage in order to empirically rule out the labour supply driven explanation. Unfortunately, I cannot use the CBP data to do this as, due to the omissions in the data, I cannot obtain a credible average wage on the sector-commuting zone level. Instead, I rely on

[11] The corresponding first stage regressions can be found in Table 2.A.14 and 2.A.15 in the appendix

Table 2.5: Firm-level mechanism: Imports from China and changes in manufacturing employment across different city sizes within an industry

	ΔL_{cj}	ΔL_{cj}	ΔL_{cj}	ΔL_{cj}
$\Delta Imp_j^{US,Ch}$	-2.77***	-6.20***	-4.03***	-3.99***
	(0.836)	(1.736)	(1.271)	(1.257)
$\Delta Imp_j^{US,Ch} \times ln(pop_c)$		1.23***	0.96***	0.94***
		(0.364)	(0.274)	(0.268)
$ln(pop_c)$		1.08***	1.35***	1.30***
		(0.245)	(0.218)	(0.208)
Time FE	Yes	Yes	Yes	Yes
Region FE	No	No	No	Yes
Industry FE (4d)	No	No	Yes	Yes
Observations	129116	129116	129116	129116
Pseudo R^2	0.02	0.04	0.20	0.20
AP F-statistic ΔImp	99.63	69.87	73.64	73.75
AP F-statistic IA	.	125.06	106.49	106.30

Note: Robust standard errors clustered at the three digit SIC level are reported in parenthesis. The regressions include fixed effects for ten sub-sectors within manufacturing and eight census regions. Regressions are weighted by initial employment in each sector-commuting zone cell. The sample includes 392 manufacturing industries in 722 commuting zones for the periods 1991 - 1999 and 1999 - 2007 that are stacked in the estimation. The population variable is demeaned such that $\Delta Imp_j^{US,Ch}$ is the effect of an import competition shock for the mean-sized commuting zone. Stars indicate significance levels the following levels *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

the wage data from the Census Integrated Public Use Micro Samples (Ruggles et al., 2017) to generate an average wage on the commuting zone level. Since census data is not available for every year before 2000 I adjust the periods to 1990 - 2000 and 2000 - 2007. In particular, I use the dataset developed by Autor et al. (2013) that provide changes in employment and wages at the commuting zone level as well changes in Chinese import competition shocks at the commuting zone level, which are defined as follows:

$$\Delta Imp_{ct}^{US} = \sum_j \frac{L_{cjt}}{L_{ct}} \frac{\Delta M_{jt}^{US,Ch}}{L_{jt}}$$
$$\Delta Imp_{ct}^{Ot} = \sum_j \frac{L_{cjt}}{L_{ct}} \frac{\Delta M_{jt}^{Ot,Ch}}{L_{jt-1}}$$

I run their baseline regression augmented with an interaction term between the import competition shock and the initial population in the commuting zone:

$$\Delta y_{ct} = \beta_0 + \beta_1 \Delta Imp_{ct}^{US} + \beta_2 L_{c90} + \beta_3 \left[\Delta Imp_{ct} \times L_{c90}\right] + \delta_r + \alpha_t + \varepsilon_{1cjt} \quad (2.24)$$

where δ_r and α_t are regional and time fixed effects. Since there is not sufficient variation in the logged population variable to identify both first stages separately, I estimate equation (2.24) using a control function approach as well as using 2SLS. The results are qualitatively the same for both estimation procedures.

The main results based on the control function approach are presented in Table 2.6.[12] The regressions on employment corroborate the earlier findings that the employment effect of an import competition shock is larger in smaller cities even when reducing the amount of identifying variation by aggregating across industries. The regressions on changes in the average wage suggest that the effect on wages only varies marginally with city size and if anything the effect is less negative in larger cities. This is in line with the labour demand driven mechanism suggested by the model and evidence against a supply-based explanation.

Across-industry trade and comparative advantage

The spatial sorting of sectors across regions, driven by the factor intensity of their input use, affects the intensity with which an average sector in a commuting zone reacts to an increase in import competition driven by a fall in trade costs. Analogously to the market access shock we get the following estimating equation:

$$\Delta log(L_c) = \beta_0 + \beta_1 L_{ct} + \beta_2 \Delta Imp_{ct}^{US,Ch} + \beta_2 \left[\Delta Imp_{ct}^{US,Ch} \times L_{ct}\right] + \delta_r + \varepsilon_{ct} \quad (2.25)$$

[12] The results using a 2SLS approach using either log population or absolute population as interaction can be found in table 2.A.16 and 2.A.17 in the appendix. The results are in line with those from the control function approach.

Table 2.6: Wage and employment regressions on the commuting zone level

	ΔL_c	Δw_c	ΔL_c	Δw_c	ΔL_c	Δw_c	ΔL_c	Δw_c
$\Delta Imp_c^{US,Ch}$	-0.7***	-0.7***	-4.5**	-1.3	-4.7**	-1.7	-3.9**	-1.7
	(0.10)	(0.24)	(1.93)	(1.25)	(2.10)	(1.26)	(1.71)	(1.59)
$\Delta Imp_c^{US,Ch} \times ln(pop_c)$			0.3**	0.1	0.3*	0.1	0.3*	0.1
			(0.15)	(0.11)	(0.17)	(0.11)	(0.14)	(0.14)
$ln(pop_c)$	-0.2**	-0.3	-0.8***	-0.4	-0.8***	-0.2	-1.0**	-0.8
	(0.09)	(0.16)	(0.30)	(0.35)	(0.28)	(0.34)	(0.40)	(0.74)
Time FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Region FE	No	No	No	No	Yes	Yes	Yes	Yes
Additional controls	No	No	No	No	No	No	Yes	Yes
FS residual	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	1444	1444	1444	1444	1444	1444	1444	1444
Pseudo R^2	0.13	0.50	0.19	0.50	0.22	0.54	0.41	0.58

Note: Robust standard errors clustered at the three digit SIC level are reported in parenthesis. The regressions are estimated using the control function approach include fixed effects for eight census regions. Regressions are weighted by initial employment in each commuting zone. The sample includes 722 commuting zones for the periods 1990 - 2000 and 2000 - 2007 that are stacked in the estimation. The population variable is demeaned such that $\Delta Imp_j^{US,Ch}$ is the effect of an import competition shock for the mean-sized commuting zone. Additional controls for the sectoral and demographic composition are included in some specifications. Stars indicate significance levels the following levels *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

where $\Delta \hat{Imp}_{ct}$ is a measure of changes in import competition, L_{ct} is the log population size of commuting zone c at the beginning of the period and δ_r are regional fixed effects to control for geographic features. I define the change in employment only based on sectoral rather than local changes as above and define the average import competition shock at the commuting zone level following Acemoglu et al. (2016):

$$\Delta log(L_c) = \sum_j \frac{L_{cj}}{L_c} \Delta log(L_j)$$

$$\Delta Imp^{Ot}_{ct} = \sum_j \frac{L_{cjt}}{L_{ct}} \frac{\Delta M^{Ot,Ch}_{jt}}{L_{jt-1}}$$

The main results are reported in Table 2.7. The results are highly statistically significant across all specifications and indicate that sectors located in more populated region experience a smaller decline in employment from the rise in Chinese exports to other countries.

Table 2.7: INDUSTRY-LEVEL MECHANISM

	$\Delta log(L_c)$	
ΔImp_c	-0.26***	-0.23***
	(0.026)	(0.027)
$ln(pop_c)$	-0.01**	-0.01***
	(0.004)	(0.003)
$\Delta Imp_c \times ln(pop_c)$	0.02**	0.02**
	(0.008)	(0.007)
Time FE	Yes	Yes
Region FE	No	Yes
Observations	1444	1444
Pseudo R^2	0.74	0.77

Note: Robust standard errors clustered at the commuting zone level in parenthesis. Regional fixed effects for eight regions within the US. Regressions are weighted by initial employment in each commuting zone. The sample includes 722 commuting zones for the periods 1991 - 1999 and 1999 - 2007 that are stacked in the estimation. The population variable is demeaned such that the constants represent the mean trade shocks for different time periods. Stars indicate significance levels the following levels *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

2.8 Conclusion

This paper documented a positive correlation between international economic integration and regional inequality within advanced economies. I also present three related stylized facts documenting higher export intensity in denser cities, which is both due to a within-sector and an across-sector margin. To microfound this aggregate correlation and the stylized facts I propose an economic geography model with spatial sorting of heterogeneous firms and heterogeneous sectors across different city sizes that features an open economy equilibrium with trade due to firm heterogeneity and endowment-driven comparative advantage. The model provides two channels that microfound the macro prediction that a decrease in trade cost increases the spatial concentration of economic activity, one on the firm level and one on the industry level. Firstly, within-industry trade reallocates market share and employment from less to more productive firms, since these more productive firms benefit more from agglomeration externalities, they are relatively located in larger cities. Hence, in the model this reallocation increases spatial concentration. Secondly, specialization due to endowment-driven comparative advantage increases employment in physical and human capital-intensive sectors for advanced economies. These sectors are located relatively more in denser cities as the relative price of capital to labour decreases with city density. Hence, in the model this reallocation increases spatial concentration.

I find empirical support for both channels proposed by the model using exogenous changes in export market access for French firms. Firstly, firms located in larger cities increase sales more following an increase in export market access. Secondly, sectors located in larger cities increase sales by more than those located in smaller cities. I additionally test the model predictions empirically using the rise in Chinese import competition for the United States. I find strong support for both mechanisms in this context as well.

I also provide evidence on the mechanisms as suggested by the model. For the sector-level mechanism including the sectoral skill intensity decreases the reduced-form correlation between export intensity and employment density significantly both statistically and economically. Including firm productivity in the estimations for the firm-level mechanism decreases the association between employment density and export intensity but only marginally. This suggest that there are export-specific agglomeration benefits, beyond the productivity benefits of agglomeration that contribute to firms in denser locations being more export intensity. Exploiting the structure of the model I find that variable trade costs indeed fall with city density while fixed trade cost increase with city density.

An additional implication of the model that I plan to explore in future work is that we overestimate the welfare effects of trade as long as we ignore its spatial implications. Estimating the gains from trade based on changes in tradables production and productivity alone does not account for the welfare losses due to the increase in congestion costs caused by increased spatial concentration.

This paper has provided causal evidence for two different channels through which international integration increases regional inequality and spatial concentration in advanced economies. While the previous literature has provided ample evidence for important distributional effects of trade across different skill groups, regional heterogeneity has been much less studied. These findings have important policy implications as they provide an additional margin for redistribution if the government aims to redistribute the aggregate gains from trade.

Appendix

2.A Additional specifications

Table 2.A.1: Cross-country correlation between trade openness and regional inequality/spatial concentration

	Regional inequality	
	Unweighted	Weighted by population
Openness	0.03^a (0.011)	0.04^b (0.021)
Year FE	Yes	Yes
Country FE	Yes	Yes
Observations	359	351
Pseudo R^2	0.95	0.91

Note: Robust standard errors clustered by country and year in parenthesis. The sample is an unbalanced panel of 26 countries for the period 1999 to 2014. Statistical significance levels are indicated by ^{c} for $p < 0.10$, ^{b} for $p < 0.05$, ^{a} for $p < 0.01$.

Table 2.A.2: Regressions corresssponding to panel (a) and (b) in figure 2.2

	Panel (a)		Panel (b)	
	Weighted (s)	Unweighted	Weighted (s)	Unweighted
log(emp dens)	0.023^a (0.0062)	0.003^a (0.0011)	0.009^c (0.0050)	0.002^b (0.0011)
Sector FE	No	No	Yes	Yes
Controls	Yes	Yes	Yes	Yes
Observations	136558	136558	136558	136558
Pseudo R^2	0.15	0.01	0.43	0.10

Cross-section in 1995. Standard errors clustered at the sector (four digit) level and the czone level in parenthesis. Controls are dummies for distance deciles to the Western and the Spanish border, and the Atl. and Med. coast. Statistical significance levels are indicated by ^{c} for $p < 0.10$, ^{b} for $p < 0.05$, ^{a} for $p < 0.01$.

Table 2.A.3: Regressions corresssponding to panel (c) figure 2.2

	Sectoral export intensity	
	Sales weighted	Unweighted
$log(empdens_{c95})$	0.015^a (0.0059)	0.012^b (0.0045)
Conrols	Yes	Yes
Observations	352	352
Pseudo R^2	0.63	0.20

Twenty-year difference from 1995 to 2015. Robust standard errors in parenthesis. Controls are dummies for distance deciles to the Western and the Spanish border, and the Atlantic and Mediterranean coast. Statistical significance levels are indicated by ^{c} for $p < 0.10$, ^{b} for $p < 0.05$, ^{a} for $p < 0.01$.

Table 2.A.4: Regressions corressponding to panel (a) in figure 2.2

	Share of export sales				
	Weighted (firm sales)				
log(emp dens)	0.013^b	0.017^b	0.005	0.010^b	0.033^a
	(0.0050)	(0.0070)	(0.0062)	(0.0049)	(0.0089)
Region FE (2d)	No	Yes	No	No	No
Region FE (1d)	Yes	No	No	No	No
Controls	No	No	No	Yes	Yes
Observations	136558	136558	136558	127207	117581
Pseudo R^2	0.08	0.21	0.00	0.04	0.17

Cross-section in 1995. Standard errors clustered at the sector (four digit) level and the czone level in parenthesis. Controls are dummies for distance deciles to the Western and the Spanish border, and the Atlantic and Mediterranean coast. Column (1) is the baseline, column (2) clusters at the sector and region level, (3) only at the sector level. Column (4) includes regional FE rather than standard controls, and (5) does not include any controls or FEs. Column (6) restricts the sample to czones where the export intensities between 5 and 50 %. Column (7) excludes Paris. Statistical significance levels are indicated by ^{c} for $p < 0.10$, ^{b} for $p < 0.05$, ^{a} for $p < 0.01$.

Table 2.A.5: Regressions corresssponding to panel (a) in figure 2.2

	Log(exports)		Exporter dummy		Log(exports) if exporting			Log(export share) if exporting		
	W (s)	Unw.	W (s)	Unw.	W (s)	W (e)	Unw.	W (s)	W (e)	Unw.
log(emp dens)	0.299^a (0.0941)	0.040^b (0.0188)	0.017^a (0.0054)	0.013^a (0.0035)	0.205^a (0.0711)	0.260^a (0.0658)	-0.081^a (0.0265)	0.070^b (0.0321)	0.073^a (0.0153)	0.012 (0.0270)
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	136558	136558	136558	136558	29212	29212	29212	29212	29212	29212
Pseudo R^2	0.11	0.02	0.03	0.03	0.20	0.52	0.02	0.07	0.16	0.01

Cross-section in 1995. Standard errors clustered at the sector (four digit) level and the czone level in parenthesis. Controls are dummies for distance deciles to the Western and the Spanish border, and the Atlantic and Mediterranean coast. Statistical significance levels are indicated by ^{c} for $p < 0.10$, ^{b} for $p < 0.05$, ^{a} for $p < 0.01$.

Table 2.A.6: Regressions corresssponding to panel (a) in figure 2.2

	Share of export sales								
log(emp dens)	0.023^a (0.0062)	0.023^a (0.0060)	0.023^a (0.0085)	0.023^b (0.0101)	0.023^b (0.0093)	0.023^a (0.0083)	0.023^a (0.0086)	0.023^b (0.0094)	0.023^a (0.0074)
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Std. err.	Ind, CZ	CZ	Ind × CZ	Ind, Reg	Reg	Ind × Reg	Ind	Ind (3d)	Rob.
Observations	136558	136558	136558	136558	136558	136558	136558	136558	136558
Pseudo R^2	0.15	0.15	0.15	0.15	0.15	0.15	0.15	0.15	0.15

Cross-section in 1995. Standard errors clustered at the sector (four digit) level and the czone level in parenthesis. Controls are dummies for distance deciles to the Western and the Spanish border, and the Atlantic and Mediterranean coast. Statistical significance levels are indicated by ^{c} for $p < 0.10$, ^{b} for $p < 0.05$, ^{a} for $p < 0.01$.

Table 2.A.7: Alternative specifications for regressions in table 2.5: Different controls

	Δ_{20} log(sales)							
	Unweighted		Sales$_{95}$ weighted		Unweighted		Sales$_{95}$ weighted	
Δ_{20}log(MA$_j$)	0.009		0.065		0.004		0.064	
	(0.0410)		(0.0444)		(0.0405)		(0.0448)	
Δ_{20}log(MA$_j$) ×	0.043^b	0.022^b	0.058^a	0.029^a	0.045^a	0.024^b	0.058^a	0.030^a
log(dens emp$_{c95}$)	(0.0167)	(0.0097)	(0.0150)	(0.0106)	(0.0170)	(0.0099)	(0.0149)	(0.0107)
log(emp dens)	-0.032^a	-0.018^b	-0.048^a	-0.026^a	-0.020^b	-0.007	-0.047^a	-0.026^a
	(0.0095)	(0.0070)	(0.0097)	(0.0099)	(0.0092)	(0.0071)	(0.0091)	(0.0098)
Sector FE	No	Yes	No	Yes	No	Yes	No	Yes
Reg FE (1d)	No	No	No	No	Yes	Yes	Yes	Yes
Observations	23355	23355	23355	23355	23122	23121	23122	23121
Pseudo R^2	0.01	0.06	0.02	0.12	0.02	0.08	0.02	0.12

Twenty-year difference from 1995 to 2015. Standard errors clustered at the sector (four digit) level and the czone level in parenthesis. Variables are winsorized at the 3rd and 97th percentile for each sector. Controls are dummies for distance deciles to the Western and the Spanish border, and the Atlantic and Mediterranean coast. Statistical significance levels are indicated by ^{c} for $p < 0.10$, ^{b} for $p < 0.05$, ^{a} for $p < 0.01$.

Table 2.A.8: Alternative specifications for regressions in table 2.5: Stacked ten year differences

	Δ_{10} log(sales)			
	Unweighted		Sales$_{95}$ weighted	
Δ_{10} log(MA$_{jt}$)	0.029 (0.0596)		0.060 (0.0695)	
Δ_{10} log(MA$_{jt}$) × log(dens emp$_{c95}$)	0.039^a (0.0129)	0.020^b (0.0094)	0.040^b (0.0185)	0.011 (0.0136)
log(emp dens)	-0.010^a (0.0027)	-0.005^b (0.0021)	-0.017^a (0.0044)	-0.010^b (0.0044)
Year FE	Yes	Yes	Yes	Yes
Sector FE	No	Yes	No	Yes
Controls	Yes	Yes	Yes	Yes
Observations	94969	94969	94969	94969
Pseudo R^2	0.02	0.05	0.03	0.11

Stacked ten-year difference from 1995 to 2005 and 2005 to 2015. Standard errors clustered at the sector-year (four digit) level and the czone-year level in parenthesis. Variables are winsorized at the 3rd and 97th percentile for each sector. Controls are dummies for distance deciles to the Western and the Spanish border, and the Atlantic and Mediterranean coast. Statistical significance levels are indicated by ^{c} for $p < 0.10$, ^{b} for $p < 0.05$, ^{a} for $p < 0.01$.

Table 2.A.9: Alternative specifications for regressions in table 2.5: Different controls

	Δ_{20} log(sales)			
	Unweighted		Sales$_{c95}$ weighted	
Δ_{20}log(MA$_j$)	0.69^a (0.147)	0.78^a (0.161)	0.63^a (0.150)	0.75^a (0.182)
Δ_{20}log(MA$_j$) × log(dens emp$_{c95}$)	0.29^a (0.105)	0.29^a (0.094)	0.53^a (0.134)	0.50^a (0.112)
Region FE	No	Yes	No	Yes
Controls	No	No	No	No
Observations	352	352	352	352
Pseudo R^2	0.11	0.18	0.18	0.26

Twenty-year difference from 1995 to 2015. Robust standard errors in parenthesis. Variables are winsorized at the 3rd and 97th percentile for each sector. Controls are dummies for distance deciles to the Western and the Spanish border, and the Atlantic and Mediterranean coast. Statistical significance levels are indicated by ^{c} for $p < 0.10$, ^{b} for $p < 0.05$, ^{a} for $p < 0.01$.

Table 2.A.10: Alternative specifications for regressions in table 2.5: Stacked ten year differences

	Δ_{10} log(sales)			
	Unweighted		Sales$_{c95}$ weighted	
Δ_{10}log(MA$_{jt}$)	0.36^a (0.049)	0.39^a (0.053)	0.31^a (0.084)	0.31^a (0.081)
Δ_{10}log(MA$_{jt}$) × log(dens emp$_{c95}$)		0.06^c (0.034)		0.03 (0.047)
log(emp dens)	-0.01^a (0.004)	-0.01^a (0.004)	-0.02^a (0.005)	-0.02^a (0.006)
Year FE	Yes	Yes	Yes	Yes
Controls	Yes	Yes	Yes	Yes
Observations	704	704	704	704
Pseudo R^2	0.19	0.20	0.18	0.18

Stacked ten-year difference from 1995 to 2005 and 2005 to 2015. Standard errors clustered at the year level in parenthesis. Variables are winsorized at the 3rd and 97th percentile for each sector. Controls are dummies for distance deciles to the Western and the Spanish border, and the Atlantic and Mediterranean coast. Statistical significance levels are indicated by ^{c} for $p < 0.10$, ^{b} for $p < 0.05$, ^{a} for $p < 0.01$.

Table 2.A.11: Alternative specifications for regressions in table 2.7: Stacked ten year differences, different controls

	Δ_{10} log(sales)			
	Unweighted		Sales$_{c95}$ weighted	
Δ_{10}log(MA$_{jt}$)	0.38^a (0.051)	0.38^a (0.052)	0.24^a (0.072)	0.26^a (0.077)
Δ_{10}log(MA$_{jt}$) × log(dens emp)	0.06^c (0.035)	0.06^c (0.033)	0.08^c (0.044)	0.07 (0.047)
Region FE	No	Yes	No	Yes
Controls	No	No	No	No
Observations	704	704	704	704
Pseudo R^2	0.06	0.02	0.10	0.07

Stacked ten-year difference from 1995 to 2005 and 2005 to 2015. Standard errors clustered at the year level in parenthesis. Variables are winsorized at the 3rd and 97th percentile for each sector. Controls are dummies for distance deciles to the Western and the Spanish border, and the Atlantic and Mediterranean coast. Statistical significance levels are indicated by ^{c} for $p < 0.10$, ^{b} for $p < 0.05$, ^{a} for $p < 0.01$.

Table 2.A.12: Regressions corressponding to figure 2.3

	$Log(rel\tau_{icj})$		$Log(rel\bar{\tau}_{cj})$
log(emp dens$_c$)	-0.027^a (0.0078)	-0.001 (0.0078)	-0.023^a (0.0081)
Ind FE	Yes	Yes	Yes
Observations	25481	25481	25481
Pseudo R^2	0.73	0.14	0.79

Cross-section in 1995. Sample only includes exporting firms. Standard errors clustered at the sector (four digit) level and the czone level in parenthesis. Controls are dummies for distance deciles to the Western and the Spanish border, and the Atlantic and Mediterranean coast. $Log(rel\bar{\tau}_{cj})$ is the export sales weighted variable trade cost. Statistical significance levels are indicated by ^{c} for $p < 0.10$, ^{b} for $p < 0.05$, ^{a} for $p < 0.01$.

Table 2.A.13: Regressions corressponding to figure 2.4

	City fixed effects estimated from Probit model							
	From unnweighted probit				From weighted probit			
log(emp dens$_c$)	1.56	1.51^a	1.19	1.12^a	0.41	0.49^b	0.49^c	0.52^a
	(0.98)	(0.49)	(0.69)	(0.36)	(0.33)	(0.19)	(0.26)	(0.13)
Ind FE (2d)	Yes	No	Yes	No	Yes	No	Yes	No
Ind FE (4d)	No	Yes	No	Yes	No	Yes	No	Yes
midrule Observations	85403	85403	85403	85403	85403	85403	85403	85403
Pseudo R^2	0.69	0.73	0.73	0.76	0.29	0.38	0.30	0.34

Cross-section in 1995. Sample only includes firms with an exporting firm in the corrsponding sector-czone pair. Standard errors clustered at the sector level (two or four digit, depending on level of fixed effect) and the czone level in parenthesis. Controls are dummies for distance deciles to the Western and the Spanish border, and the Atlantic and Mediterranean coast. Statistical significance levels are indicated by ^{c} for $p < 0.10$, ^{b} for $p < 0.05$, ^{a} for $p < 0.01$.

Table 2.A.14: First stage regressions of the trade shock coefficient corresponding to table 2.5

	$\Delta Imp_j^{US,Ch}$	$\Delta Imp_j^{US,Ch}$	$\Delta Imp_j^{US,Ch}$	$\Delta Imp_j^{US,Ch}$	$\Delta Imp_j^{US,Ch}$	$\Delta Imp_j^{US,Ch}$	$\Delta Imp_j^{US,Ch}$
$\Delta Imp_j^{Ot,Ch}$	1.22***	1.13***	1.11***	1.11***	1.21***	1.21***	1.21***
	(0.123)	(0.137)	(0.135)	(0.134)	(0.103)	(0.142)	(0.142)
$\Delta Imp_j^{Ot,Ch}$ $\times log(pop_c)$		0.03	0.03	0.03	0.02	0.02*	0.02*
		(0.022)	(0.021)	(0.021)	(0.011)	(0.009)	(0.009)
$ln(pop_c)$		-0.01	-0.00	0.01	-0.00	-0.00	-0.00
		(0.008)	(0.008)	(0.011)	(0.005)	(0.003)	(0.003)
Time FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Sub-sector FE	No	No	Yes	Yes	No	No	No
Region FE	No	No	No	Yes	No	No	Yes
Industry FE (3d)	No	No	No	No	Yes	No	No
Industry FE (4d)	No	No	No	No	No	Yes	Yes
Observations	129116	129116	129116	129116	129116	129116	129116
Pseudo R^2	0.63	0.63	0.65	0.65	0.79	0.88	0.88

Note: Robust standard errors clustered at the three digit SIC level are reported in parenthesis. The regressions include fixed effects for ten sub-sectors within manufacturing and eight census regions. Regressions are weighted by initial employment in each sector-commuting zone cell. The sample includes 392 manufacturing industries in 722 commuting zones for the periods 1991 - 1999 and 1999 - 2007 that are stacked in the estimation. The population variable is demeaned such that $\Delta Imp_j^{US,Ch}$ is the effect of an import competition shock for the mean-sized commuting zone. Stars indicate significance levels the following levels *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table 2.A.15: First stage regressions of the trade shock coefficient corresponding to table 2.5

	$\Delta Imp_j^{US,Ch} \times ln(pop_c)$	$\Delta Imp_j^{US,Ch} \times ln(pop_c)$	$\Delta Imp_j^{US,Ch} \times ln(pop_c)$	$\Delta Imp_j^{US,Ch} \times ln(pop_c)$	$\Delta Imp_j^{US,Ch} \times ln(pop_c)$	$\Delta Imp_j^{US,Ch} \times ln(pop_c)$
$\Delta Imp_j^{Ot,Ch}$	-0.03 (0.106)	-0.10 (0.123)	-0.09 (0.123)	0.14 (0.230)	0.17 (0.317)	0.17 (0.317)
$\Delta Imp_j^{Ot,Ch}$ $\times log(pop_c)$	1.26*** (0.125)	1.26*** (0.126)	1.26*** (0.126)	1.23*** (0.128)	1.22*** (0.125)	1.22*** (0.125)
$ln(pop_c)$	0.08** (0.038)	0.09* (0.048)	0.11** (0.050)	0.10** (0.048)	0.09** (0.046)	0.09** (0.045)
Time FE	Yes	Yes	Yes	Yes	Yes	Yes
Sub-sector FE	No	Yes	Yes	No	No	No
Region FE	No	No	Yes	No	No	Yes
Industry FE (3d)	No	No	No	Yes	No	No
Industry FE (4d)	No	No	No	No	Yes	Yes
Observations	129116	129116	129116	129116	129116	129116
Pseudo R^2	0.69	0.70	0.70	0.77	0.83	0.83

Note: Robust standard errors clustered at the three digit SIC level are reported in parenthesis. The regressions include fixed effects for ten sub-sectors within manufacturing and eight census regions. Regressions are weighted by initial employment in each sector-commuting zone cell. The sample includes 392 manufacturing industries in 722 commuting zones for the periods 1991 - 1999 and 1999 - 2007 that are stacked in the estimation. The population variable is demeaned such that $\Delta Imp_j^{US,Ch}$ is the effect of an import competition shock for the mean-sized commuting zone. Stars indicate significance levels the following levels *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table 2.A.16: Wage and employment regressions on the commuting zone level using 2SLS with log population

	ΔL_c	Δw_c	ΔL_c	Δw_c	ΔL_c	Δw_c	ΔL_c	Δw_c
$\Delta Imp_c^{US,Ch}$	-0.7***	-0.7***	-4.7***	-1.6	-4.7**	-1.8	-3.6**	-1.5
	(0.11)	(0.24)	(1.75)	(1.87)	(1.87)	(1.90)	(1.63)	(1.78)
$\Delta Imp_c^{US,Ch} \times ln(pop_c)$			0.3**	0.1	0.3**	0.1	0.2*	0.1
			(0.13)	(0.16)	(0.14)	(0.16)	(0.12)	(0.15)
$ln(pop_c)$	-0.2**	-0.3*	-0.8***	-0.4	-0.8***	-0.2	-0.9**	-0.7
	(0.10)	(0.15)	(0.26)	(0.42)	(0.24)	(0.41)	(0.38)	(0.74)
Time FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Region FE	No	No	No	No	Yes	Yes	Yes	Yes
Additional controls	No	No	No	No	No	No	Yes	Yes
Observations	1444	1444	1444	1444	1444	1444	1444	1444
Pseudo R^2	0.06	0.49	0.10	0.49	0.13	0.52	0.32	0.57
AP F-statistic ΔExp	95.15	95.15	3.56	3.56	2.89	2.89	2.80	2.80
AP F-statistic IA	.	.	4.21	4.21	3.01	3.01	3.55	3.55

Note: Robust standard errors clustered at the three digit SIC level are reported in parenthesis. The regressions include fixed effects for eight census regions. Regressions are weighted by initial employment in each commuting zone. The sample includes 722 commuting zones for the periods 1991 - 1999 and 1999 - 2007 that are stacked in the estimation. The population variable is demeaned such that $\Delta Imp_j^{US,Ch}$ is the effect of an import competition shock for the mean-sized commuting zone. Stars indicate significance levels the following levels *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table 2.A.17: Wage and employment regressions on the commuting zone level using 2SLS with absolute population

	ΔL_c	Δw_c	ΔL_c	Δw_c	ΔL_c	Δw_c	ΔL_c	Δw_c
$\Delta Imp_c^{US,Ch}$	-0.66***	-0.68***	-0.87***	-0.79***	-0.89***	-0.83***	-0.85***	-0.79***
	(0.097)	(0.256)	(0.123)	(0.217)	(0.134)	(0.186)	(0.208)	(0.258)
$\Delta Imp_c^{US,Ch} \times pop_c$			0.03***	0.01*	0.03***	0.01*	0.03***	0.02**
			(0.006)	(0.008)	(0.004)	(0.008)	(0.004)	(0.009)
pop_c	0.00	-0.02***	-0.06***	-0.05**	-0.07***	-0.05*	-0.08***	-0.07*
	(0.003)	(0.004)	(0.023)	(0.023)	(0.017)	(0.028)	(0.012)	(0.037)
Time FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Region FE	No	No	No	No	Yes	Yes	Yes	Yes
Additional controls	No	No	No	No	No	No	Yes	Yes
Observations	1444	1444	1444	1444	1444	1444	1444	1444
Pseudo R^2	0.06	0.51	0.24	0.52	0.29	0.54	0.46	0.59
AP F-stat: ΔImp	97.79	97.79	78.45	78.45	68.32	68.32	38.04	38.04
AP F-stat: IA	.	.	86.97	86.97	80.60	80.60	75.02	75.02

Note: Robust standard errors clustered at the three digit SIC level are reported in parenthesis. The regressions include fixed effects for eight census regions. Regressions are weighted by initial employment in each commuting zone. The sample includes 722 commuting zones for the periods 1991 - 1999 and 1999 - 2007 that are stacked in the estimation. The population variable is defined in units of 100,000 inhabitants and demeaned such that $\Delta Imp_j^{US,Ch}$ is the effect of an import competition shock for the mean-sized commuting zone. Stars indicate significance levels the following levels *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

2.B Theory

2.B.1 Proof of proposition 1

Note that the export intensity of firm i is given by: $r_{icj}^{X,int} = r_{icj}^{X}/r_{icj}^{D}$, which is equal to 0 if a firm does not export and, if a firm exports given by:

$$r_{icj}^{X,int} = \tau_j^{1-\sigma_j} \frac{(P_j^F)^{\sigma_j-1} R_j^F}{(P_j^H)^{\sigma_j-1} R_j^D} > 0$$

Define real productivity of firms in city size c in sector j as the measure of productivity that incorporates the city-specific marginal cost, which is given by: $\varphi_c(z) = \psi(z, L_{cj}^*(z))/w(L_{cj}^*(z))^{1-\alpha_j}$ and is increasing in city size. This follows immediately from the firm optimization problem. Since firm productivity is log-supermodular in raw efficiency and city size, firms with higher raw efficiency are located in larger cities. If two firms with different raw efficiency levels were located in the same city the firm with the higher raw efficiency would have higher real productivity and therefore make higher profits. Since it is optimal for this firm to locate in larger cities this must imply higher profits and hence higher real productivity. Therefore real productivity increases with city size. Note that city size and density are isomorphic in this model.

Since real productivity is increasing in city density there is a cut-off density $(\bar{L}_{cj}^X)$ below which firms will not enter the export market and above which firms will enter the export market. Hence, export intensity of firms is given by:

$$r_{icj}^{X,int} = \frac{r_{icj}^X}{r_{icj}^D} = \begin{cases} 0 & L_c < \bar{L}_{cj}^X \\ \tau_j^{1-\sigma_j} \frac{(P_j^F)^{\sigma_j-1} R_j^F}{(P_j^H)^{\sigma_j-1} R_j^D} & \bar{L}_{cj}^X \leq L_c \end{cases}$$

such that export intensity is an increasing function of city density.

2.B.2 Proof of proposition 2

We have shown in section 2.B.1 that real productivity increases with city size. Note that as in the standard Melitz model the productivity cut-offs in each sector are determined independently of the sector aggregates. Writing the free entry and the zero profit cut-offs condition for the closed economy in terms of real productivity yields:

$$\tilde{\kappa}_{1j}\rho^{-\alpha_j(\sigma_j-1)}\left(\varphi_c(z_j^{dc})\right)^{\sigma_j-1}P_j^{\sigma_j-1}R_j - f_{P_j}\bar{c}_j = 0$$
$$\int_{z_j^{dc}}\left[\tilde{\kappa}_{1j}\rho^{-\alpha_j(\sigma_j-1)}\left(\varphi_c\right)^{\sigma_j-1}P_j^{\sigma_j-1}R_j - f_{P_j}P\right]f(z_j)dz_j = \bar{c}_j f_{E_j}$$

Combining these two equations we can derive the raw efficiency cut-off for entry:

$$f_{P_j}J(z_j^{dc}) = f_{E_j}$$

where:

$$J(z_j^{dc}) = \int_{z_j^{dc}}\left[\left(\frac{\varphi(z_j)}{\varphi(z_j)}\right)^{\sigma_j-1} - 1\right]f(z_j)dz$$

We can derivie a similar expression for the raw efficiency cut-offs in the open economy. We need to impose the parameter restriction that $\tau^{1-\sigma_j}f_{X_j} > f_{P_j}$ which ensures the the raw efficiency cut-off for entry is below the raw efficiency cut-off for exporting. Combining the free entry condition with the zero profit cut-off conditions for entry and exporting yields:

$$f_{P_j}J(z_j^{do}) + f_{X_j}J(z_j^{xo}) = f_{E_j}$$

Comparing the conditions from the closed and the open economy it follows directly that $z_j^{dc} < z_j^{do}$ from the fact that J is decreasing in z. Hence the raw efficiency cut-off is higher in the open economy and therefore the minimum city size is larger.

The density of people living in a city of size L_c is given by:

$$f_L(L_c) = \kappa_4\frac{1}{\bar{N}}\sum_{j=1}^{S}\ell_j(z_j^*(L_c))\cdot M_j f_j(z_j^*(L_c))\frac{dz_j^*}{dlL_c}$$

where $\kappa_4 = 1/((1-b)(1-\eta))$ accounts for the employment in construction. $z_j^*(L_c)$ denotes the inverse matching function in sector j that allows us to express z_j as a function of L_c. $\ell_j(z_j^*(L_c))$ is the labour demand of a firm in sector j with a productivity level such it locates in city size L_c. M_j denotes the mass of firms in sector j. $f_j(z_j^*(L_c))\frac{dz_j^*}{dL_c} = f_j(z)$ is the density of firms in sector j that decides to locate in city size L_c. It follows from the definition of this density that if the spatial distribution of employment in every sector j in the open economy first-order stochastically dominates the spatial distribution of employment in the closed economy, then the city size distribution in the open economy first-order stochastically dominates the city size distribution in the closed economy. We will now prove that this is true for every sector j using the result by Dharmadhikari and Joag-dev (1983) that $X \underset{s}{>} Y$ if the density $g(Y)$ crosses the density $f(X)$ only once and from above. So the spatial distribution of the open economy denoted by density $f_L^o(L_c)$ first-order stochastically dominates the city size distribution in the closed economy with density $f_L^c(L_c)$ if $f_L^c(L_c)$ cuts $f_L^o(L_c)$ only once and from above. The densities can be written as:

$$\begin{aligned} f_j^c(L_c) &= \frac{1}{\bar{N}} M_j^c \ell^c(z_j^*(L_c)) f(z_j^*(L_c)) \frac{dz_j^*}{dL_c} \\ &= \frac{1}{\bar{N}} \frac{\tilde{\kappa}_{1j} {\rho_c}^{-\tilde{\alpha}_j} (\sigma_j - 1)(1-\alpha_j) \frac{\psi(z_j^*(L_c), L_c)^{\sigma_j - 1}}{w(L_c))^{(\sigma_j-1)(1-\alpha_j)+1}} f(z) \frac{dz_j^*}{dL_c} dz P_j^{\sigma_j - 1} R_j^c}{\sigma_j \tilde{\kappa}_{1j} \rho^{-\tilde{\alpha}_j} S_j(z_j^{dc}) P_j^{\sigma_j-1}} \\ &= \frac{1}{\bar{N}} \frac{(\sigma_j - 1)(1-\alpha_j)}{\sigma_j} \frac{R_j^c}{S_j(z_j^{dc})} \frac{\psi(z_j^*(L_c), L_c)^{\sigma_j-1}}{w(L_c)^{(\sigma_j-1)(1-\alpha_j)+1}} f(z_j^*(L_c)) \frac{dz_j^*}{dL_c} \end{aligned}$$

Similarly for the open economy:

$$\begin{aligned} f_j^o(L_c) &= \frac{1}{\bar{N}} M_j^o \ell^o(z_j^*(L_c)) f(z_j^*(L_c)) \frac{dz_j^*}{dL_c} \\ &= \frac{1}{\bar{N}} \frac{\tilde{\kappa}_{1j} {\rho_c}^{-\tilde{\alpha}_j} (\sigma_j - 1)(1-\alpha_j) \frac{\psi(z_j^*(L_c), L_c)^{\sigma_j - 1}}{w(L_c)^{(\sigma_j-1)(1-\alpha_j)+1}} f(z_j^*(L_c)) \frac{dz_j^*}{dL_c} P_j^{\sigma_j - 1} R_j^c}{\sigma_j \tilde{\kappa}_{1j} \rho^{-\tilde{\alpha}_j} S_j(z_j^{dc}) P_j^{\sigma_j-1}} \end{aligned}$$

Let's define the difference function $h(L_c) = f_j^o(L_c) - f_j^c(L_c)$. To show first-order stochastic dominance it is sufficient to show that $h(L_c)$ is weakly positive at the minimum of the support and negative at the maximum, and only changes sign once.

$$\begin{aligned} h(L_c) =& \frac{1}{\bar{N}} \frac{(\sigma_j - 1)(1-\alpha_j)}{\sigma_j} \frac{\psi(z^*, L_c)}{w(L_c)^{(\sigma_j-1)(1-\alpha_j)+1}} \frac{dz_j^*}{dL_c} \\ & \times \left(\frac{(\mathbb{1}_d^o(z^*) + \mathbb{1}_x^o(z^*)\tau^{1-\sigma_j}) R_j^o}{S_j(z_j^{do})\tau^{1-\sigma_j} S_j(z_j^{xo})} - \frac{\mathbb{1}_d^c(z^*) R_j^c}{S_j(z_j^{dc})} \right) \end{aligned}$$

Note that if $\mathbb{1}_A^k(z^*(L_c)) = \mathbb{1}_A^k(z^*(L_c + \Delta L_c))$ with $A = c, o$ and $k = d, x$ then $sign(h((L_c)) = sign(h(L_c + \Delta L_c))$. This relies on the result that the matching function is the same in the closed and the open economy. So changes in the sign of $h(L_c)$ that indicate that the density functions cut each other can only occur at the points where the indicator functions change. So we will separately analyse the sign in the four intervals intervals between the different cut-offs: $[0, z_j^{dc}), [z_j^{dc}, z_j^{do}), [z_j^{do}, z_j^{xo}), [z_j^{xo}, \infty)$.[13]

For the first interval we know that all indicator functions are zero since firms with a raw efficiency draw below z_j^{dc} will not enter any market.

$$h_1(L_c) = 0 \qquad for \quad z \in [0, z_j^{dc})$$

For values of z in the interval $[z_j^{dc}, z_j^{do})$, we know that $\mathbb{1}_d^o(z^*) = \mathbb{1}_x^o(z^*) = 0$ and $\mathbb{1}_d^c(z^*) = 1$, such that:

$$h_2(L_c) = \frac{1}{\bar{N}} \frac{(\sigma_j - 1)(1 - \alpha_j)}{\sigma_j} \frac{\psi(z^*, L_c)}{w(L_c)^{(\sigma_j-1)(1-\alpha_j)+1}} \frac{dz_j^*}{dL_c} \left(\frac{-R_j^c}{S_j(z_j^{dc})} \right) < 0$$

For the interval $[z_j^{do}, z_j^{xo})$ firms in the open economy become active as well with $\mathbb{1}_x^o(z^*) = 0$ and $\mathbb{1}_d^o(z^*) = \mathbb{1}_d^c(z^*) = 1$:

$$h_3(L_c) = \frac{1}{\bar{N}} \frac{(\sigma_j - 1)(1 - \alpha_j)}{\sigma_j} \frac{\psi(z^*, L_c)}{w(L_c)^{(\sigma_j-1)(1-\alpha_j)+1}} \frac{dz_j^*}{dL_c} \times \left(\frac{R_j^o}{S_j(z_j^{do})\tau^{1-\sigma_j} S_j(z_j^{xo})} - \frac{R_j^c}{S_j(z_j^{dc})} \right)$$

whose sign is ambiguous. I will therefore consider both possibilities that $h(L_c)$ is positive or negative on the interval $[z_j^{do}, z_j^{xo})$.

Note that $h(L_c)$ on the interval $[z_j^{xo}, \infty)$ (denoted h_4) is strictly larger than h_3:

$$h_4(L_c) = \frac{1}{\bar{N}} \frac{(\sigma_j - 1)(1 - \alpha_j)}{\sigma_j} \frac{\psi(z^*, L_c)}{w(L_c)^{(\sigma_j-1)(1-\alpha_j)+1}} \frac{dz_j^*}{dL_c} \times \left(\frac{(1 + \tau^{1-\sigma_j} R_j^o}{S_j(z_j^{do})\tau^{1-\sigma_j} S_j(z_j^{xo})} - \frac{R_j^c}{S_j(z_j^{dc})} \right)$$

Therefore if $h_3 > 0$ then $h_4 > 0$. This concludes the proof for first-order stochastic dominance if $h_3 > 0$.

If $h_3 < 0$, then $h_4 > 0$ has to be true because both $f_j^o(L_c)$ and $f_j^c(L_c)$ are density function over the same support such that one cannot be larger than the other for its entirety. This concludes the proof for first-order stochastic dominance if $h_3 < 0$, which concludes the proof of the proposition.

[13] The fact that $z_j^{do} < z_j^{xo}$ follows directly from imposing $\tau^{1-\sigma_j} f_{X_j} > f_{P_j}$

2.B.3 Proof of proposition 3

Note that in the absence of firm heterogeneity the trade component of the model simplifies to a Krugman (1980) model with Heckscher-Ohlin type comparative advantage in the spirit of Romalis (2004). To isolate the effects of differences in factor intensities we assume no differences in Hicks-neutral productivity, transport costs or the elasticity of substitution across sectors or countries.

Under these assumptions, we get the following expressions for the price of a variety, the sectoral price index and the quantity produced by each firm (each variable is denoted for sector j in country H and symmetric for all sector-country combinations):

$$p_j^H = \frac{\sigma}{\sigma - 1} \frac{1}{\psi} \rho_H^{\alpha_j} \bar{w}^{1-\alpha_j} w(L_{cj}^*)^{1-\alpha_j} \tag{2.26}$$

$$P_j^H = \left[n_j^H (p_j^H)^{1-\sigma} + n_j^F (\tau p_j^F)^(1 - \sigma) \right]^{\frac{1}{1-\sigma}} \tag{2.27}$$

$$q_j^H = q_j^F = \frac{(\sigma - 1) f}{w_{cj}^{1-\alpha_j}} \tag{2.28}$$

Since more capital-intensive sectors are located in larger cities, we want to show that the export intensity of sector j is higher than the export intensity of sector k ($r_j^{X,int} > r_k^{X,int}$) if sector j is more capital intensive than sector k ($\alpha_j > \alpha_k$) and the country is capital abundant.

Revenue in sector j from serving the foreign and domestic market are given by:

$$r_j^X = n_j^H p_j^H \tau^{1-\sigma_j} R_j^F (P_j^F)^{\sigma-1} \tag{2.29}$$

$$r_j^D = n_j^H p_j^H R_j^H (P_j^H)^{\sigma-1} \tag{2.30}$$

Export intensity in sector j is therefore given by:

$$r_j^{X,int} = \frac{r_j^X}{r_j^D} = \tau^{1-\sigma} \frac{R_j^F}{R_j^H} \left(\frac{P_j^F}{P_j^H} \right)^{\sigma-1} \tag{2.31}$$

The relative export intensity in sector j relative to sector k is given by:

$$r_{j/k}^{X,int} = \frac{r_j^{X,int}}{r_k^{X,int}} = \left(\frac{P_j^F P_k^H}{P_j^H P_k^F} \right)^{\sigma-1} \tag{2.32}$$

Using the definition of the price index (equation 2.27), it follows that $r_j^{X,int} > r_k^{X,int}$ if:

$$1 < \frac{n_j^H}{n_j^F}\frac{n_k^F}{n_k^H}\left(\frac{p_j^F}{p_j^H}\frac{p_k^H}{p_k^F}\right)^{\sigma-1} \tag{2.33}$$

For this inequality to hold it is sufficient that the relative number of Home firms is higher in capital intensive industries ($\frac{n_j^H}{n_j^F} > \frac{n_k^H}{n_k^F}$), and the relative price of varieties produced in home is lower in capital intensive industries ($\frac{p_j^H}{p_j^F} < \frac{p_k^H}{p_k^F}$).

Inserting this expression for the price of varieties (equation 2.26), it follows that $\frac{p_j^H}{p_j^F} < \frac{p_k^H}{p_k^F}$ holds if:

$$\frac{\rho_H}{\bar{w}_H} < \frac{\rho_F}{\bar{w}_F} \tag{2.34}$$

i.e. the relative price of capital is lower in home (H), the capital-abundant country, then in foreign (F).

Next we will show that in the trade equilibrium the locally abundant factors are relatively cheap. The factor market clearing conditions are given by:

$$\bar{w}^H\bar{L}^H = (\alpha_1\xi_1 w_{c1}^{-1}s_1 + \alpha_2\xi_2 w_{c2}^{-1}s_2)(R^H + R^F) \tag{2.35}$$

$$\rho^H\bar{K}^H = ((1-\alpha_1)\xi_1 s_1 + (1-\alpha_2)\xi_2 s_2)(R^H + R^F) \tag{2.36}$$

$$\bar{w}^F\bar{L}^F = (\alpha_1\xi_1 w_{c1}^{-1}(1-s_1) + \alpha_2\xi_2 w_{c2}^{-1}(1-s_2))(R^H + R^F) \tag{2.37}$$

$$\rho^F\bar{K}^F = ((1-\alpha_1)\xi_1(1-s_1) + (1-\alpha_2)\xi_2(1-s_2))(R^H + R^F) \tag{2.38}$$

Home is endowed with more capital and Foreign is endowed with more labour. For the full employment conditions to hold Home has to either have a larger share of the capital-intensive industry or to use capital more intensively in each industry. From the cost minimization problem of the firm and the resulting factor demands it follows that Home will only use capital more intensively in any industry if $\rho^H/\bar{w}^H < \rho^F/\bar{w}^F$. The share of home firms in world revenues in sector j is defined as:

$$s = \frac{n_j^H p_j^H q_j^H}{n_j^H p_j^H q_j^H + n_j^F p_j^F q_j^F}$$

Solving for s yields:

$$s = \frac{(R^H + \tau^{2-2\sigma}R^F) - \tilde{p}_j\tau^{1-\sigma}(R^H + R^F)}{(1+\tau^{2-2\sigma})(R^H + R^F) - (\tilde{p}^{\sigma} + \tilde{p}^{-\sigma})\tau^{1-\tau}(R^H + R^F)} \tag{2.39}$$

Home will only have a larger share of the capital-intensive industry if the price of varieties in the capital-intensive sector are cheaper in Home than in Foreign, which is only the case if $\rho^H/\bar{w}^H < \rho^F/\bar{w}^F$. Hence capital must be relatively cheaper in the Home country and the relative price of varieties in the capital intensive sector in the home country is cheaper than in the labour intensive sector:

$$\frac{p_j^H}{p_j^F} < \frac{p_k^H}{p_k^F} \tag{2.40}$$

which concludes the first half of the proof.

Next, I show that the relative number of Home firms is higher in capital intensive industries ($\frac{n_j^H}{n_j^F} > \frac{n_k^H}{n_k^F}$). Using monopoly pricing (2.26), the price index (2.27) and the quantity in equilibrium (2.28), we can express the relative number of firms in home as follows:

$$\frac{n_j^H}{n_j^F} = \frac{R^H + \tau^{2-2\sigma}R^F - \tilde{p}_j^{\sigma}\tau^{1-\sigma}(R^H + R^F)}{\tilde{p}_j(R^F + \tau^{2-2\sigma}R^H) - \tilde{p}_j^{1-\sigma}\tau^{1-\sigma}(R^H + R^F)} \tag{2.41}$$

where $\tilde{p}_j = p_j^H/p_j^F$ is the relative price of varieties in sector j produced in home relative to foreign, which is smaller in capital-intensive sectors than in labour-intensive sectors, as shown above. Since the relative number of firms (in Home) declines in the relative price of varieties, and the relative price of varieties is lower in the capital-intensive sectors, the relative number of firms is larger in the capital intensive sector. This concludes the second part of the proof, showing that the capital-intensive sector is more export intensive in the capital-abundant country. Since more capital-intensive sectors are located in larger cities this implies that sectors located in larger cities are more export intensive.

2.B.4 Proof of proposition 4

Note that in the absence of firm heterogeneity the trade component of the model simplifies to a Krugman (1980) model with Heckscher-Ohlin type comparative advantage in the spirit of Romalis (2004).To isolate the effects of differences in factor intensities we assume no differences in Hicks-neutral productivity, transport costs or the elasticity of substitution across sectors or countries.

As shown above (section 2.B.3) we can write the share of home firms' in world revenues (s) as (equation 2.39):

$$s = \frac{(R^H + \tau^{2-2\sigma} R^F) - \tilde{p}_j \tau^{1-\sigma}(R^H + R^F)}{(1+\tau^{2-2\sigma})(R^H + R^F) - (\tilde{p}^{\sigma} + \tilde{p}^{-\sigma})\tau^{1-\tau}(R^H + R^F)} \tag{2.42}$$

The share of firms of a given sector located in Home decreases in the relative price of varieties in that sector, as can be intutitively seen by evaluating the derivative at $\tilde{p} = 1$:

$$\left.\frac{\partial s}{\partial \tilde{p}}\right|_{\tilde{p}=1} = \frac{-\sigma\tau^{1-\sigma}}{(\tau^{1-\sigma}-1)^2} < 0$$

Note that the relative price of varieties is fully determined by the factor prices in the two countries (see equation 2.26), which themselves depend on the abundance of factors. Next we will show that in the trade equilibrium the locally abundant factors are relatively cheap and hence Home will capture a larger share of the market in the capital-intensive sector, while Foreign will predominantly export the labour-intensive good. The factor market clearing conditions are given by:

$$\bar{w}^H \bar{L}^H = (\alpha_1 \xi_1 w_{c1}^{-1} s_1 + \alpha_2 \xi_2 w_{c2}^{-1} s_2)(R^H + R^F) \tag{2.43}$$

$$\rho^H \bar{K}^H = ((1-\alpha_1)\xi_1 s_1 + (1-\alpha_2)\xi_2 s_2)(R^H + R^F) \tag{2.44}$$

$$\bar{w}^F \bar{L}^F = (\alpha_1 \xi_1 w_{c1}^{-1}(1-s_1) + \alpha_2 \xi_2 w_{c2}^{-1}(1-s_2))(R^H + R^F) \tag{2.45}$$

$$\rho^F \bar{K}^F = ((1-\alpha_1)\xi_1(1-s_1) + (1-\alpha_2)\xi_2(1-s_2))(R^H + R^F) \tag{2.46}$$

Home is endowed with more capital and Foreign is endowed with more labour. For the full employment conditions to hold Home has to either have a larger share of the capital-intensive industry or to use capital more intensively in each industry. From equation (2.39) we know that Home will only have a larger share of the capital-intensive industry if the price of varieties in the capital-intensive sector are cheaper in Home than in Foreign, which is only the case if $\rho^H/\bar{w}^H < \rho^F/\bar{w}^F$. From the cost minimization problem of the firm and the resulting factor demands it follows that Home will only use capital more intensively in any industry if $\rho^H/\bar{w}^H < \rho^F/\bar{w}^F$. Hence capital will be relatively cheaper in the Home country, which will export the capital-intensive good.

Next, we compare the factor allocation within Home across the autarky and the trade equilibrium. The factor market clearing conditions under autarky are given by:

$$\bar{w}^{HA}\bar{L}^{HA} = (\alpha_1\xi_1 w_{c1}^{-1} + \alpha_2\xi_2 w_{c2}^{-1})R^{HA} \tag{2.47}$$

$$\rho^{HA}\bar{K}^{HA} = ((1-\alpha_1)\xi_1 + (1-\alpha_2)\xi_2)R^{HA} \tag{2.48}$$

Combining factor market clearings in Home across the two equilibria (equations 2.35, 2.36, 2.47 and 2.48), we can show that the price of capital relative to labour is higher under trade if the following regularity condition hold:

$$\frac{(1-\alpha_1)}{(1-\alpha_2)}\frac{\alpha_2}{\alpha_1} < \frac{w_{c1}}{w_{c2}}$$

which ensures that the wage premium that firms in larger cities pay is small enough so that it does not imply factor intensity reversals across sectors. This condition holds under all reasonable parameter values. Given these differences in factor prices both sectors will use labour more intensively, which implies that the capital-intensive sector has to be larger and has a higher demand for both factors under the trade equilibrium to ensure full employment of factors. From the matching function it follows that the capital-intensive sector is located in a larger city than the labour-intensive sector. Hence, the re-allocation of employment from the labour- to the capital-intensive sector implies a reallocation in space to a larger city such that the spatial distribution of population in the open economy first-order stochastically dominates the spatial distribution of population in the closed economy.

3
Internal Migration and the Growth of Cities in Post-Apartheid South Africa[1]

[1] A different version of this chapter has been published as "Migration and Urbanisation in Post-Apartheid South Africa", IZA Discussion Paper No. 10113, with Chris Parsons and Ferdinand Rauch and has been accepted for publication by the *World Bank Economic Review*

Abstract

Although Africa has experienced rapid urbanization in recent decades, we know little about the process of urbanization across the continent. We exploit a natural experiment, the abolition of South African pass laws, to explore how exogenous population shocks affect the spatial distribution of economic activity. Under apartheid, black South Africans were severely restricted in their choice of location and many were forced to live in homelands. Following the abolition of apartheid they were free to migrate. Given a migration cost in distance, a town nearer to the homelands will receive a larger inflow of people than a more distant town following the removal of mobility restrictions. Drawing upon this exogenous variation, we study the effect of migration on urbanization in South Africa. While we find that on average there is no endogenous adjustment of population location to a positive population shock, there is heterogeneity in our results. Cities that start off larger do grow endogenously in the wake of a migration shock, while rural areas that start off small do not respond in the same way. This heterogeneity indicates that population shocks lead to an increase in urban relative to rural populations. Overall, our evidence suggests that exogenous migration shocks can foster urbanization in the medium run.

3.1 Introduction

Africa is the least urbanized continent, but its urbanization rate is catching up. The pace of urbanization is remarkable and the continent is due to overtake Asia as the fastest urbanizing region of the world within a decade (United Nations, 2014). Managing the challenges of this rapid transformation represents a key policy challenge and yet the evidence base, particularly in the case of Sub-Saharan Africa, remains limited. One central question for policy makers is to what degree this process can be managed by policy as opposed to being determined by fundamentals alone.

To illustrate this issue, consider a town experiencing an exogenous migration shock. Theoretically it can evolve in just three ways. First, the town's population could shrink back to the initial population level, i.e. mean-revert. Such a reaction would be consistent with an optimal urban network of relative city sizes, where relative sizes might be driven by location fundamentals. Secondly, the town's population could simply remain at the new increased population level and not adjust endogenously to the shock. In this case, the distribution of city sizes would be path dependent. Thirdly, the city could grow further. This would be consistent with a theory of agglomeration effects and multiple equilibria, where an initial population shock moves the town onto a new population trajectory growth path from one equilibrium to another. If city sizes behave according to the first scenario, policies to affect the location of people would be ineffective, while in the other two, policies that induce migration can in turn affect urbanization.

In this paper, we study how cities in South Africa behave having been exposed to exogenous population shocks following the abolition of apartheid. Under the apartheid regime, the black South African population was severely restricted in its mobility. Large parts of the population were forced to live in so-called homelands and townships. En-route to the democratic transition in 1994, these restrictions were lifted and in June 1991 black South Africans could move freely. Substantial internal migration flows resulted, which led to increased urbanization during the 1990s and

2000s (see Figure 3.2.1 below). We use the fact that the locations of the homelands resulted from a long historical process beginning in the 18th century (Lapping, 1986), which makes it plausible that, conditional on covariates, their location is quasi-random with respect to economic conditions today. Assuming the subsequent migration outflows from the homelands behave according to some migration cost in distance, we are able to exploit the exogenous variation from this positive migration shock to identify the effect of increased internal migration on the distribution of population in South Africa. In other words, assuming migration costs increase with distance, *ceteris paribus*, a town physically located nearer to a homeland is assumed to have received a larger inflow of previously mobility-restricted black migrants.[2] Hence, while the homelands are crucial to our empirical design, we do not study the development of population within the homelands.

Our main findings are threefold. First, we show that the distance to homelands is a strong predictor of black population growth in the years following the end of apartheid (i.e. our "first stage"). Second, we show that on average, an exogenous increase in population in a town leads to an increase in population by just that amount, in the medium to long run. This suggests that on average, the population distribution follows a path dependent process. Third, we find heterogeneous responses to exogenous population shocks across rural and urban areas. Population levels in areas with initially high population densities experience further agglomeration, i.e. exogenous immigration leads to population growth. Only in rural areas do we continue to find path dependence.[3] This suggests that a positive exogenous population shock generates a 'Matthew effect' ('those who have will be given'), as densely populated areas gain population relative to sparsely populated areas. We further investigate this heterogeneity by examining how the effect varies with both initial population density and the reduced-form magnitude

[2] Even if the distance cost of migration are small, there would be a distance coefficient if migrants "radiate" from their origin (Rauch, 2016).

[3] While standard models of trade and urbanization typically do not predict path dependence, recent studies that have found path-dependent behavior for small and medium sized towns include Bleakley and Lin (2012) and Michaels and Rauch (2018).

of the shock. For a given initial density, a larger shock leads to higher endogenous population growth. This is consistent with the idea that a significant shock is required to push a locality from its current equilibrium onto a new trajectory. These results imply that policies aiming to foster migration can further trigger urban agglomeration forces in high density areas.

South Africa's history lends itself to studying our research question and the country maintains excellent census data, both before and after Apartheid. One important limitation of the census data however, is that after apartheid many changes were made to various geographical and other definitions, which limits the comparability of our data before and after 1994, although the population data can be matched with some confidence on a level as fine as wards. Another drawback of our data on the regional level is that it does not identify internal migrants explicitly, so we have to infer differences in migration as differences in population growth conditional on covariates that account for differences in fertility and mortality. A third shortcoming is that no reliable information for population in homelands is available. For our purposes, information from outside homelands is sufficient. Hence we are unable to pinpoint the underlying micro-mechanisms driving our results.

The remainder of this chapter is organized as follows. Section 3.2 details the historical development of South Africa thereby providing evidence for the quasi-random location of the homelands. Section 3.3 discusses the related literature and introduces the theoretical thought experiment that serves as a framework for the empirical analysis presented in Section 3.4. The heterogeneous responses to a positive population shock are discussed in Section 3.5, and Section 3.6 concludes.

3.2 Historical background

Around two-thirds of South Africa's total population live in urban areas, making it one of the most urbanized countries in Africa. In the second half of the 20th century, urbanization in South Africa was shaped by the apartheid policy of the

National Party government (1948-1994). Apartheid - literally meaning "apart-ness" - was by its very nature a spatial concept (Christopher, 2001). The government aimed to completely separate the black and non-black populations.[4] Policies ranged from installing two town hall bathrooms to segregating city quarters and creating native reserves, the so-called homelands (or 'bantustans') that were to become independent states for the black population.

Segregation and mobility restrictions imposed on the black population had a long tradition in South Africa dating back to at least the 18th century (Lapping, 1986). The support for apartheid policies in the run-up to the 1948 elections, especially among poor white South Africans, resulted from the increasing black urbanization rate during the preceding decades. These dynamics derived from the expansion of manufacturing and labor shortages resulting from World War II (Ogura, 1996). It was generally believed that the problem of white poverty was linked to increasing black urbanization. The Native Economic Commission (1930-32) provides an example as it explicitly names black urbanization as a cause for greater levels of unemployment among low-skilled white people (Beinart, 2001, p.122). One of the main goals of the apartheid policies was therefore to prevent and reverse black urbanization, or to put it in the words of the Stallard Commission (1922): *'The Native should only be allowed to enter urban areas, [...], when he is willing to enter and to minister to the needs of the white man, and should depart therefrom when he ceases so to minister.'* (Feinstein, 2005, p.152). The policies that took shape after 1948 were therefore unique in aiming to achieve complete spatial and social segregation and were achieved by mobilizing significant government resources and displacing large numbers of black South Africans.

In order to control the movement of the black population, the government restricted blacks' rights to own land and their legal ability to settle where they wished. The literature distinguishes two dimensions of separation, 'urban apartheid' and

[4] We use the same terminology for racial categories as the census, namely 'Black' or 'African', 'Colored', 'Asian/Indian', and 'White', where the last three categories make up to the 'Non-black' category.

'grand apartheid' (Christopher, 2001). Urban apartheid aimed at creating separate quarters for that part of the black population that was allowed to stay permanently in urban areas. Grand apartheid rather aimed at moving the majority of the black population - that was not needed as laborers in white urban areas - to native reserves.

The three main measures to implement 'grand' and 'urban apartheid' were the Group Areas Act (1950), the Pass Laws Act (1952) and the Population Registration Act (1950). The latter assigned a population group to each citizen, which largely defined an individual's political and social rights. The Group Areas Act assigned a native reserve to each black population group and enabled the government to remove people that were not living in the area assigned to their population group. To control population flows and black urbanization in particular, the government relied on a pass system. The Pass Laws Act forced every black African to carry an internal passport at all times.[5] If a black African could not present their passport demonstrating their right to be in a particular region, they were subject to arrest.

These strictly enforced laws significantly constrained the distribution of population in space as well as the process of urbanization. According to the Surplus People Project (1985),[6] the South African government forcefully relocated at least 3.5 million people between 1960 and 1983. Additionally, several hundred thousand arrests were made every year under the pass laws (Beinart, 2001, p.158f). Table 3.2.1 displays the share of the black population living in urban and rural areas within South Africa and the homelands from 1950 to 1980. While the proportion living in urban areas in South Africa stayed roughly constant over the three decades, the proportion living in rural areas decreased by around 15%, while the homelands experienced a commensurate increase. These movements resulted in densely populated areas in the homelands that can be defined as urban in terms of population densities, but not in terms of public service delivery or industrial development. This 'dislocated

[5] It built on pre-apartheid legislation including the Natives Urban Areas Act from 1923 and Natives Urban Areas Consolidation Act from 1945, which forced every black man in urban areas to carry passes at all times.

[6] The Surplus People Project was a non-governmental organization that documented forced removals through the apartheid government.

urbanization' (Beinart, 2001), driven by government decisions instead of economic fundamentals, provides evidence of the substantial impact that the apartheid policies had on the distribution of population. Overall, while apartheid policies failed to reverse the level of urbanization of black South Africans, they were able to stop the trend towards increasing urbanization driven by economic growth and instead channel urbanization dynamics away from (white) cities and towards the homelands.

Table 3.2.1: Descriptive statistics on the population distribution

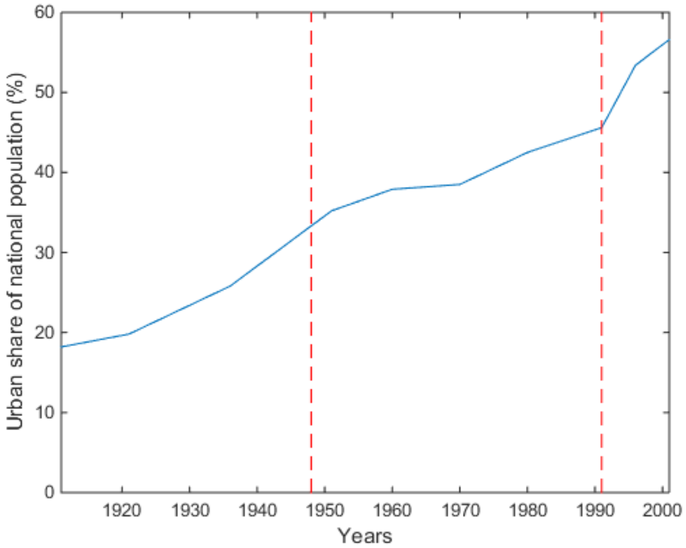
Distribution of black population across area types				Share of urbanised population across population groups				
Year	Urban	Rural	Homelands	Year	White	Colored	Indian	Black
1950	25.4	34.9	39.7	1951	78	65	78	27
1960	29.6	31.3	39.1	1960	84	68	83	32
1970	28.1	24.5	47.4	1980	88	75	91	49
1980	26.7	20.6	52.7	1991	91	83	96	58

Source: Surplus People Project (1985, p.18)

Table 3.2.1 also shows the share of the population living in urban areas during apartheid by population group. The three non-black population groups were already far more urbanized in 1951 and by 1991 around 90% of the non-black population resided in urban areas. The black population was predominantly living in rural areas in 1951 and urbanized until 1991, but remained significantly less urbanized than the other three population groups. As previously emphasized, this urbanization was heavily influenced by government policies that kept the black population out of urban areas in 'white' South Africa and engineered urbanization in the homelands. During the 1990s, urbanization rapidly increased (see Figure 3.2.1). Since the non-black population was almost entirely urbanized in 1991, this is evidence of large domestic migration flows of the black population.

Given the historical context, two main concerns arise regarding the proposed research design, which uses distance to the nearest homeland as an instrument for migration. First, that the location of homelands is non-random and that these

Figure 3.2.1: Urban share of the national population (%), 1911-2001



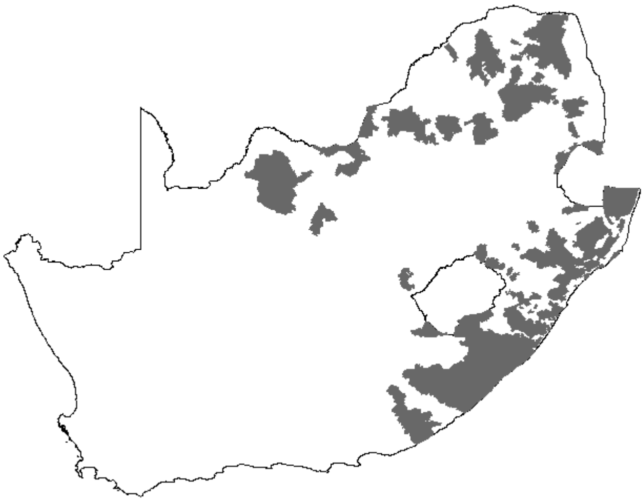
Note: Data from Turok (2012), vertical dashed lines mark the apartheid regime of the National Party (1948-1991)

could have for instance been located nearer to large industrial centers to serve as labor reservoirs. Secondly, that the constraint on internal mobility was binding.

The homelands established under apartheid (see Figure 3.2.2) were confined to areas designated as native reserves under the Native Land Act in 1913. This land comprised 7% of the overall area of South Africa and was already largely inhabited by the black population at the time, as the government was unwilling to expropriate white farmers. Hence the land allocation in 1913 failed to transfer large tracts of land between the different population groups and merely legally consolidated the distribution of land that had emerged predominantly through the European conquest of African land (Neame, 1962, p.40f). Since land was largely conquered for agricultural purposes, the African land reserves were of relatively low quality.

In 1913, South Africa was predominantly an agricultural economy with just two important industries - gold mining around Johannesburg and diamond mining around Kimberley. These industries established a system of migrant labor. Both found it optimal to change their entire workforce on a regular basis - every three

Figure 3.2.2: HOMELANDS (BANTUSTANS) ESTABLISHED UNDER APARTHEID



Source: Authors' own work. Bantustan boundary data from the Directorate: Public State Land Support via Africa Open Data

to six months - and wanted workers' families to remain in reserves. This allowed firms to pay low wages since the workers' families were supposed to find alternative work in the reserves (e.g. subsistence farming) which also reflected the (very low) opportunity cost of the worker. Additionally, they were able to send sick or injured workers back to the reserves where their tribe would take care of them (Lapping, 1986, p.26). This suggests that there was no need for specifically located labor reservoirs when the homelands where established. Therefore, no significant economic considerations appeared to have motivated the location of homelands, except for perhaps agricultural factors.

The 1936 Land Act and subsequent Government initiatives aimed at consolidating native territories to make them viable as independent states. There were no attempts to relocate them for economic reasons. One possible economic reason would be the proximity of cheap labor. Instead of relocating the homelands, the government created black townships such as Soweto to serve as labor reservoirs. If a homeland was conveniently located, many inhabitants commuted to work in white cities (KwaMashu and Umlazi in the homeland KwaZulu provide an example). There were therefore no incentives to relocate homelands as alternative ways to increase the pool of cheap labor proved more convenient.

A second concern when analyzing the switch from the constrained equilibrium for the black population under apartheid to the unconstrained equilibrium, is whether this constraint was binding. There are several observations that suggest that the constraint was indeed binding and that the switch to an unconstrained equilibrium was a significant shock to the distribution of population. First, the homelands were much poorer than other parts of South Africa. In 1985, GDP per capita in the homelands varied between 600 and 150 Rands, an order of magnitude below the 7,500 Rand estimated for the rest of South Africa (Christopher, 2001, p.93). Secondly, while more than 90% of whites and Indians lived in urban areas in 1986, less than 60% of blacks did, and we observe a large jump in urbanization starting in the 1990s. Thirdly, while keeping blacks out of urban areas was one of the major goals of apartheid policy, the absolute level of the black population in 'white' urban areas nevertheless increased. This suggests that strong urban attraction pulled blacks into urban areas, while apartheid reduced the rate of urbanization (Feinstein, 2005, p.157).

3.3 Related literature

This paper relates to a number of literatures, in particular to the increased interest in cities and urban planning from the perspective of economic development. While urbanization already plays a central role in many of the seminal contributions in the early development economics literature (see for example Lewis (1954), Ranis and Fei (1961), Harris and Todaro (1970)), there have been a number of recent empirical and theoretical contributions studying the determinants of urbanization, as well as its effect on economic growth. Recent work by Henderson (2005) suggests that urban growth may be a necessary condition for GDP growth. Potts (2012) and Gollin et al. (2016) draw upon census data and economic theory to show the importance of natural resources as a determinant of city sizes in Africa, and raise questions about differences between urbanization in Africa and elsewhere.

We add to this literature by studying how the urban system in South Africa reacts to an exogenous population shock. The main policy question we address here is the degree to which population flows within a country can be managed in the medium to long run. To illustrate this policy question, consider a town of initial population N_0, in a country with no population growth, that is given an exogenous population increase of Δ to $N_0 + \Delta$ people. There are only three ways in which the population of this town can respond in the long run. First, there could be mean reversion to the original relative population level, such that the long run population is smaller than $N_0 + \Delta$. Secondly, there could be a random walk process that generates path dependence, such that the long run expected value of the size of the town is now $N_0 + \Delta$. Thirdly, it could be that the additional population generates agglomeration effects and triggers a process in which it gains a long run population greater than $N_0 + \Delta$.

These three possibilities can be captured in an economic model following Henderson (1974), as demonstrated recently by Bleakley and Lin (2015). Let us consider a simple version of these models here to illustrate the key point.

In this model, agents derive utility from locating in particular areas. In equilibrium, there cannot be any gains from mobility, such that the utilities of all agents have to be equal across locations. Utility stems from the difference between the agglomeration ($A(N)$) and the congestion cost ($C(N)$) curves that are both functions of population density (N).[7] Spatial utility in region i is defined as: $U(N^i) = A(N^i) - C(N^i)$. The agglomeration curve summarizes the consumption gains from a greater number of varieties as well as higher wages resulting from productivity gains due to agglomeration effects. The congestion cost curve is determined by rents and commuting costs. The population allocation equilibrium

[7] In the context of this stylized model, we use changes in the population level and changes in population density interchangeably, since we consider a fixed amount of space.

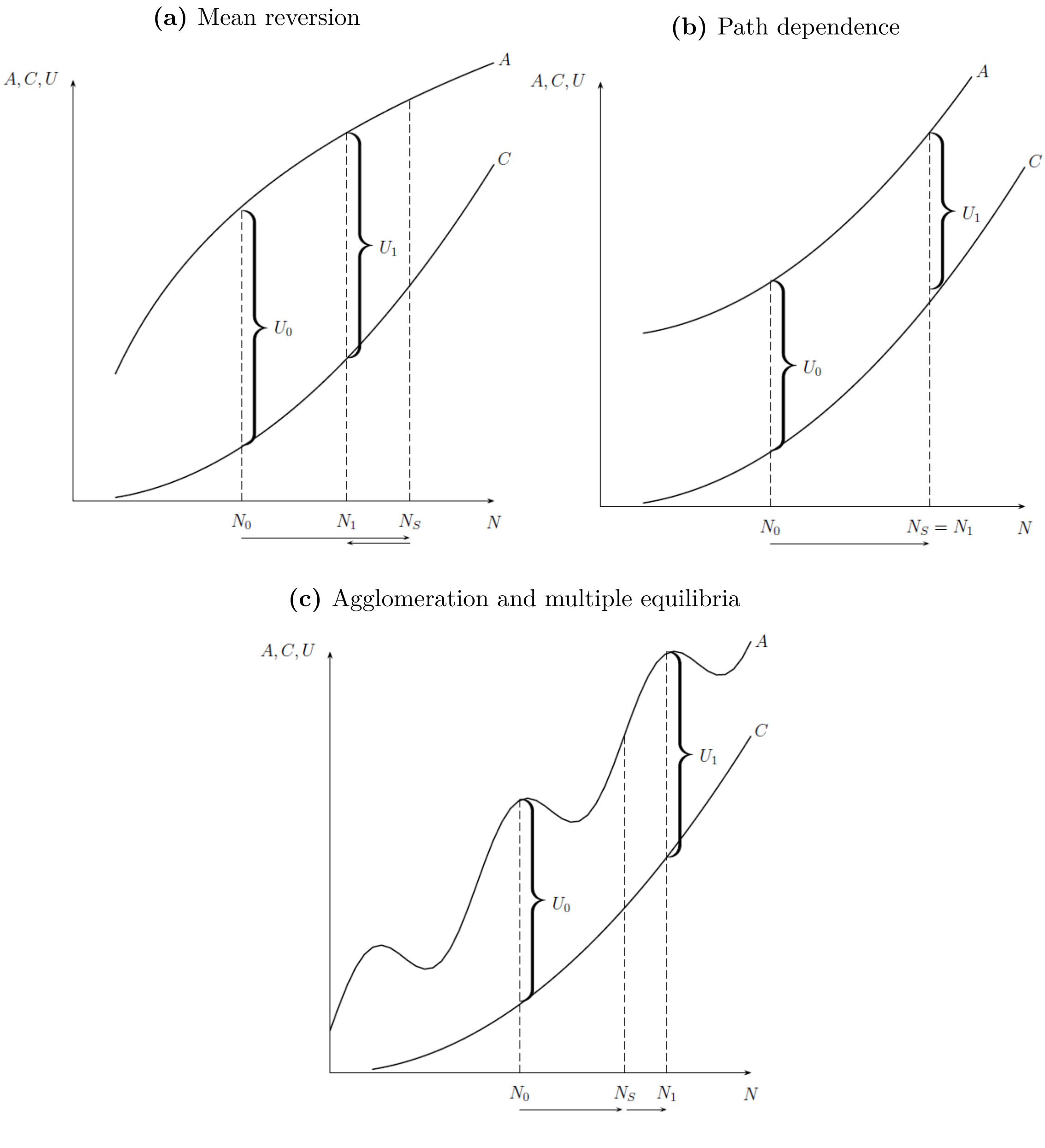
is determined by the indifference condition that the spatial utility from locating in a certain area has to be equalized across all K areas: $U(N^1) = ... = U(N^K)$. When assuming a particular functional form for one of the two functions, we can infer characteristics of the shape of the other function from the three hypotheses outlined above.

There are no intuitive guidelines as to the shape of the agglomeration curve as a function of population density $A(N)$. The agglomeration function could be non-monotonic, as new industries emerge to replace others when the population level crosses some threshold. This could result in significant changes in the structure of the local economy. For the congestion cost curve $C(N)$ on the other hand, given finite space, it is plausible to assume that it is increasing ($C'(N) > 0$) in population density, convex ($C''(N) > 0$) and tending to infinity after a certain population density threshold has been reached ($lim_{N \longrightarrow \bar{N}} C(N) = \infty$). Given these assumptions on the congestion cost function, different shapes for the agglomeration function follow from the three hypotheses outlined above.

The definition of equilibrium implies that the utility across locations has to be equal before the shock hits.[8] Population movements reacting to the exogenous shock again have to equalize the utility across locations to attain a new equilibrium. In the empirical analysis, all areas are treated by a population shock with varying intensities that depend upon their geographical proximity to the homelands. The utility level in the initial equilibrium is denoted by $U(N_0)$. $U(N_S)$ is the utility level after the shock and $U(N_1)$ is the utility level of the new equilibrium after agents have adjusted their location decisions. By the definition of equilibrium, $U(N_0)$ and $U(N_1)$ have to be equal across all locations (treated and control), while $U(N_S)$ is not related to an equilibrium and can therefore vary across locations.

[8] We are implicitly assuming that 'white' South Africa was in spatial equilibrium before the positive population shock. However, this is obviously not the case given 'urban apartheid' and the movement restrictions imposed on the non-white population within South Africa. Since these restrictions were in place in all of 'white' South Africa they are orthogonal to the population treatment and so we abstract from them in the model for simplicity.

Figure 3.3.1: Modified Henderson model with gains from agglomeration and congestion costs



(a) Mean reversion

(b) Path dependence

(c) Agglomeration and multiple equilibria

The population level mean reverts (Panel A in Figure 3.3.1)[9] if the utility at the new population level $U(N_S)$ is below the utility in the initial equilibrium $U(N_0)$ that can be attained in the untreated areas. Agents move from treated areas to the control areas until the utilities are equalized across both types of locations $U(N_1^T) = U(N_1^C)$. This leads to a reduction of population below N_S. This implies that the slope of the agglomeration function has to be locally shallower than the

[9] Note that the graphs only display the evolution of population in treated areas.

slope of the congestion cost function. The evolution of population is path dependent (Panel B) if the utility at the population level after the shock is equal to the utility in the initial equilibrium $U(N_0) = U(N_S)$. This implies that there are no gains from moving between control and treatment areas and that therefore there is no endogenous adjustment of location decisions, such that the new population level is an equilibrium population level: $N_S = N_1$. Since the difference between the agglomeration and congestion functions at N_0 is equal to the difference at N_S, the slopes of the two functions between N_0 and N_S have to be equal. If this property holds globally, then there are infinitely many equilibria of the spatial distribution of population. In the case of agglomeration (Panel C), agents move from control areas to treated areas. This implies that gains from migration exist, such that the utility level after the shock has to be greater than the utility in the initial equilibrium: $U(N_0) < U(N_S)$. Utility could not be strictly increasing in population between N_0 and N_S however, because that would imply the existence of gains from migration at N_0. The existence of such gains would contradict the definition of an equilibrium, such that N_0 could not be an equilibrium. For N_0 to be an equilibrium therefore, utility has to be non-monotonic, implying the existence of multiple equilibria for the spatial distribution of population. In order for the utility function to be non-monotonic, the slope of the agglomeration function has to be non-monotonic.[10] This model demonstrates the three possible cases between which we aim to distinguish in this paper. Since the *local* slope of agglomeration and congestion function determine which case applies, the model also suggests that the reaction might be different at towns of different initial population densities and motivates us to study heterogeneity along this dimension. It is clear that understanding the slope of these curves is of central importance to policy makers trying to adjust the size of cities and towns.

[10] Note that the functional form displayed in Panel C is just one example of a broad class of possible agglomeration functions. In particular, it is not necessary for the slope of $A(N)$ to be locally negative for the existence of multiple equilibria.

This setup relates to a large empirical literature that studies how exogenous shocks to cities affects the long-run development of affected areas. Studying the population of Japan, Davis and Weinstein (2002) find that population tends to mean-revert after a shock thereby concluding that location fundamentals play an important role. Studies by Brakman et al. (2004) for post-war Germany and by Miguel and Roland (2011) for Vietnam arrive at similar results using destruction resulting from wars. Bleakley and Lin (2012) analyse path dependence by studying the development of towns that experienced a negative shock to their fundamentals. Their main result is that former portage cities maintained their historical importance, a finding consistent with recent results from local positive population shocks from German refugees after World War II that were highly persistent and could not be explained by location fundamentals (Schumann, 2014). Using the same natural experiment, Peters (2017) shows that income per capita, overall manufacturing employment and the entry of new plants are positively correlated with refugee inflows in Germany after World War II.

We contribute to this literature in several ways. First, our study is the first to analyze a large scale and indeed positive population shock. The aforementioned studies analyzing the effects of war, find evidence of path-dependence but cannot isolate whether this is driven by natural fundamentals, sunk investments, social networks, capital unaffected by shocks, or gains from agglomerations. Bleakley and Lin (2012) provide evidence that it is not driven by location fundamentals, but cannot distinguish between other factors. Since we analyze a positive population shock in which incoming migrants have neither social networks nor private sunk investments, we are able to isolate the effect of gains from agglomeration. Secondly, we provide evidence from a credible natural experiment that is well-identified and are able to draw upon a much larger sample in comparison with most studies in this literature. Thirdly, we are the first to provide evidence from Africa, a region that is amongst the most rapidly urbanizing regions in the world, the continent in which such policies related questions matter most. Fourthly, many of the previous studies focus solely upon urban areas, whereas we are able to look at both rural and urban areas and the differences between the two.

There are other studies that exploit the exogenous variation resulting from apartheid policies to study the development of South Africa after 1994, see for example de Kadt and Sands (2016), de Kadt and Larreguy (2018) and Dinkelman (2011, 2017). We are the first to use this natural experiment to study the causal economic effect of internal migration and how it effects the distribution of population across space. This paper, thereby, adds quasi-experimental evidence to the literatures that examine the determinants of the uneven distribution of population across space and the relationship between city size and population growth (e.g. Black and Henderson (2003), Eeckhout (2004), Duranton (2007), Rauch (2013) and Rossi-Hansberg and Wright (2007)). After Auerbach (1913) observed that the size distribution of cities follows a power law, there have been many attempts to explain this persistent empirical regularity (often referred to as Zipf's Law, after Zipf (1949)). Following the theoretical work by Gabaix (1999) who showed that Zipf's Law emerges naturally if cities have equal relative growth rates (Gibrat's Law), an extensive empirical literature on the distribution of population has developed. The majority of empirical studies find urban systems tend to obey Gibrat's Law and that city size is uncorrelated with population growth, while others find departures from Gibrat's Law even for cities (Soo (2007), González-Val et al. (2013), and Holmes and Lee (2010)). Michaels et al. (2012) emphasize the importance of structural transformation for urbanization. In their long-run study of population growth in the United States from 1880 to 2000, they find that areas with high initial population density obeyed Gibrat's Law, i.e. subsequent population growth was uncorrelated with initial population density. Our research contributes to this literature by demonstrating the heterogeneous reaction of towns of different size to a population shock.

3.4 Empirical analysis

Data

In order to empirically test the three hypotheses, we make use of a unique geographically referenced South African census dataset. It contains observations for the years 1991, 1996, 2001 and 2011 at the ward level and hence bridges across the democratic transition in 1994. This dataset consists of two parts. First, it contains publicly available census data aggregated to the ward level for the censuses in 1996, 2001 and 2011 provided by Statistics South Africa. This allows us to distinguish between the short-, medium- and long-run effects of the exogenous population shock. Secondly, it contains data from the last census under the apartheid government in 1991. De Kadt and Sands (2016) matched a partial enumerator area map from the census in 1991 with the 100% sample of the individual level census data made available by DataFirst at the University of Cape Town and aggregated it to the 2011 ward level. This last census was implemented in March 1991. This timing is crucial as the Native Land Act, the Population Registration Act and the Group Areas Act were repealed in June 1991.

While the Pass Laws Act had already been repealed in 1986 and although identification would be cleaner if data from before 1986 were available, this timeline does not pose a major threat to our identification strategy. This is because the Group Areas Act and the Population Registration Act were still in place, and as such the black population was still severely constrained in its choice of residence until June 1991.[11]

[11] The data from 1991 does not cover the entirety of South Africa. One general drawback of the dataset is that it does not cover the homelands. This does not affect the analysis since we only look at areas outside the homelands. Another more relevant drawback is that there are a few areas that are not covered within South Africa (see Figure 3.A.1 appendix 3.A). This is due to two reasons. First, Statistics South Africa only has a partial map of the census enumeration areas in 1991. Therefore, part of the census data cannot be geographically referenced. Secondly, due to violent turmoil at the time, some areas could not be visited by enumerators and no data are available on a granular level. This is potentially beneficial for our analysis since we exclude areas with high racial tension, which could otherwise bias our results. So, while this reduces the number of observations and therefore the statistical power in the empirical analysis, our parameter

Identification

Distance to the nearest homeland is used as an instrument for migration flows in order to causally identify the effect of migration on population distribution. Figure 3.4.1 shows that the relationship between this distance and population growth between 1991 and 2011 is strongly negative, both at short and longer distances. In this figure we pool neighboring observations into discrete bins to improve clarity. We specify 100 bins in total, which puts around 20 observations into each bin. The log linear specification seems a good fit for the data. The validity of the instrument relies on the conditional quasi-random allocation of homelands, which has been argued for in Section 3.2. The assumption may be violated for areas adjacent to the homelands however, as they are likely to be affected by economic spillovers from the neighboring homeland in a variety of ways that are not related to the cost of out-migration from the homelands. To adjust for this problem, we exclude areas within 10 km from the homelands as a robustness check to ensure that the estimates are not driven by local economic spillovers.

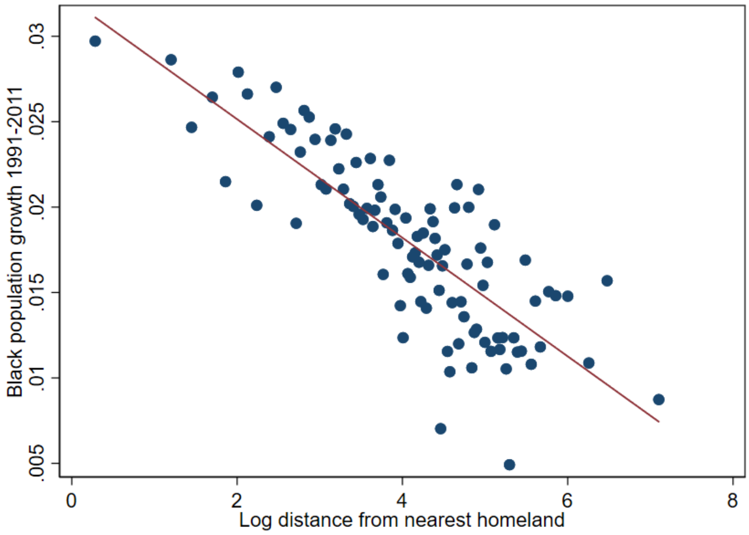
For our instrument to be informative, the cost of migration has to increases substantially with distance, which would imply that a town located nearer to the homelands *ceteris paribus* receives more migrants than a town further away. This assumption, consistent with the gravity framework, is a common assumption in the migration literature and the informativeness can be tested empirically by looking at the partial correlation between the instrument and the endogenous variable.[12] The informativeness of distance as an instrument crucially depends upon the level of fixed effects chosen, which affects the variation in the data. As shown in Table 3.A.1 in appendix 3.A, the informativeness of the instrument decreases almost monotonically in the granularity of the fixed effects.[13] A trade-off therefore exists

estimates will remain consistent.

[12] For example, Peri (2012) uses distance to the Mexican border as an instrument for the intensity of migration to different US states.

[13] This is intuitive, since for example when using municipality level fixed effects, the identifying variation of the instruments explains in which part of Johannesburg migrants are going to settle. This is likely to be uncorrelated with distance to homelands especially for urban areas.

Figure 3.4.1: First stage



Note: Relationship between distance to the homelands and black population growth, conditional on the covariates used in the main specification in Table 3.5.1. Data is collapsed into 100 bins, representing roughly 20 wards each.
Source: Author's analysis based on South African census data.

between accounting for local unobservables and retaining sufficient identifying variation, in order to ensure that our instrument remains informative. We include province level fixed effects in order to account for different trends and policies across provinces while allowing for sufficient spatial variation.[14]

Estimation

To estimate the causal effect of migration on the distribution of population, the following system is estimated using two-stage least squares (2SLS):

$$\Delta N^{B}_{i,t} = \alpha_2 + log(distance_i)\pi + \boldsymbol{X_{i,1991}}\gamma_2 + \delta_p + \upsilon_m \tag{3.1}$$

$$\Delta N_{i,t-1991} = \alpha_1 + \widehat{\Delta N^{B}_{i,t}}\beta + \boldsymbol{X_{i,1991}}\gamma_1 + \delta_p + \epsilon_m \tag{3.2}$$

where $\Delta N_{i,t-1991}$ denotes overall population growth in ward i between 1991 and

[14] Provinces are equivalent to states in the US.

t and $\Delta N^{B}_{i,t-1991}$ denotes black population growth. Our controls include: population groups, population density, education, the gender ratio, employment and income ($\boldsymbol{X_{i,1991}}$) and province-level fixed effects (δ_p) (see Table 3.4.1). The errors are clustered at the municipality level.[15] The ward level is the lowest geographical level that can be tracked consistently over time and municipalities are the lowest level of local government. Distance is defined as the distance to the nearest homeland measured from centroid to centroid. Since no measure of domestic migration is available in the census data, black population growth conditional on fixed effects and covariates is used as a proxy for black migration.[16] A dummy variable for Cape Town is also included as the municipality is a special case in terms of location, politics and demographics and hence migration patterns. The Western Cape was the only province where the African National Congress did not come first in the general elections in 1994. Until today, it has not achieved the political dominance in the province or the municipality of Cape Town that it has in the rest of the country. In terms of demographics, there is a much higher white and especially colored population in Cape Town, more than anywhere else in the country. Most importantly, there is a lot of circular migration from the Eastern and the Northern Capes into Cape Town. These migration dynamics potentially distort our identification strategy such that we include a dummy for Cape Town, which significantly increases the predictive power of our instrument. The results are robust to not including a dummy for Cape Town.

Linking theory and the variable of interest in the empirical estimation

In order to test the three competing hypotheses outlined in Section 3.3, it proves crucial to link the predictions from the hypotheses to the parameter of interest β. If the underlying process was driven by mean reversion, then the effect of

[15] Using Conley (1999) standard errors to account for spatial correlation yields similar results.

[16] As has been used previously as a proxy for migration status, see for example Czaika and Kis-Katos (2009).

Table 3.4.1: Summary statistics of included variables

VARIABLES	(1) N	(2) mean	(3) sd	(4) min	(5) max
Excluded instrument					
log distance	2,093	4.092	1.564	0.0529	6.746
Endogenous variables					
ΔBlack Population (1991-1996)	2,093	-1.081	20.75	-615.4	0.200
ΔBlack Population (1991-2001)	2,093	0.0310	0.0381	-0.200	0.100
ΔBlack Population (1991-2011)	2,093	0.0179	0.0207	-0.168	0.0499
Dependent variables					
ΔTotal Population (1991-1996)	2,093	-1.787	29.81	-829	0.200
ΔTotal Population (1991-2001)	2,093	0.0348	0.0417	-0.358	0.1000
ΔTotal Population (1991-2011)	2,093	0.0214	0.0213	-0.168	0.0500
ΔNonblack Population (1991-1996)	2,093	-0.709	18.51	-773.8	0.192
ΔNonblack Population (1991-2001)	2,093	0.00380	0.0211	-0.294	0.0960
ΔNonblack Population (1991-2011)	2,093	0.00354	0.00989	-0.0669	0.0425
Province fixed effects					
Eastern Cape	2,093	0.100	0.301	0	1
Free State	2,093	0.0994	0.299	0	1
Gauteng	2,093	0.172	0.378	0	1
KwaZulu-Natal	2,093	0.145	0.352	0	1
Limpopo	2,093	0.0674	0.251	0	1
Mpumalanga	2,093	0.102	0.302	0	1
North West	2,093	0.0726	0.260	0	1
Northern Cape	2,093	0.0717	0.258	0	1
Western Cape	2,093	0.170	0.375	0	1

Source: Author's analysis based on South African census data.

the exogenous population shock as measured by β would be less than one and decreasing over time, as the shock dissipates through the urban system. In the case of path dependence, β would be expected to be equal to one in all periods. In the agglomeration scenario, β would be significantly greater than one.

In order to assign these theoretical interpretations to the estimated parameters, we have to avoid using percentage growth rates in the endogenous variable and in the dependent variables in the second stage. Using percentage growth would make the shock a function of the share of black population that already lives in

Table 3.4.1 - continued

	(1)	(2)	(3)	(4)	(5)
VARIABLES	N	mean	sd	min	max
Control variables (from 1991 census in logs)					
Male share	2,093	0.506	0.0826	0	1
Population group ratio	2,093	0.275	0.321	0	1
Population density	2,093	4.647	2.478	0.0148	10.40
Total population	2,093	8.282	1.197	0.693	10.45
Black population	2,093	6.712	1.997	0	9.944
Employed	2,093	7.261	1.306	0	9.635
Unemployed	2,093	4.977	1.389	0	8.183
Not economically active	2,093	7.665	1.258	0	9.925
No schooling	2,093	6.864	1.154	0	9.317
Some primary schooling	2,093	6.807	1.188	0	9.135
Finished primary school	2,093	5.414	1.168	0	8.255
Some secondary schooling	2,093	6.792	1.333	0	9.585
Finished secondary school	2,093	5.952	1.578	0	9.404
Higher education	2,093	3.442	1.899	0	8.283
No income	2,093	7.618	1.250	0	9.932
Income: R1-499	2,093	3.655	1.391	0	7.349
Income: R500-699	2,093	3.266	1.285	0	6.852
Income: R700-999	2,093	3.775	1.246	0	6.952
Income: R1000-1499	2,093	4.610	1.246	0	7.594
Income: R1500-1999	2,093	4.604	1.228	0	7.489
Income: R2000-2999	2,093	5.188	1.302	0	7.856
Income: R3k-4k	2,093	5.111	1.280	0	7.953
Income: R5k-6k	2,093	4.706	1.259	0	8.084
Income: R7k-9k	2,093	4.786	1.396	0	8.357
Income: R10k-14k	2,093	4.874	1.498	0	9.125

Source: Author's analysis based on South African census data.

the area, which makes the interpretation we are looking for impossible. Therefore we instead define our variables of interest as absolute growth rates relative to the overall population in period t where t corresponds to 1996, 2001 and 2011. We incorporate this normalizing factor since the size of the population shock should be measured relative to the overall population rather than in absolute terms to get a good understanding of its impact.

Pre-trends

Given the empirical setting of this exercise we expect the population growth effects to take place after 1994, but not before. A natural test is to see if indeed population growth is independent of the distance to homelands before 1994. Our dataset does not cover any year other than 1991 in the period pre-1994, and so we can't use it for this purpose. Instead we use the Dysturb dataset (Giraut and Vacchiani-Marcuzzo, 2013), a dataset that maps population in comparable units over time in South Africa. Dysturb provides data at two different levels of aggregation, 'urban agglomeration' (UA) and 'magisterial district' (MD). We use the UA dataset because unlike the MD the units used here are defined consistently over time. In Table 3.4.2 we regress population growth on the distance to homelands variable. In columns 4, 5 and 6 we control for initial population, in the first three columns we do not. In columns 1 and 4 we use all units with non-missing data in 1980 and 1991. In columns 2 and 5 we use all units with non-missing data in 1991 and 2001. In columns 3 and 6 we use all units with non-missing data in 1980, 1991 and 2001 to make sure the difference in coefficients between columns is not driven by sample selection. The table shows that we find the expected negative correlation between distance to the homeland and population growth for 1991-2001, but not for 1980-1991.

3.5 Results

Baseline results

Table 3.5.1 summarizes the main results and Table 3.5.2 provides further results from different sub-samples as robustness checks. Each cell of the tables summarizes one regression.[17] The Angrist and Pischke (2009) F-statistic of the first stage is well above the rule of thumb threshold of 10 for all specifications for the

[17] I.e. each cell in the first row of Table 3.5.1 summarizes the causal partial effect of exogenous migration of black population between 1991 and 1996 on the overall population growth rate between 1991 and 1996.

Table 3.4.2: Pre-trend regressions

	(1) $\Delta pop_{80,91}$	(2) $\Delta pop_{91,01}$	(3) $\Delta pop_{91,01}$	(4) $\Delta pop_{80,91}$	(5) $\Delta pop_{91,01}$	(6) $\Delta pop_{91,01}$
$log(dist)$	-0.002 (0.0095)	-0.017*** (0.0066)	-0.018** (0.0072)	-0.001 (0.0096)	-0.018*** (0.0066)	-0.020*** (0.0071)
$log(pop_{80})$				0.009 (0.0097)		
$log(pop_{91})$					-0.020*** (0.0072)	-0.023*** (0.0080)
Observations	160	207	158	160	207	158

Notes. Sample varies according to data availability for different periods. The sample in columns 3 and 6 consists of observations with data for both periods. Robust standard errors in parentheses.Coefficients that are significantly different from *zero* at the 90% level of confidence are marked with a *; at the 95% level, a **; and at the 99% level, a ***.

Source: Author's analysis based on South African census data.

medium and long horizon.[18] Weak instrument problems only arise for the short period between 1991-1996 and we will not discuss these parameter estimates. The increased explanatory power over the longer time horizons is consistent with the fact that migration decisions only adjust intermittently to a change in policy, such as the end of apartheid.

In the OLS regression (Table 3.5.1, Column 1), we cannot reject the null hypothesis that the coefficient is different from one in the short-run (1991-1996). For the two subsequent periods on the other hand, the coefficient estimates are well below one. This suggests that black population growth occurred in areas with low population growth of the incumbent population and vice versa, since if there was no reaction by the incumbent population an increase in population by one would lead to a coefficient of one mechanically. These results should not be assigned a causal interpretation however, since the result could be driven by unobserved shocks that induce black in-migration and white out-migration or vice versa jointly. The baseline results from the two-stage least squares estimation suggest that the coefficient is not different from one at any horizon such that there is no causal

[18] The corresponding first stage regressions for Tables 3.5.1 and 3.5.2 as well as the other main tables shown here are reported in appendix 3.A.

Table 3.5.1: OLS AND 2SLS BASELINE REGRESSIONS

	(1) OLS	(2) 2SLS
	Population growth	Population growth
Panel A: Population growth rates (1991-1996)		
ΔBlack Population	1.126	0.360
	(0.0814)	(0.862)
FS AP F-Stat	-	2.52
Panel B: Population growth rates (1991-2001)		
ΔBlack Population	0.899^{*}	1.061
	(0.0393)	(0.115)
FS AP F-Stat	-	29.98
Panel C: Population growth rates (1991-2011)		
ΔBlack Population	0.895^{***}	0.993
	(0.0236)	(0.0873)
FS AP F-Stat	-	42.95
Province fixed effects	Yes	Yes
Controls	Yes	Yes
Observations	2093	2093

Notes. This Table displays estimates of equation (3.2) in the text. Each cell presents estimates from a separate regression. The baseline sample consists of all wards inside South Africa for which 1991 data is available. The standard errors are clustered on the municipality level. There are 201 clusters. The outcome variable is absolute overall population growth in the relevant time period divided by overall population. The relevant time periods are 1991-1996 in Panel A, 1991-2001 in Panel B and 1991-2011 in Panel C. Controls include variables on education, income, population group, population density and employment in 1991. There are nine provinces for which fixed effects are included. The estimated coefficients for the first stage regressions are reported in appendix 3.A. Coefficients that are significantly different from *one* at the 90% level of confidence are marked with a *; at the 95% level, a **; and at the 99% level, a ***.

Source: Author's analysis based on South African census data.

effect from exogenous black migration on aggregate migration decisions of non-black incumbents. This is evidence that an exogenous population shock is absorbed without an endogenous reaction of the population level. The results suggest that the effect of an exogenous population shock on the aggregate long-run equilibrium of the population distribution is consistent with the theoretical notion of path-dependence (Hypothesis 2). We note that coefficients estimated for 1991-2001 and 1991-2011 are not statistically different from one another, which could suggest that the migration transition period had converged to a new steady state not long after 2001.

Table 3.5.2: 2SLS regressions using different sub-samples

	(1) Dummy for Johannesburg	(2) No dummy for Cape Town	(3) Dummies for all metro areas	(4) Drop within 10 km	(5) Drop < 5% white	(6) Drop < 10% white	(7) Drop distance ≥ 6	(8) District FE	(9) Municipality level
Panel A: Population growth rates (1991-1996)									
ΔBlack Population	0.360	0.373	0.315	0.615	23.37	20.75	0.345	-0.652	0.814***
	(0.861)	(0.850)	(0.909)	(0.442)	(32.85)	(25.53)	(0.853)	(1.926)	(0.058)
FS AP F-Stat	2.52	2.53	2.32	1.36	0.25	0.31	2.61	1.67	10.27
Panel B: Population growth rates (1991-2001)									
ΔBlack Population	1.069	1.072	1.034	1.066	1.169	1.231	1.008	0.948	1.039
	(0.112)	(0.132)	(0.134)	(0.189)	(0.144)	(0.173)	(0.108)	(0.084)	(0.154)
FS AP F-Stat	32.23	22.96	26.65	12.51	30.67	25.55	30.55	34.68	11.17
Panel C: Population growth rates (1991-2011)									
ΔBlack Population	1.000	1.007	0.957	1.046	1.049	1.062	0.9726	0.870	0.926
	(0.086)	(0.0992)	(0.115)	(0.131)	(0.139)	(0.146)	(0.088)	(0.077)	(0.151)
FS AP F-Stat	44.67	32.52	31.28	20.69	31.09	28.89	42.81	36.59	12.91
District fixed effects	No	No	No	No	No	No	No	Yes	No
Province fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	2093	2093	2093	1790	1374	1137	1730	2093	203

Notes. This Table displays estimates of equation (3.2) in the text for different sub-samples. Column headings denote sub-sample used in each specification. Each cell presents estimates from a separate regression. The standard errors are clustered on the municipality level. There are 201 clusters. All columns are estimated using 2SLS where the natural log of distance to the nearest homeland is used to instrument for absolute black population growth in the relevant time period divided by the overall population. The outcome variable is absolute population growth in the relevant time period divided by the overall population. The relevant time periods are 1991-1996 in Panel A, 1991-2001 in Panel B and 1991-2011 in Panel C. Controls include variables on education, income, population group, population density and employment in 1991. There are nine provinces for which fixed effects are included. The estimated coefficients for the first stage regressions are reported in appendix 3.A. Coefficients that are significantly different from *one* at the 90% level of confidence are marked with a *; at the 95% level, a **; and at the 99% level, a ***. 95% confidence intervals are in brackets.
Source: Author's analysis based on South African census data.

In addition to the baseline regressions, we report several regressions based on different sub-samples as robustness checks (Table 3.5.2). We include a dummy for Johannesburg in Column 1 as the largest metropolitan area and industrial center to ensure that it is not driving the results. In Column 2 we remove the dummy variable for Cape Town that we usually include, which does not significantly affect results. In Column 3 we include separate fixed effects for all the metropolitan areas in our sample, which again does not seem to change our results significantly. As outlined above, we exclude areas close to the homelands, since for these localities, distance to the nearest homeland could affect them not only through migration, but also through local economic spillovers (Column 4). We also exclude areas with a low white population share in 1991 because the migration restrictions under apartheid might have been less binding for these areas (Columns 5 and 6). As a further robustness check, we exclude the areas in the upper tail of the distance distribution in Column 7 to ensure that the high number of observations in the upper tail of the distance distribution does not skew the results. In Column 8 we report results using district instead of province fixed effects. We also aggregate wards up to the municipality level and run a separate regression to test whether the results are robust to using a different level of aggregation (Column 9). These robustness tests using different sub-samples as reported in Table 3.5.2 corroborate our baseline findings since none of the coefficients significantly deviates from one. When comparing the timing of the effect, both the coefficients and statistical power seem fairly similar for the periods 1991-2001 and 1991-2011. The short run result for 1991-1996 is weaker, both in the first stage statistical power and in the magnitude of the second stage result. This might suggest that the migration took longer than the first year after apartheid to converge, while the new equilibrium was largely reached by 2001, and so did not change to 2011.

One concern is that fertility or mortality differences may influence these results. To investigate these concerns we repeat the entire exercise from in Table 3.5.2 for people of working age population only. These results are in Table 3.5.3. Here we define working age as the population that is aged between 15 and 64. All coefficients are similar to their counterpart in Table 3.5.2.

Table 3.5.3: 2SLS regressions using working-age population only

	(1) Baseline	(2) Dummy for Johannesburg	(3) No Cape Town dummy	(4) Dummies for metro areas	(5) Drop within 10 km	(6) Drop < 5% white	(7) Drop < 10% white	(8) Drop dist. ≥ 6	(9) District FE	(10) Municipality level
Panel A: Population growth rates (1991-1996)										
ΔBlack Population	1.311	1.339	1.305	1.175	1.480	1.433	1.694	1.272	1.026	1.128
	(0.206)	(0.212)	(0.227)	(0.242)	(0.383)	(0.315)	(0.479)	(0.206)	(0.182)	(0.361)
FS AP F-Stat	17.27	17.76	14.07	14.45	7.06	14.05	8.46	15.91	18.01	4.46
Panel B: Population growth rates (1991-2001)										
ΔBlack Population	1.070	1.076	1.085	1.048	1.060	1.189	1.254	1.020	0.945	1.032
	(0.120)	(0.118)	(0.139)	(0.143)	(0.196)	(0.157)	(0.185)	(0.116)	(0.0862)	(0.187)
FS AP F-Stat	30.50	32.49	22.85	27.03	11.84	28.83	24.73	30.84	34.50	9.32
Panel C: Population growth rates (1991-2011)										
ΔBlack Population	0.990	0.996	1.006	0.947	1.021	1.055	1.067	0.931	0.876	0.918
	(0.0803)	(0.0787)	(0.0919)	(0.107)	(0.125)	(0.134)	(0.140)	(0.0805)	(0.0788)	(0.175)
FS AP F-Stat	42.36	43.38	31.55	31.37	19.32	31.24	28.54	41.98	36.75	11.53
District fixed effects	No	No	No	No	No	No	No	No	Yes	No
Province fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	2093	2093	2093	2093	1790	1374	1137	1730	2093	203

Notes. This Table displays estimates of equation (3.2) in the text using working-age population rather than overall population for different sub-samples. This includes everyone aged 15 to 64. Column headings denote sub-sample used in each specification. Each cell presents estimates from a separate regression. The standard errors are clustered on the municipality level. There are 201 clusters. All columns are estimated using 2SLS where the natural log of distance to the nearest homeland is used to instrument for absolute black population growth in the relevant time period divided by the overall population. The outcome variable is absolute population growth in the relevant time period divided by the overall population. The relevant time periods are 1991-1996 in Panel A, 1991-2001 in Panel B and 1991-2011 in Panel C. Controls include variables on education, income, population group, population density and employment in 1991. There are nine provinces for which fixed effects are included. The estimated coefficients for the first stage regressions are reported in appendix 3.A. Coefficients that are significantly different from *one* at the 90% level of confidence are marked with a *; at the 95% level, a **; and at the 99% level, a ***. 95% confidence intervals are in brackets.
Source: Author's analysis based on South African census data.

In our main regressions we measure population growth on the right hand side in the same time period as on the left hand side. This may be measured with noise if the incumbent population only reacts to the population shock with a lag as opposed to instantaneously. In order to make sure that these potential dynamics do not distort our results, we test for them in an alternative specification. Table 3.5.4 reports the result of a specification where we run the first stage for black population growth for the period 1991 to 2001 and the second stage for overall population growth for the period 2001 to 2011. Intuitively, this regression picks up whether an exogenous population shock during the period 1991 to 2001 affects overall population growth in the subsequent period. In this specification a coefficient equal to 0 indicates path dependence while a coefficient smaller or larger than 0 indicates mean reversion or multiple equilibria. The fact that none of the coefficients in Table 3.5.4 is significantly different from 0 indicates that the dynamic response is also consistent with path dependence.

Table 3.5.4: Alternative specification using different time periods

	Full sample		Working-age population	
	(1) Δ Overall population	(2) Δ Non-black population	(3) Δ Overall population	(4) Δ Non-black population
ΔBlack Population (1991 - 2001)	0.167 (0.139)	0.0319 (0.484)	0.246 (0.153)	-0.035 (0.045)
FS AP F-Stat	29.97	29.97	30.50	30.50
Province fixed effects	Yes	Yes	Yes	Yes
Controls	Yes	Yes	Yes	Yes
Observations	2093	2093	2093	2093

Notes. This table reports 2SLS results. In the second stage we regress different population growth measures for the period 2001 - 2011 on predicted black population growth in the previous period (1991 - 2001). In the first stage we instrument black population growth using distance to the nearest homeland as in the baseline specification. Controls include variables on education, income, population group, population density and employment in 1991. There are nine provinces for which fixed effects are included. Coefficients that are significantly different from zero at the 90% level of confidence are marked with a *; at the 95% level, a **; and at the 99% level, a ***. 95% confidence intervals are in brackets.
Source: Author's analysis based on South African census data.

Overall, the empirical results provide strong evidence for path dependence (Hypothesis 2). These results suggest that, in the aggregate, there is no evidence for multiple equilibria and a non-monotonic agglomeration curve or mean-reverting behavior. The evidence in favor of path-dependence is consistent with an agglomeration function that has the same slope as the congestion function or high costs of migration as found by Imbert and Papp (2018) for temporary labor migration in India. This corroborates the dynamics found by Bleakley and Lin (2012) for fall line cities in the US and Michaels and Rauch (2018) for Roman cities in France and Britain.

Heterogeneity

This Section discusses how the causal effect of migration on population growth varies across two dimensions, the initial population density or level of urbanization of an area, and the size of the exogenous population shock.

First, we are interested in how the causal effect varies with initial population densities. In this case, theory does not provide clear guidance. Due to the convexity of the cost curve, the causal effect of migration could be decreasing in initial population density because the additional costs generated by new migrants reduce the utility level of incumbents. At the same time, Michaels et al. (2012) show that long-run population growth in the US is smaller for low initial population densities and increases with population density after a cut-off of 7 people per km^2. Such a result would be consistent with an agglomeration curve that is much steeper in urban than in rural areas. This could suggest that in densely populated areas, exogenous migration leads to a larger increase in population than in less densely populated areas.

In order to estimate how the effect of a positive population shock varies across initial population densities, we define dummy variables for high initial population densities and for high initial shares of urbanized households. The results reported in Table 3.5.5 show that there is a positive and significant interaction between high

Table 3.5.5: Heterogeneity with respect to the initial population density and level of urbanization

	(1) High population density dummy	(2) High urban share dummy
Panel A: Population growth rates (1991-1996)		
ΔBlack Population	1.026	-6.333
	(1.425)	(11.35)
High initial urban share dummy ×		7.589
ΔBlack Population		(11.56)
High initial population density dummy ×	0.509	
ΔBlack Population	(1.133)	
FS AP F-Stat: ΔBlack Population	0.81	0.07
FS AP F-Stat: Urban interaction	-	0.07
FS AP F-Stat: Density interaction	0.45	-
Panel B: Population growth rates (1991-2001)		
ΔBlack Population	1.106	1.002
	(0.138)	(0.111)
High initial urban share dummy ×		0.178**
ΔBlack Population		(0.084)
High initial population density dummy ×	0.681**	
ΔBlack Population	(0.284)	
FS AP F-Stat: ΔBlack Population	26.54	36.98
FS AP F-Stat: Urban interaction	-	23.70
FS AP F-Stat: Density interaction	18.71	-
Panel C: Population growth rates (1991-2011)		
ΔBlack Population	0.957	0.986
	(0.092)	(0.082)
High initial urban share dummy ×		0.040
ΔBlack Population		(0.062)
High initial population density dummy ×	0.348**	
ΔBlack Population	(0.152)	
FS AP F-Stat: ΔBlack Population	44.53	46.50
FS AP F-Stat: Urban interaction	-	27.49
FS AP F-Stat: Density interaction	17.67	-
Province fixed effects	Yes	Yes
Controls	Yes	Yes
Observations	2093	2093

Notes. This Table displays estimates of equation (3.2) in the text with an additional interaction term. Each column displays one specification. The standard errors are clustered on the municipality level. There are 201 clusters. All columns are estimated using 2SLS. Absolute black population growth divided by the overall population and the same term interacted with a dummy for high population density in 1991 or for high urban share of households are the endogenous variables. Log distance to the nearest homeland and log distance to the nearest homeland times a dummy for high population density in 1991 or high urban share of households are used as instruments for the endogenous variables. An area is defined as having a high initial population density if it is among the 25% most dense areas. An area is defined as having a high urban share if it is among the areas with the 75% highest share of urban households in 1991. The outcome variable is absolute population growth in the relevant time period divided by the overall population. The relevant time periods are 1991-1996 in Panel A, 1991-2001 in Panel B and 1991-2011 in Panel C. Controls include variables on education, income, population group, population density and employment in 1991. There are nine provinces for which fixed effects are included. The estimated coefficients for the first stage regressions are reported in appendix 3.A. Coefficients on the interaction terms that are significantly different from zero at the 90% level of confidence are marked with a *; at the 95% level, a **; and at the 99% level, a ***.
Source: Author's analysis based on South African census data.

initial densities and population shocks. This suggests that areas with high initial densities experience a significant endogenous inflow of population as a reaction to the exogenous population shock while others do not. This effect exists in the medium and long-run but is economically stronger in the medium run for population density. It looses significance in the long-run for the share of urbanized households. The results of both specifications indicate that the population dynamics induced by a positive population shock differ between less densely populated rural areas and highly populated urban areas. While we cannot reject path dependence for rural areas, there is a significant and positive effect in urban areas suggesting that an exogenous population shock leads to endogenous immigration. This is suggestive evidence for the existence of multiple equilibria within urban areas.

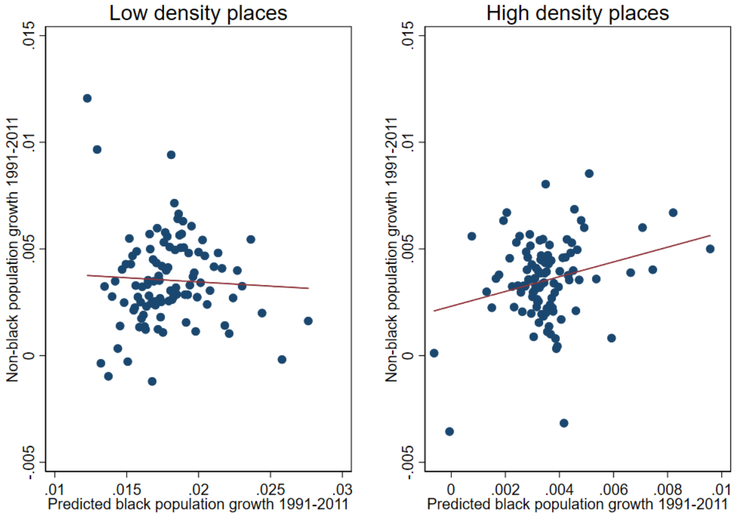
We show the effects graphically in Figure 3.5.1, which corresponds closely to Panel C in Table 3.5.5. We aggregate observations into bins to increase clarity, we specify 100 bins in total for each of the two plots. While low density places show no reaction to the migration shock, and so follow path-dependence, in high density places we see evidence of agglomeration economies: Inflows of people lead to the rest of the population to positively react.

We next consider the size of the population shock as an additional dimension of heterogeneity. So we estimate how overall population growth varies with the size of the shock and initial population density. In order to do so, we combine the deciles of the two distributions and estimate 100 distinct conditional means:

$$\begin{aligned}\Delta N_{i,2011} = &\sum_{j=1}^{10}\sum_{k=1}^{10}\beta_{j,k}\Big[1[\text{if } \mathit{Popden}_{i,1991} \text{ in decile } j] \times 1[\text{if } \mathit{distance}_i \text{ in decile } k]\Big] \\ &+ \gamma' X_{i,1991} + \delta_p + \epsilon_m \qquad (3.3)\end{aligned}$$

The $\beta_{j,k}$s are the coefficients of interest and estimate how conditional population growth varies by the deciles of the initial population density distribution and the size of the shock distribution. The size of the shock is measured using the reduced

Figure 3.5.1: Population reaction for high and low density places



Note: This figure displays the relationship between predicted population growth and its impact for the period 1991-2011. Definitions correspond to those in Table 3.5.5.
Source: Author's analysis based on South African census data.

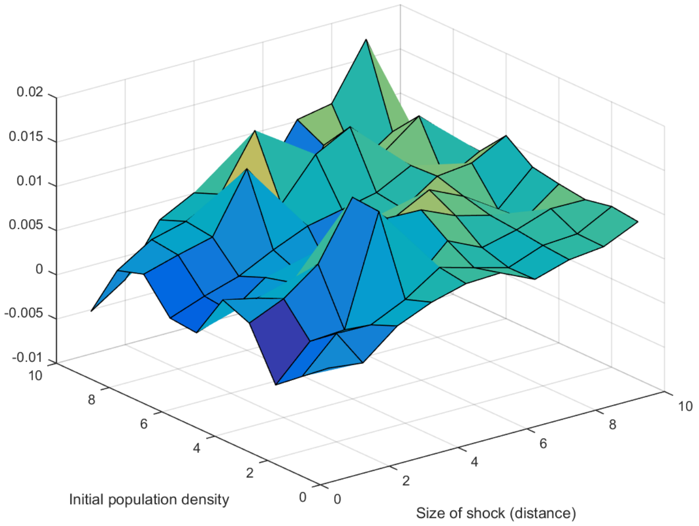
form, i.e. distance to the nearest homeland. While the estimates do not provide the same clear cut causal evidence as the two stage least squares approach they are indicative as to how the effect of the exogenous population shock varies with the size of the shock and the initial density. The results displayed in Figure 3.5.2 suggest that for a given initial density an increase in the size of a shock results in a higher population growth rate. This is in line with the idea that it requires a substantial shock to switch between equilibria.

The fact that the population of more densely populated areas increases relative to less densely populated areas could be interpreted as a 'Matthew effect'[19] of an exogenous population shock, where areas rich in population gain over-proportionally from a positive population shock.

In the context of the modified Henderson model presented in Section 3.3, this result suggests that the shape of the agglomeration function is different between urban and rural areas for the relevant population levels. In rural areas, the gains from

[19] 'For unto every one that hath shall be given, and he shall have abundance: but from him that hath not shall be taken even that which he hath.' Matthew 25:29, (American Bible Society, 1999).

Figure 3.5.2: The effect of initial density and size of shock for population growth



Note: This figure displays the $\beta_{j,k}$ coefficients resulting from estimating equation (3.3) in the main text. The size of the shock increases along the x-axis. It starts off with the highest decile of the distance distribution going to the decile with the lowest values (i.e. those closest to a homeland). Similarly, the first value on the y-axis corresponds to those wards in the lowest decile of the initial population density distribution while the last one contains the highest decile. The z-axis displays differences in the conditional mean of population growth in the period 1991-2011.
Source: Author's analysis based on South African census data.

agglomeration are below the increased congestion cost if the population increases exogenously. In urban areas, the gains from an increase in population seem to be equal to the additional costs. The gains from agglomeration therefore seem to be much larger in urban areas than in less densely populated rural areas.

While the simple model easily accommodates this heterogeneity in the agglomeration and congestion cost curves across different initial population densities, it remains silent on its origin. There a number of ways this heterogeneity can be microfounded. Such an agglomeration function emerges naturally in a two-sector economic geography model or labor market models that distinguish high- and low-skilled workers with different production technologies in urban and rural labor markets. Consider a simple economic geography model where the agricultural sector produces food using a fixed endowment of land and labor under a technology with decreasing returns to labor. The industrial sector, consisting of manufacturing and services, produces consumption goods using capital and labor with external agglomeration economies. Labor is perfectly mobile across sectors and locations. Areas with low population densities are predominantly agricultural, while urban areas are predominantly industrial. If an exogenous population shock hits both urban and rural areas, the marginal product of labor decreases in rural areas and generates displacement effects because the real wage decreases. This dynamic arises naturally from the assumption that there is only a fixed amount of land available for agricultural production. In urban areas, an increase in the labor force generates higher investment in capital (assuming a constant real interest rate set in world markets). Therefore, the marginal product of labor does not fall and might even increase due to external economies of scale. This generates agglomeration effects or a path-dependent evolution of population in urban areas.

A similar result emerges in a standard model used in the migration literature (e.g. Borjas (1999) and Kremer and Watt (2006)) that distinguishes between low- and high-skilled labor used in production in urban areas. The production in rural areas only uses low-skilled labor and the fixed amount of land as inputs with the same technology as above. In urban areas, low- and high-skilled labor are used as complements in production with a constant returns to scale technology. In this framework, the population shock we analyze in the data is best approximated by an increase of unskilled labor, since the apartheid government only provided a bare minimum of schooling to the black population (Feinstein, 2005, p.159f). In the model, an increase in unskilled labor increases the wage for high-skilled labor and the rents for capital. If the supply of capital is elastic, this leads to an increase in capital and an inflow of skilled workers such that all factor prices return to their initial equilibrium values. Therefore, an exogenous increase in the number of unskilled workers attracts skilled workers such that the population level of urban areas experiences agglomeration and a shift towards a new equilibrium.

3.6 Conclusion

We study the effect of an exogenous migration shock generated by the abolition of migration restrictions for the black population on the distribution of population in South Africa. There are three ways in which an area can react to an exogenous population shock that arise from different theories describing the distribution of population in space. The population level of an area could mean revert towards its initial level, it could remain at the new population level (path dependence) or it could grow further, i.e. agglomerating population, suggesting the existence of multiple equilibria. The empirical results presented in this paper suggest that in the aggregate, the reaction of the population level to an exogenous population shock is consistent with path dependence. This potentially has important policy implications. If the population level of a region is path dependent, a temporary policy measure that induces migration can permanently change the distribution of population.

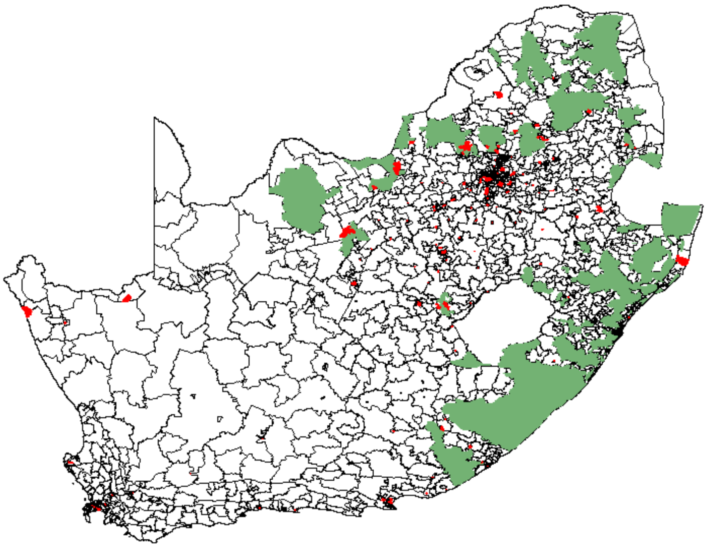
Additionally, we find that the reaction of an area to an exogenous population shock varies with the initial population density. In rural areas with low initial population densities, the effect of an exogenous population shock is significantly smaller than in urban areas with high population densities. In urban areas, the dynamics of the population level are consistent with agglomeration. We provide evidence that for a given initial population density a larger exogenous population shock leads to more endogenous immigration. In the context of the modified Henderson model, this result shows that the agglomeration curve in rural areas is much more concave than in urban areas and it also suggests that it's slope is non-monotonic. These results are consistent with a simple economic geography model in which production in rural areas features decreasing returns to labor due to a fixed endowment of land usable for agricultural purposes. A steeper agglomeration function in urban areas also emerges in a standard model from the migration literature that features complementarities between low- and high-skilled labor in urban, but not in rural areas. If an exogenous population shock hits both rural and urban areas, these differing dynamics increase the share of the population living in cities.

Appendix

3.A Additional specifications

This appendix contains all the first stage regressions corresponding to the tables in the chapter, as well as one graph discussed briefly in the main text.

Figure 3.A.1: Missing wards



Note: This map displays all wards in South Africa outside of the former homelands. Those wards with in red/black are missing. The former homelands are colored in green/grey.

Source: Authors' own work using data from the Directorate: Public Sate Land Support via Africa Open Data.

Table 3.A.1: Summary of first stage regressions for the baseline specifications with different fixed effects

	(1) No FE	(2) Province FE	(3) District FE	(4) Municipality FE
Panel A: Population growth rates (1991-1996)				
log distance	0.308	-0.940	-0.936	-1.148
	(0.365)	(0.593)	(0.723)	(0.993)
Panel B: Population growth rates (1991-2001)				
log distance	-0.007***	-0.007***	-0.006***	-0.003**
	(0.001)	(0.001)	(0.001)	(0.001)
Panel C: Population growth rates (1991-2011)				
log distance	-0.004***	-0.004***	-0.004***	-0.002**
	(0.000)	(0.001)	(0.001)	(0.001)
Level of Fixed effects	No	Province	District	Municipality
Controls	Yes	Yes	Yes	Yes
Observations	2093	2093	2093	2093

Note: This Table displays estimates of equation (3.1) in the main text. Column headings denote different specification. Each cell presents estimates from a separate regression. The standard errors are clustered on the municipality level. There are 201 clusters. All columns are estimated using OLS where the natural log of distance to the nearest homeland is the variable of interest. The outcome variable is absolute black population growth in the relevant time period divided by the overall population. The relevant time periods are 1991-1996 in Panel A, 1991-2001 in Panel B and 1991-2011 in Panel C. Controls include variables on education, income, population group, population density and employment in 1991. Fixed effects at varying levels are included. Coefficients that are statistically significant at the 90% level of confidence are marked with a *; at the 95% level, a **; and at the 99% level, a ***. Standard errors in parentheses. *Source*: Author's analysis based on South African census data.

Table 3.A.2: First stage regressions corresponding to Table 3.5.1

	(1) Black population growth
Panel A: Population growth rates (1991-1996)	
log distance	-0.940
	(0.593)
Panel B: Population growth rates (1991-2001)	
log distance	-0.007***
	(0.001)
Panel C: Population growth rates (1991-2011)	
log distance	-0.004***
	(0.001)
Fixed effects	Yes
Controls	Yes
Observations	2093

Notes. This Table displays estimates of equation (3.1) in the main text. Each cell presents estimates from a separate regression. The standard errors are clustered on the municipality level. There are 201 clusters. All columns are estimated using OLS where the natural log of distance to the nearest homeland is the variable of interest. The outcome variable is absolute black population growth in the relevant time period divided by the overall population. The relevant time periods are 1991-1996 in Panel A, 1991-2001 in Panel B and 1991-2011 in Panel C. Controls include variables on education, income, population group, population density and employment in 1991. There are nine provinces for which fixed effects are included. Coefficients that are statistically significant at the 90% level of confidence are marked with a *; at the 95% level, a **; and at the 99% level, a ***. Standard errors in parentheses.
Source: Author's analysis based on South African census data.

Table 3.A.3: Summary of first stage regressions corresponding to Table 3.5.2

	(1) Dummy for Johannesburg	(2) No dummy for Cape Town	(3) Dummies for all metro areas	(4) Drop within 10 km	(5) Drop $< 5\%$ white	(6) Drop $< 10\%$ white	(7) Drop distance ≥ 6	(8) District FE	(9) Municipality level
Panel A: Population growth rates (1991-1996)									
log distance	-0.949	-0.951	-1.109	-0.806	0.0761	0.106	-1.004	-0.945	-0.948
	(0.592)	(0.591)	(0.722)	(0.687)	(0.152)	(0.189)	(0.616)	(0.719)	(0.700)
Panel B: Population growth rates (1991-2001)									
log distance	-0.007***	-0.006***	-0.006***	-0.006***	-0.008***	-0.007***	-0.007***	-0.007***	-0.006***
	(0.001)	(0.001)	(0.001)	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Panel C: Population growth rates (1991-2011)									
log distance	-0.004***	-0.003***	-0.003***	-0.004***	-0.003***	-0.004***	-0.003***	-0.004***	-0.003***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
District fixed effects	No	No	No	No	No	No	No	Yes	No
Province fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	2093	2093	2093	1790	1374	1137	1730	2093	203

Notes. This Table displays estimates of equation (3.1) in the main text. Column headings denote different specification. Each cell presents estimates from a separate regression. The standard errors are clustered on the municipality level. There are 201 clusters. All columns are estimated using OLS where the natural log of distance to the nearest homeland is the variable of interest. The outcome variable is absolute black population growth in the relevant time period divided by the overall population. The relevant time periods are 1991-1996 in Panel A, 1991-2001 in Panel B and 1991-2011 in Panel C. Controls include variables on education, income, population group, population density and employment in 1991. There are nine provinces for which fixed effects are included. The estimated coefficients for the first stage regressions are reported in the appendix. Coefficients that are statistically significant at the 90% level of confidence are marked with a *; at the 95% level, a **; and at the 99% level, a ***. Standard errors in parentheses.
Source: Author's analysis based on South African census data.

Table 3.A.4: Summary of first stage regressions corresponding to Table 3.5.3

	(1) Baseline	(2) Dummy for Johannesburg	(3) No dummy for Cape Town	(4) Dummies for all metro areas	(5) Drop within 10 km	(6) Drop < 5% white	(7) Drop < 10% white	(8) Drop distance ≥ 6	(9) District FE	(10) Municipality level
Panel A: Population growth rates (1991-1996)										
log distance	-0.003***	-0.003***	-0.003***	-0.003***	-0.003***	-0.003***	-0.003***	-0.003***	-0.003***	-0.004*
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.002)
Panel B: Population growth rates (1991-2001)										
log distance	-0.007***	-0.007***	-0.006***	-0.006***	-0.006***	-0.007***	-0.007***	-0.007***	-0.006***	-0.008**
	(0.001)	(0.001)	(0.001)	(0.001)	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)	(0.003)
Panel C: Population growth rates (1991-2011)										
log distance	-0.004***	-0.004***	-0.003***	-0.003***	-0.004***	-0.003***	-0.003***	-0.003***	-0.004***	-0.004***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
District fixed effects	No	No	No	No	No	No	No	No	Yes	No
Province fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	2093	2093	2093	2093	1790	1374	1137	1730	2093	203

Notes. This Table displays estimates of equation (3.1) in the main text using working age population. Column headings denote different specification. Each cell presents estimates from a separate regression. The standard errors are clustered on the municipality level. There are 201 clusters. All columns are estimated using OLS where the natural log of distance to the nearest homeland is the variable of interest. The outcome variable is absolute black population growth in the relevant time period divided by the overall population. The relevant time periods are 1991-1996 in Panel A, 1991-2001 in Panel B and 1991-2011 in Panel C. Controls include variables on education, income, population group, population density and employment in 1991. There are nine provinces for which fixed effects are included. The estimated coefficients for the first stage regressions are reported in the appendix. Coefficients that are statistically significant at the 90% level of confidence are marked with a *; at the 95% level, a **; and at the 99% level, a ***. Standard errors in parentheses.
Source: Author's analysis based on South African census data.

Table 3.A.5: First stage regressions of specification with interaction term corresponding to table 3.5.5

	(1)	(2)	(3)	(4)
	ΔBlack pop growth	ΔBlack pop growth × high population density dummy	ΔBlack pop growth	ΔBlack pop growth × high urban share dummy
Panel A: Population growth rates (1991-1996)				
log distance	-1.250*	-0.821	-1.426*	-1.388
	(0.727)	(0.656)	(0.846)	(0.841)
log distance × high population density dummy	1.426	1.564	-	-
	(1.130)	(1.253)	-	-
log distance × high urban share dummy	-	-	0.710	0.815
	-	-	(0.736)	(0.707)
Panel B: Population growth rates (1991-2001)				
log distance	-0.007***	0.001***	-0.009***	0.004**
	(0.001)	(0.0003)	(0.001)	(0.002)
log distance × high population density dummy	0.002*	-0.004***	-	-
	(0.001)	(0.001)	-	-
log distance × high urban share dummy	-	-	0.003***	-0.009***
	-	-	(0.001)	(0.002)
Panel C: Population growth rates (1991-2011)				
log distance	-0.004***	0.000	-0.004***	0.002*
	(0.001)	(0.000)	(0.001)	(0.001)
log distance × high population density dummy	0.001	-0.002***	-	-
	(0.001)	(0.001)	-	-
log distance × high urban share dummy	-	-	0.001**	-0.005***
	-	-	(0.001)	(0.001)
Province fixed effects	Yes	Yes	Yes	Yes
Controls	Yes	Yes	Yes	Yes
Observations	2093	2093	2093	2093

Notes. This Table displays estimates of equation (3.1) in the main text with an additional interaction term. Column headings denote different specifications. The standard errors are clustered on the municipality level. There are 201 clusters. All columns are estimated using OLS where the natural log of distance to the nearest homeland and the same term times a dummy for high initial population density or high initial urban share of households are the variables of interest. The outcome variable are absolute black population growth divided by the overall population times a dummy for high initial population density or high initial share of urban households in 1991 and absolute black population growth divided by the overall population. The relevant time periods are 1991-1996 in Panel A, 1991-2001 in Panel B and 1991-2011 in Panel C. Controls include variables on education, income, population group, population density and employment in 1991. There are nine provinces for which fixed effects are included. The estimated coefficients for the first stage regressions are reported in the appendix. Coefficients that are statistically significant at the 90% level of confidence are marked with a *; at the 95% level, a **; and at the 99% level, a ***. Standard errors in parentheses.

Source: Author's analysis based on South African census data.

4

Trading Opportunities and Settlement Formation in the Iron Age[1]

[1] A different version of this chapter has been published as "Of Mice and Merchants: Trade and Growth in the Iron Age", NBER Working Paper 24825, with Stephan Maurer, Jörn-Steffen Pischke and Ferdinand Rauch, and is currently prepared for resubmission at the *Review of Economic and Statistics*

Abstract

We study the causal connection between trade and development using one of the earliest massive trade expansions: the first systematic crossing of open seas in the Mediterranean during the time of the Phoenicians. We construct a measure of connectedness along the shores of the sea. We relate connectedness to local growth, which we measure using the presence of archaeological sites. We find an association between better connected locations and archaeological sites during the Iron Age, at a time when sailors began to cross open water routinely on a big scale. We corroborate these findings at the level of the world.

4.1 Introduction

We investigate to what degree trading opportunities affected economic development at an early juncture of human history. In addition to factor accumulation and technical change, Smithian growth due to exchange and specialization is one of the fundamental sources of growth. An emerging literature on the topic is beginning to provide compelling empirical evidence for a causal link from trade to growth. We contribute to this literature and focus on one of the earliest massive expansions in maritime trade: the systematic crossing of open seas in the Mediterranean at the time of the Phoenicians from about 900 BC. We relate trading opportunities, which we capture through the connectedness of points along the coast, to early development as measured by the presence of archaeological sites. We find that locational advantages for sea trade matter for the presence of Iron Age cities and settlements, and thus helped shape the development of the Mediterranean region, and the world.

A location with more potential trading partners should have an advantage if trade is important for development. The particular shape of a coast has little influence over how many neighboring points can be reached from a starting location within a certain distance as long as ships sail mainly close to the coast. However, once sailors begin to cross open seas, coastal geography becomes more important: Some coastal points are in the reach of many neighbors while others can reach only few. The general shape of the coast and the location of islands matters for this. We capture these geographic differences by dividing the Mediterranean coast into grid cells, and calculating how many other cells can be reached within a certain distance. Parts of the Mediterranean are highly advantaged by their geography, e.g. the island-dotted Aegean and the "waist of the Mediterranean" at southern Italy, Sicily, and modern Tunisia. Other areas are less well connected, like most of the straight North African coast, parts of Iberia and southern France, and the Levantine coast.

We relate our measure of connectivity to the number of archaeological sites found near any particular coastal grid point. This is our proxy for economic development.

It is based on the assumption that more human economic activity leads to more settlements and particularly towns and cities. When these expand and multiply there are more traces in the archaeological record. We find a pronounced relationship between connectivity and development in our dataset for the Iron Age around 750 BC, when the Phoenicians began to systematically traverse the open sea. We have less evidence whether there was any relationship between connectivity and sites for earlier periods when the data on sites are poorer. Connectivity might already have mattered during the Bronze Age when voyages occurred at some frequency, maybe at more intermediate distances. Our interpretation of the results suggests that the relationship between coastal geography and settlement density, once established in the Iron Age, persists through the classical period. This is consistent with a large literature in economic geography on the persistence of city locations. While our main results pertain to the Mediterranean, where we have good information on archaeological sites, we also corroborate our findings at a world scale using population data for 1 AD from McEvedy and Jones (1978) as outcome.

Humans have obtained goods from far away locations for many millennia. While some of the early trade involved materials useful for tools (like the obsidian trade studied by Dixon, Cann, and Renfrew 1968), as soon as societies became more differentiated a large part of this early trade involved luxury goods doubtlessly consumed by the elites. Such trade might have raised the utility of the beneficiaries but it is much less clear whether it affected productivity as well. Although we are unable to measure trade directly, our work sheds some light on this question. Since trade seems to have affected the growth of settlements even at an early juncture this suggests that it was productivity enhancing. The view that trade played an important role in early development has recently been gaining ground among both economic historians and archaeologists; see e.g. Temin (2006) for the Iron Age Mediterranean, Algaze (2009) for Mesopotamia, Barjamovic et al. (2019) for Assyria, and Temin (2017) for Ancient Rome.

Our approach avoids issues of reverse causality and many confounders by using a geography based instrument for trade. In fact, we do not observe trade itself but effectively estimate a reduced form relationship, relating opportunities for trade directly to economic development. This means that we do not necessarily isolate the effect of the exchange of goods per se. Our results could be driven by migration or the spread of ideas as well, and when we talk about "trade" we interpret it in this broad sense. We do believe that coastal connectivity captures effects due to maritime connections. It is difficult to imagine any other channel why geography would matter in this particular manner, and we show that our results are not driven by a variety of other geographic conditions.

Since we do not use any trade data we avoid many of the measurement issues related to trade. We measure trading opportunities and development at a fine geographic scale, hence avoiding issues of aggregation to a coarse country level. Both our measure of connectedness and our outcome variable are doubtlessly crude proxies of both trading opportunities and of economic development. This will likely bias us against finding any relationship and hence makes our results only more remarkable.

The periods we study, the Bronze and Iron Ages, were characterized by the rise and decline of many cultures and local concentrations of economic activity. Many settlements and cities rose during this period, only to often disappear again. This means that there were ample opportunities for new locations to rise to prominence while path dependence and hysteresis may have played a lesser role compared to later ages. The political organization of the Mediterranean world prior to the Romans was mostly local. The Egyptian Kingdoms are the main exception to this rule but Egypt was mostly focused on the Nile and less engaged in the Mediterranean. As a result, institutional factors were less important during the period we study.

There is a large literature on trade and growth. Canonical studies are the investigations by Frankel and Romer (1999) and Redding and Venables (2004). These papers use distance from markets and connectivity as measured by gravity relationships to capture the ease with which potential trading partners can be reached. However, these measures do not rely purely on geography but conflate economic outcomes like population and output, which are themselves affected by the development process.

The more recent literature has circumvented this by analyzing exogenous events related to changes in trade. Most similar to our study are a series of papers which also exploit new trade relationships arising from discoveries, the opening of new trade routes, and technological change. Acemoglu et al. (2005) link Atlantic trade starting around 1,500 AD to the ensuing shift in the focus of economic activity in Europe from the south and center of the continent to the Atlantic periphery. Redding and Sturm (2008) focus on the natural experiment created by the division and reunification in Germany, which changed the access to other markets sharply for some locations but not others. A similar natural experiment is employed by Feyrer (2009a), who uses exogenous variation in sea distance created by the temporary closure of the Suez Canal. Various papers exploit the availability of new transport technologies; Feyrer (2009b) uses air transport, Donaldson (2018) and Donaldson and Hornbeck (2016) use railroads, and Pascali (2017) steam ships. These papers generally find that regions whose trading opportunities improved disproportionately saw larger income growth. That we find similar results for a much earlier trade expansion suggests that the productivity benefits of trade have been pervasive throughout history.

Our paper also relates to a literature on how changes in locational fundamentals shape the location of cities (Davis and Weinstein (2002), Bleakley and Lin (2012), Bosker and Buringh (2017), Hanlon (2017), Michaels and Rauch (2018)). Our contribution to this literature is to give evidence on one of the most important locational fundamentals, market access. In a world with multiple modes of transport for the transportation of different goods, it is typically hard to measure market access and changes of market access of a city. Our measure relates to a world where much long distance trade took place on boats, which makes it easier to isolate a measure of market access.

Also closely related is the paper by Ashraf and Galor (2011). They relate population density in various periods to the relative geographic isolation of a particular area. Their interest is in the impact of cultural diversity on the development process, and they view geographic isolation effectively as an instrument for cultural homogeneity. Similar to our measure, their geographic isolation measure is a measure of connectivity of various points around the world. They find that better connected (i.e. less isolated) countries have lower population densities for every period from 1 to 1,500 AD, which is the opposite of our result. Our approach differs from Ashraf and Galor (2011) in that we only look at locations near the coast and not inland locations. They control for distance to waterways in their regressions, a variable that is strongly positively correlated with population density. Hence, our results are not in conflict with theirs.

Our paper is also related to a number of studies on prehistoric Mediterranean connectivity and seafaring. MacEvedy (1967) creates a measure of “littoral zones” using coastal shapes. He produces a map which closely resembles the one we obtain from our connectivity measure but does not relate geography directly to seafaring. This is done by Broodbank (2006), who overlays the connectivity map with archaeological evidence of the earliest sea-crossings up to the end of the last Ice Age. He interprets the connections as nursery conditions for the early development of nautical skills, rather than as market access, as we do for the later Bronze and Iron Ages.

Also related is a literature in archaeology using network models connecting archaeological sites; Knappett et al. (2008) is an example for the Bronze Age Aegean. Barjamovic et al. (2019) conduct a similar exercise for Assyria based on a gravity model. None of these papers relate to the changes arising from open sea-crossings, which is the focus of our analysis. Temin (2006) discusses the Iron Age Mediterranean through the lens of comparative advantage trade but offers no quantitative evidence as we do.

4.2 Brief history of ancient seafaring in the Mediterranean

The Mediterranean is a unique geographic space. The large inland sea is protected from the open oceans by the Strait of Gibraltar. The tectonics of the area, the African plate descending under the Eurasian one, have created a rugged northern coast in Europe and a much straighter one in North Africa. Volcanic activity and the more than 3,000 islands also tend to be concentrated towards the north. The climatic conditions in the Mediterranean are generally relatively favorable to agriculture, particularly in the north. The Mediterranean is the only large inland sea with such a climate (Broodbank, 2006). Its east-west orientation facilitated the spread of agriculture from the Levant (Jared, 1997). Despite these common features, the size of the Mediterranean and an uneven distribution of natural resources also implies great diversity. Purcell and Horden (1999) stress that the area consists of many micro-regions. Geography and climate make the Mediterranean prone to risks such as forest fires, earthquakes, plagues of locusts, droughts, floods, and landslides. As a consequence, trade networks that allow to moderate shocks are of great mutual interest in the region, and trade has played a central role since its early history.[2]

Clear evidence of the first maritime activity of humans in the Mediterranean is elusive. Crossings to islands close to the mainland were apparently undertaken as far back as 30,000 BC (Fontana Nuova in Sicily). In a careful review of the evidence, Broodbank (2006) dates more active seafaring to around 10,000 BC based on the distribution of obsidian (a volcanic rock) at sites separated by water (see Dixon et al. (1968)). This points to the existence of active sea-faring of hunter-gatherer societies, and suggests that boats must have traveled distances of 20-35 kilometers around that time. We have no evidence on the first boats but they were likely made from skin and frame or dugout canoes.

[2] The following discussion mainly draws on Abulafia (2011) and Broodbank (2013).

Agriculture around the Mediterranean began in the Levant some time between 9,500 BC and 8,000 BC. From there it spread initially to Anatolia and the Aegean. Signs of a fairly uniform Neolithic package of crops and domesticated animals can be found throughout the Mediterranean. The distribution of the earliest evidence of agriculture, which includes islands before reaching more peripheral parts of the mainland, suggests a maritime transmission channel.

The Neolithic revolution did not reach Iberia until around 5,500 BC. By that time, many islands in the Aegean had been settled, there is evidence for grain storage, and metal working began in the Balkans. Because of the uneven distribution of ores, metals soon became part of long range transport. Uncertainty must also have been a reason for the formation of networks. Trade networks facilitated both comparative advantage based exchange and insurance. The first archaeological evidence of a boat also stems from this period: a dugout canoe, about 10 m long, at La Marmotta north of Rome. A replica proved seaworthy and allowed travel of 20 - 25 km per day in a laden boat.

The Levant, which was home to the first cities, remained a technological leader in the region, yet there is little evidence of sea-faring even during the Copper Age. This changed with the rise of large scale societies in Mesopotamia and Egypt. Inequality in these first states led to rich elites, who soon wished to trade with each other. Being at the cross-roads between these two societies, the Levant quickly became a key intermediary.

Two important new transport technologies arrived in the Mediterranean around 3,000 BC: the donkey and the sail. The donkey was uniquely suited to the climatic conditions and rugged terrain around the Mediterranean (better than camels or horses). Donkeys are comparable in speed to canoes. Sailboats of that period could be around 5-10 times faster in favorable conditions, ushering in a cost advantage of water transport that would remain intact for many millennia to come. The land route out of Egypt to the Levant (“The Way of Horus”) was soon superseded by sea routes leading up the Levantine coast to new settlements like Byblos, with Levantine traders facilitating much of Egypt's Mediterranean trade. Coastal communities began to emerge all the way from the Levant via Anatolia to the Aegean and Greece.

There is no evidence of the sail spreading west of Greece at this time. Canoes, though likely improved into high performance water craft, remained inferior to sail boats but kept facilitating maritime transport in the central and western Mediterranean. The major islands there were all settled by the early Bronze Age. While not rivaling the maritime activity in the eastern Mediterranean, regional trade networks arose also in the west. One example is the Beaker network of the 3rd Millennium BC; most intense from southern France to Iberia, with fewer beakers found in the western Maghreb, northern Italy, and Sardinia but also stretching all the way into central Europe, the Baltic, and Britain. Land routes probably dominated but sea trade must have played a role. The Cetina culture of the late 3rd Millennium BC in the Adriatic is another example. Occasional sea-crossings up to 250 km were undertaken during this period.

A drying spell around 2,200 BC and decline in Egypt disrupted the active maritime network in the eastern Mediterranean and the population it supported. The oldest known shipwreck in the Mediterranean at the island of Dokos in southern Greece dates from this period. The 15 meters long boat could carry a maximum weight of 20 tons. The wreck contained largely pottery, which was likely the cargo rather than carrying liquids, and also carried lead ingots. The ship probably was engaged in local trade.

Decline in the eastern Mediterranean soon gave rise to new societies during the 2nd millennium BC: palace cultures sprang up all over the eastern Mediterranean. Minoan Crete and Mycenae in Greece were notable examples but similar cities existed along the Anatolian coast and in the Levant. The palaces did not simply hold political power, but were centers of religious, ceremonial, and economic activity. At least initially, craftsmen and traders most likely worked for the palace rather than as independent agents. Sail boats still constituted an advanced technology, and only the concentration of resources in the hands of a rich elite made their construction and operation possible. The political reach of the palaces at coastal sites was local; larger polities remained confined to inland areas as in the case of Egypt, Babylon, or the Hittite Empire.

An active trade network arose again in the eastern Mediterranean stretching from Egypt to Greece during the Palace period. The Anatolian land route was replaced by sea trade. Some areas began to specialize in cash crops like olives and wine. A typical ship was still the 15 m, 20 ton, one masted vessel as evidenced by the Uluburn wreck found at Kas in Turkey, dating from 1,450 BC. Such vessels carried diverse cargoes including people (migrants, messengers, and slaves), though the main goods were likely metals, textiles, wine, and olive oil. Evidence for some of these was found on the Uluburun wreck; other evidence comes from archives and inscriptions akin to bills of lading. Broodbank (2013) suggests that the value of cargo of the Uluburun ship was such that it was sufficient to feed a city the size of Ugarit for a year. Ugarit was the largest trading city in the Levant at the time with a population of about 6,000 - 8,000. This highlights that sea trade still largely consisted of high value luxury goods. The Ugarit archives also reveal that merchants operating on their own account had become commonplace by the mid 2nd millennium. Levantine rulers relied more on taxation than central planning of economic activities. Trade was both risky and profitable; the most successful traders became among the richest members of their societies.

Around the same time, the Mycenaeans traded as far as Italy. Sicily and the Tyrrhenian got drawn into the network. While 60 - 70 km crossings to Cyprus or Crete and across the Otranto Strait (from Greece to the heel of Italy) were commonplace, coast hugging still prevailed among sailors during the 2nd millennium BC. After crossing the Otranto Strait, Greek sailors would continue along the coast of the Bay of Taranto, the instep of Italy's boot, as is suggested by the distribution of Greek pottery at coastal sites. Indigenous sea-farers from the central

Mediterranean now joined these routes, and the sail finally entered the central Mediterranean around 1,200 BC. While there were no big breakthroughs, naval technology also improved in the late 2nd millennium. Better caulking and keels added to sea-worthiness (Abulafia, 2011), while brail rigging and double prows improved maneuverability. Most notably, latitude sailing was developed and allowed sailors to steer a straight east-westerly course. "This was a leap in the scope of connections, a permanent shift in Mediterranean history and a crucial stage in tying together the basin's inhabitants across the soon-to-be shrinking sea," observes Broodbank (2013, p. 431) before warning that "we should not exaggerate, nor anticipate, the importance of such connections at this early juncture. Not until the Iron Age did relations become close enough to fundamentally reshape the culture and economies of outlying regions." (p. 441)

A new period of decline around 1,200 BC reduced the power of Egypt, wiped out cities like Ugarit, and ended the reign of the last palace societies in the eastern Mediterranean. In the more integrated world that the eastern Mediterranean had become, troubles spread quickly from one site to others. The Bronze Age came to an end with iron coming on the scene. Rather than being technologically all that much superior to bronze, iron ore was far more abundant and widespread than copper and hence much more difficult to monopolize. As was the case many times before, decline and change opened up spaces for smaller players and more peripheral regions. Cyprus flourished. Many Levantine cities recovered quickly. Traders from the central Mediterranean also expanded. Traditionally, decline during the Bronze Age collapse was often blamed on the anonymous "Sea Peoples." Modern scholarship seems to challenge whether these foreigners were simply just raiders and pirates, as the Egyptians surely saw them, rather than also entrepreneurial traders who saw opportunities for themselves to fill the void left by the disappearance of imperial connections and networks. Some of these new interlopers settled in the Levant (Broodbank, 2013).

While there is much academic debate about the origin of the Phoenicians, there is little doubt that the Levantine city states which had taken in these migrants were the origin of a newly emerging trade network. Starting to connect the old Bronze Age triangle formed by the Levantine coast and Cyprus, they began to expand throughout the entire Mediterranean after 900 BC. The Phoenician city states were much more governed by economic logic than was the case for royal Egypt. One aspect of their expansion was the formation of enclaves, often at nodes of the network. Carthage and Gadir (Cadiz) are prime examples but many others existed. At least initially these were not colonies; the Phoenicians did not try to dominate local populations. Instead, locals and other settlers were invited to pursue their own enterprise and contribute to the trading network. The core of the network consisted of the traditional sea-faring regions, the Aegean and the Tyrrhenian. The expanding trade network of the early 1st millennium BC did not start from scratch but encompassed various regional populations. Tyrrhenian metal workers and Sardinian sailors had opened up connections with Iberia at the close of the 2nd millennium. But the newly expanding network not only stitched these routes together, it also created its own, new, long-haul routes.

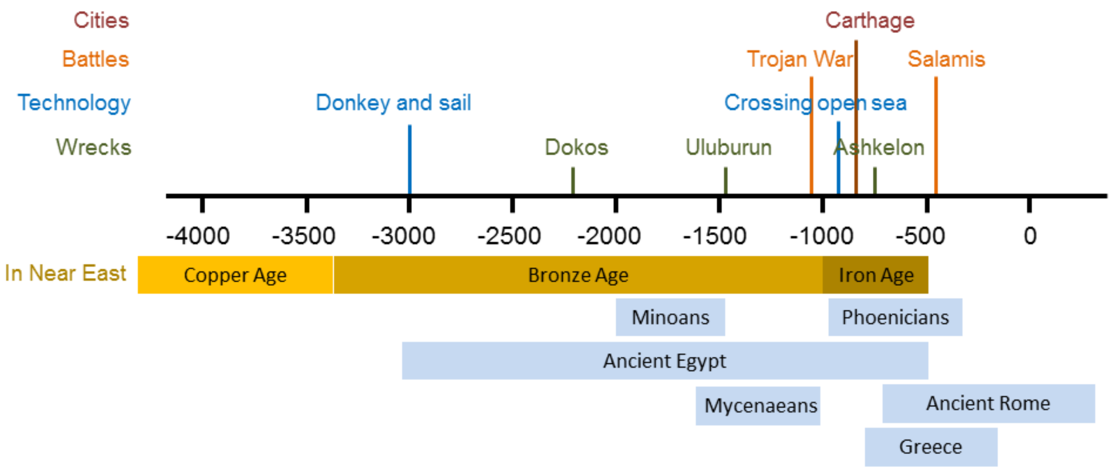
These new routes began to take Phoenician and other sailors over long stretches of open sea. While this had long been conjectured by earlier writers like Braudel (2001, writing in the late 1960s) and Sherratt and Sherratt (1993), contemporary scholars are more confident. Cunliffe (2008) writes about the course of a Phoenician sailor: “Beyond Cyprus, for a ship’s master to make rapid headway west there was much to be said for open-sea sailing. From ... the western end of Cyprus he could have sailed along the latitude to the south coast of Crete ... where excavation has exposed a shrine built in Phoenician fashion. Traveling the same distance again ..., once more following the latitude, would have brought him to Malta” (p. 275-276), a route which became known as the “Route of the Isles.” Abulafia (2011) describes their seafaring similarly: “The best way to trace the trading empire of the early Phoenicians is to take a tour of the Mediterranean sometime around 800 BC. ... Their jump across the Ionian Sea took them out of the sight of land, as did their trajectory from Sardinia to the Balearics; the Mycenaeans had tended to crawl round the edges of the Ionian Sea past Ithaka to the heel of Italy, leaving pottery behind as clues, but the lack of Levantine pottery in southern Italy provides silent evidence of the confidence of Phoenician navigators." (p. 71).

This involved crossing 300 - 500 km of open sea. One piece of evidence for sailing away from the coast are two deep sea wrecks found 65 km off the coast of Ashkelon (Ballard et al., 2002). Of Phoenician origin and dating from about 750 BC, the ships were 14 meters long, and each carried about 400 amphorae filled with fine wine. These amphorae were highly standardized in size and shape. This highlights the change in the scale and organization of trade compared to the Uluburun wreck with its diverse cargo. It also suggests an early form of industrial production supporting this trade.

An unlikely traveler offers a unique lens on the expansion of trade and the density of connections which were forged during this period. The house mouse populated a small area in the Levant until the Neolithic revolution. By 6,000 BC, it had spread into southern Anatolia before populating parts of north eastern Africa and the Aegean in the ensuing millennia (there were some travelers on the Uluburun ship). There were no house mice west of Greece by 1,000 BC. Then, within a few centuries, the little creature turned up on islands and on the mainland throughout the central and western Mediterranean (Cucchi et al., 2005).

The Phoenicians might have been at the forefront of spreading mice, ideas, technology, and goods all over the Mediterranean but others were part of these activities. At the eve of Classical Antiquity, the Mediterranean was constantly criss-crossed by Greek, Etruscan, and Phoenician vessels as well as smaller ethnic groups. Our question here is whether this massive expansion in scale led to locational advantages for certain points along the coast compared to others, and whether these advantages translated into the human activity which is preserved in the

Figure 4.2.1: Timeline



archaeological record. A brief, rough time line for the period we investigate is given in figure 4.2.1.

4.3 Data and key variables

For our Mediterranean dataset we compute a regular grid of 10 × 10 kilometers that spans the area of the Mediterranean and the Black Sea based on a coastline map of the earth from Bjorn Sandvik's public domain map on world borders.[3] We use a Lambert Azimuthal Equal Area projection, with the coordinates 39N, 18.5E as reference point, which is close to the center of the part of the map we study. No projection avoids distortions completely but this one works well for the study of a limited geographical area. The distances of the edges of our 10×10 km grid are close to the true distances: Even at points furthest from the reference points, such as Gibraltar in the west and Sinai in the east, measurement error of both vertical and horizontal lines remains within less than 2 percent of true distances.

[3] We use version 3, available from `http://thematicmapping.org/downloads/world_borders.php`.

We define a grid-cell as coastal if its centroid is within 5 km of a coastline. Grid-cells whose centroid is more than 5 km away from a landmass are classified as sea, the remaining cells are classified as land. Our estimation dataset consists of all coast cells and all land cells within 50 km of a coast cell. Each cell is an observation. There are 11,999 cells in this dataset of which 3,352 are coastal.

We compute the distance between coastal point i and coastal point j moving only over water d_{ij} using the cost distance command in ArcGIS. Our key variable in this study, called c_{di}, measures the number of other coastal cells which can be reached within shipping distance d from cell i. Destinations may include islands but we exclude islands which are smaller than $20km^2$. We also create separate measures, one capturing only connectedness to islands, and a second measuring connectedness to other points on the mainland coast. While we use straight line or shortest distances, we realize that these would have rarely corresponded to actual shipping routes. Sailors exploited wind patterns and currents, and often used circular routes on their travels (Arnaud 2007). Our measure is not supposed to mimic sailing routes directly but simply capture opportunities.[4]

Figure 4.3.1 displays the measure c_{500} for a distance of 500 km; darker points indicate better connected locations. Measures for other distances are strongly positively correlated and maps look roughly similar. The highest connectedness appears around Greece and Turkey partly due to the islands, but also western Sicily and the area around Tunis. The figure also highlights substantial variation of the connectedness measure within countries. The grid of our analysis allows for spatial variation at a fine scale. Figure 4.3.2 shows a histogram of the log connectedness measure for a distance of 500 km. The modes in the rightmost part of the histogram are associated with points in the Aegean.

[4] We do not attempt to use wind patterns to calculate sailing times. Leidwanger (2013), combining modern data on wind speeds and prevailing directions with the sailing logs from sea trials with the replica of a 3rd century BC wreck on a Piraeus to Cyprus route, is an attempt to do this for a small area a few hundred kilometers across off the Turkish coast. He discusses shortcomings and problems with this approach. His work illustrates how far away we still are from being able to extend an exercise like this to an area like the entire Mediterranean.

Figure 4.3.1: Connectedness in the Mediterranean for a 500 km distance

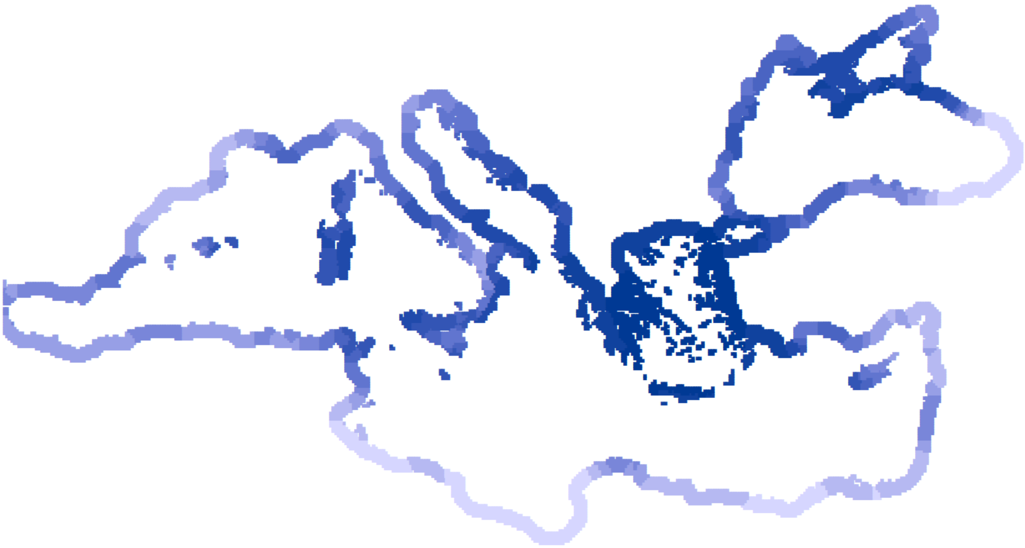
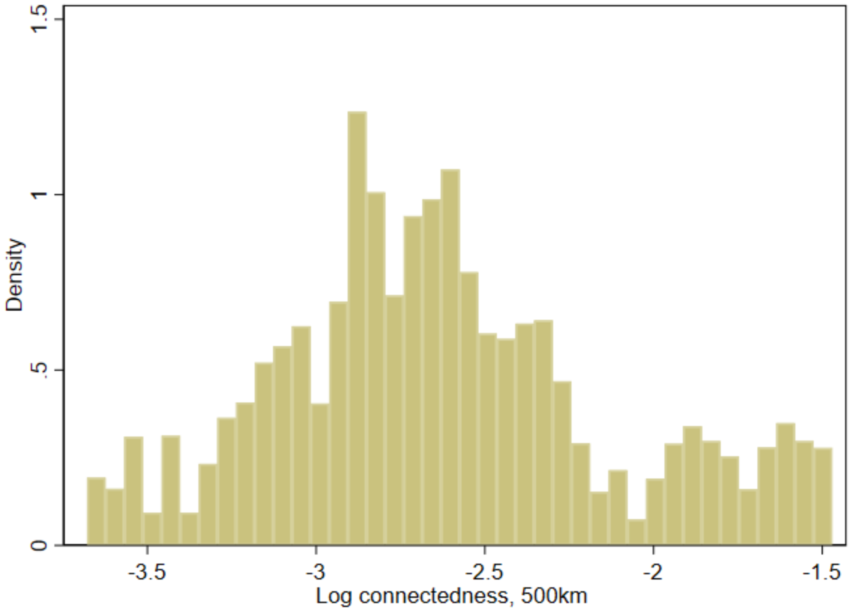


Figure 4.3.2: Distribution of log connectedness at 500 km distance

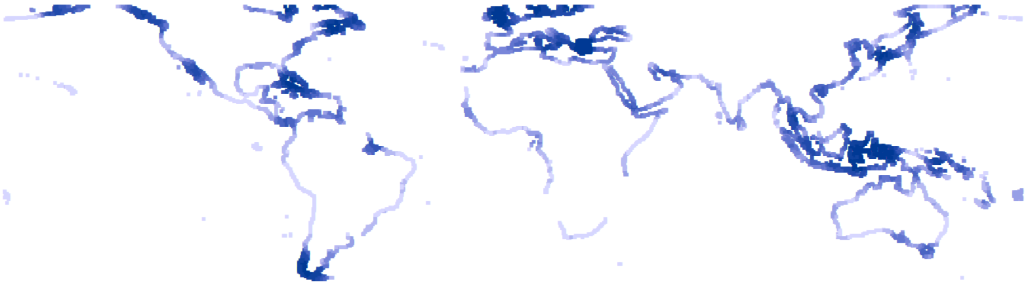


We interpret the measure c_d as capturing connectivity. Of course, coastal shape could proxy for other amenities. For example, a convex coastal shape forms a bay, which may serve as a natural harbor. Notice that our 10 × 10 km grid is coarse enough to smooth out many local geographic details. We will capture bays 50 km across but not those 5 km across. It is these more local features which are likely more relevant for locational advantages like natural harbors. Our grid size also smooths out other local geographic features, like changes in the coastline which have taken place over the past millennia, due, for example, to sedimentation. The broader coastal shapes we capture have been roughly constant for the period since 3,000 BC, which we study (Agouridis, 1997).

Another issue with our measure of connectivity is whether it only captures better potential for trade or also more exposure to external threats like military raids. Overall, it was probably easier to defend against coastal attacks than land-based ones (e.g. Cunliffe (2008, p. 447)) so this may not be a huge concern. But at some level it is obvious that openness involves opportunities as well as risks. In this respect we measure the net effect of better connectivity.

We also compute a global dataset based on a global grid, using a Cylindrical Equal Area projection. We increase the cell size to 50 × 50 kilometers. This is for computational convenience, but also our outcome variable at the global level varies only at the country level and thus spatial precision is less relevant than in the Mediterranean dataset. While we define our global connectedness measure for the whole world, our analysis focuses on the part of the world between -60 degrees and 60 degrees latitude, as units outside that range are unlikely candidates for early urbanization for climatic reasons. In the Southern Hemisphere there is no landmass apart from the Antarctic below 60 degrees, while in the Northern Hemisphere 60 degrees is close to Helsinki, Aberdeen, and Anchorage, well north of climatic conditions particularly favorable to early settlement. We again compute the distance from each coastal grid point to each other coastal grid point by moving only over water. Figure 4.3.3 shows the global connectedness measure c_{500}. The

Figure 4.3.3: Connectedness in the world for a 500 km distance



most connected coastal points are located again near Greece, but also in Southeast Asia, Chile, Britain, and Northern Canada, while Western Africa and Eastern South America have few well connected coastal points.[5]

We measure economic development by counting archaeological sites of settlements. Historians and archaeologists have long debated to what extent the material evidence that has been discovered is representative of actual historical conditions. On one end of the spectrum are warnings like that of Manning (2018, p. 64) that "archaeological evidence, especially for settlement history, is extremely uneven for the first millennium BCE." The idea of a "positivist fallacy" of "making archaeological prominence and historical importance into almost interchangeable terms: in equating what is observable with what is significant" goes back to at least Snodgrass (1992, p. 38). At the other end are optimists such as Broodbank (2013), who concludes that "only a single imbalance is so devastating that it threatens

[5] We only show the connectedness measure for countries where we also have outcome data, hence some countries have missing cells in figure 4.3.3.

to undermine the integrity of the overall study of the Mediterranean. This is the dearth of information on the early societies of the Mediterranean North Africa" (p. 37). We deal with the North African exceptionalism by showing results excluding the North African coast. But Broodbank concludes that "the low archaeological profile of much of Mediterranean North Africa may not entirely be due to a lack of prospection ... In the coming chapters we shall encounter several indications that this was indeed the case" (2013, p. 39).

Whether the archaeological record is representative of history is one issue, another is to obtain a quantitatively useful snapshot of the archaeological record. Our data on settlements for our period of investigation come from the Pleiades Project, an electronic database (Bagnall and Talbert, 2014) at the University of North Carolina, the *Stoa Consortium*, and the *Institute for the Study of the Ancient World* at New York University maintained jointly by the *Ancient World Mapping Center*.[6] The Pleiades dataset is a gazetteer for ancient history. It draws on multiple sources to provide a comprehensive summary of the current knowledge on geography in the ancient world. The starting point for the database is the *Barrington Atlas of the Greek and Roman World* (Talbert, 2000); but it is an open source project and material from multiple other scholarly sources has been added.[7]

The Pleiades data consists of three different databases of which we use the "pleiades-places" dataset. It offers a categorization as well as an estimate of the start and end date for each place. We only keep units that have a defined start and end date, and limit the dataset to units that have a start date before 500 AD. We use two versions of these data, one more restricted (which we refer to as "narrow") and the other more inclusive ("wide"). In the narrow one we only keep units that contain the word "urban" or "settlement" in the categorization. These words can appear alongside other categorizations of minor constructions,

[6] Available at `pleiades.stoa.org`. We use a version of the dataset downloaded in September 2017.

[7] Various historians have assured us that the *Barrington Atlas* is probably the most representative source for the period we are studying.

such as bridge, cemetery, lighthouse, temple, villa, and many others. In the "wide" measure, we include any man-made structure, excluding only natural landmarks (e.g. rivers) and administrative units.[8]

Some of the entries in the Pleiades dataset are located more precisely than others. The dataset offers a confidence assessment consisting of the classifications precise, rough, and unlocated. We only keep units with a precisely measured location.[9] For both datasets, as we merge the Pleiades data onto our grid we round locations to the nearest 10 × 10 kilometers and are thus robust to some minor noise.

Since the Pleiades data is originally based on the *Barrington Atlas* it covers sites from the classical Greek and Roman period well and adequate coverage seems to extend back to about 750 BC. Coverage of older sites seems much more limited as the number of sites with earlier start dates drops precipitously. For example, our wide dataset has 1,565 sites in 750 BC and 5,707 in 1 AD but only 142 in 1,500 BC. While economic activity and populations were surely lower in the Bronze Age, there are likely many earlier sites missing in the data. As a consequence, our estimation results with the Pleiades data for earlier periods may be less reliable.[10]

Our measure of urbanization for a given cell is the number of sites that exist at time t and fall into that cell. We prefer a count of sites over an indicator given that it is scale invariant with respect to the grid size. The maximum number of sites in a cell for the narrow Pleiades measure is 5 but for 98.5% of the cells the value is either 0 or 1.

For our global results, we have only a single early outcome measure: population in 1 AD from McEvedy and Jones (1978). This is the same data as used by Ashraf and Galor (2011b) for a similar purpose. Population density is measured at the level of modern countries, and our sample includes 123 countries.

[8] The raw Pleiades dataset contains some sites that are duplicates and/or have been moved to the errata section of Pleiades. We drop those sites from our analysis.

[9] An exception to this are roads and canals, which typically cannot be interpreted as a single point, and where we therefore also include rough locations.

[10] In Appendix 4.A we present some alternative estimates based on the much earlier *Archaeological Atlas of the World* (Whitehouse et al., 1975), which is more focused on the pre-Classical era but has problems of its own.

4.4 Specification and results

We run regressions of the following type:

$$u_{it} = c_{di}\beta_{dt} + X_i\gamma_t + e_{it}, \tag{4.1}$$

where u_{it} is the urbanization measure for grid point i, c_{di} is the log of the connectivity measure for distance d, and X_i are grid point control variables. For coastal cells, connectivity is simply the connectivity of the respective coastal cells. For inland cells, we assign the connectivity level of the closest coastal cell. We only measure connectivity of a location, not actual trade. Hence, when we refer to trade this may refer to the exchange of goods but could also encompass migration and the spread of ideas. u_{it} measures the number of archaeological sites in each cell and year, which we view as proxy for the GDP of an area. Growth manifests itself both in terms of larger populations as well as richer elites in a Malthusian world. We would expect that the archaeological record captures exactly these two dimensions.

We start by using only linear variables for latitude and longitude as control variables. Latitude captures climatic variation due to the north-south gradient of the region. Climatic conditions also vary in the east-west orientation since proximity to the Atlantic moderates weather variability (Manning 2018, p. 85), and the longitude variable controls for this. Since some of our cells are up to 50 km inland, we also consider distance to the coast as an additional control variable, as well as distance to the Fertile Crescent. This may be important because agriculture spread from the Fertile Crescent throughout the Mediterranean Basin, and various authors have linked the timing of the Neolithic Revolution to later development (Diamond 1997; Hibbs and Olsson 2004; Comin, Easterly, and Gong 2010). We explore dropping the Aegean, to address concerns that our results may be driven exclusively by developments around the Greek islands, by far the best connected area in the Mediterranean. We also show results dropping North Africa to address

Table 4.4.1: Balancing checks

Dependent variable	(1)	(2)	(3)	(4)	(5)	(6)
Agricultural productivity	0.46	0.00	0.53	0.07	0.16	-0.17
(following Galor and Özak (2016))	(0.08)	(0.10)	(0.14)	(0.16)	(0.11)	(0.09)
Ruggedness	0.19	0.15	0.06	-0.05	-0.29	-0.13
(following Nunn and Puga (2012))	(0.14)	(0.19)	(0.29)	(0.28)	(0.16)	(0.16)
River proximity	-3.02	-2.86	-4.40	-3.83	-2.46	-2.94
	(1.73)	(2.14)	(2.96)	(3.33)	(2.09)	(2.19)
Mines proximity	-0.36	0.11	-0.12	0.42	-1.95	-0.03
	(0.37)	(0.74)	(1.21)	(1.47)	(0.74)	(0.67)
Wind	0.32	1.05	-0.52	0.24	0.68	1.20
	(0.16)	(0.23)	(0.30)	(0.34)	(0.17)	(0.22)
Observations	11999	11999	10049	10049	9448	9448
Controls:						
Longitude and latitude	X	X	X	X	X	X
Distance to coast and Fertile Crescent		X		X		X
Dropping Aegean			X	X		
Dropping North Africa					X	X

Coefficients from regressions of various dependent variables on 500 km log connectedness. Standard errors clustered at the level of 200×200 km cells, in parentheses.

concerns that there may be fewer archaeological sites in North Africa due to a relative lack of exploration. This may spuriously correlate with the fact that the coast is comparatively straight. We cluster standard errors at the level of a grid of 200×200 km following Bester, Conley and Hanson (2011). Using a 400×400 km grid as cluster variable results in very similar standard errors.

Our measure of connectedness depends only on coastal and maritime geography and therefore is plausibly exogenous. However, it might be spuriously correlated with other factors that affect early growth, such as agricultural productivity, topographic conditions, or rivers, which provide inland connections. Those factors are hard to measure precisely. Hence, instead of including them on the right-hand side of our regression equation as control variables, we follow the suggestion of Pei, Pischke and Schwandt (2017) and show that they are not systematically related to our measure of coastal connectivity.

The results of these balancing regressions are shown in table 4.4.1. In the first row, we relate connectedness to agricultural productivity, which we construct using data from the FAO-GAEZ database and following the methodology of Galor and Özak (2016): We convert agroclimatic yields of 48 crops in $5' \times 5'$ cells under rain-fed irrigation and low levels of input into caloric yields and assign the maximal caloric yield of the closest $5' \times 5'$ to our grid cells. In the second row, we use the measure of ruggedness from Nunn and Puga (2012), averaged over our 10×10 km cells. Both ruggedness and agroclimatic conditions are standardized to have mean 0 and standard deviation 1. The third row looks at distance to the nearest river. For this, we used Wikipedia to create a list of all rivers longer than 200 km and geocoded their paths from FAO Aquamaps, dropping tributaries. We then calculate the distance from each cell to the nearest river, capping it at 50 km. To make the interpretation easier, we then take the negative of this measure, so that a positive coefficient on connectedness would mean that well-connected cells are closer to rivers. We use distance to the nearest mine, using data from the Oxford Roman Economy Project (2017), coding distance in the same way as for rivers. For wind, we use the AMI Wind on ERS-1 Level 4 Monthly Gridded Mean Wind Fields provided by the Centre de Recherche et d'Exploitation Satellitaire (CERSAT), at IFREMER, Plouzané (France). This dataset contains monthly average wind speeds over oceans on a 1x1 degree grid. We average wind speed over the sailing period from March to October, using the data for 1993. Each coast cell is then assigned the value of the closest wind grid cell.

Column (1) in table 4.4.1 starts by showing the results of balancing regressions just controlling for latitude and longitude. Column (2) also adds a control for distance to the Fertile Crescent and the distance to the coast. Neither agricultural productivity, ruggedness, nor distance to rivers or mines seem to have a large association with our measure of connectedness once we control for the distance to the coast and the Fertile Crescent. The exception is wind speed, which correlates positively with connectedness.

Columns (3) and (4) show that dropping the Aegean from the sample sometimes leads to bigger associations but also impairs precision. When we control for distance to the coast and Fertile Crescent in the sample without the Aegean, associations between the balancing variables and connectedness tend to be small and insignificant, including for wind speed. The only exception is distance to rivers but this relationship is very imprecise. Outside of North Africa, a slight negative association between connectedness and agricultural productivity arises with controls. We are comforted by the fact that our measure of connectedness does not appear to be related to the five variables examined in the table in a systematic way across subsamples. This is especially true once we control for distance to the coast and the Fertile Crescent. As a result, we will use all of latitude, longitude, and distance to the coast and Fertile Crescent as controls in the analyses that follow.

4.4.1 Basic results

In table 4.4.2, we start by showing results for connections within 500 km and the settlement counts in 750 BC from our two datasets. At this time, we expect sailors to make extensive use of direct sea connections, and hence the coefficients β_{dt} from equation (4.1) should be positive. This is indeed the case for a wide variety of specifications. We find stronger results in the wide Pleiades data, and the association is highly significant. The magnitude of these estimates is large. Increasing the connectedness of a cell by one percent increases the number of archaeological sites by around 0.002. Table 4.4.2 reports the means of the dependent variables. A hundred percent increase in connectedness more than doubles the site count in the wide Pleiades data, suggesting an elasticity above one. The coefficient is slightly lower for the narrow site definition. Coefficients decrease in magnitude when we drop the Aegean in column (2), but they remain positive and substantial, indicating that the Aegean alone was not driving the results in column (1). Dropping North Africa in column (3) makes little difference compared to the original results.

Table 4.4.2: Basic results

Dependent variable	Dep. var. mean	(1)	(2)	(3)
Pleiades wide 750BC	0.130	0.207 (0.056)	0.102 (0.043)	0.203 (0.056)
Pleiades narrow 750BC	0.103	0.156 (0.048)	0.074 (0.035)	0.155 (0.048)
Observations		11999	10049	9448
Controls:				
Longitude and latitude		X	X	X
Distance to coast and Fertile Crescent		X	X	X
Dropping Aegean			X	
Dropping North Africa				X

Coefficients from regressions on 500 km log connectedness. Standard errors clustered at the level of 200×200 km cells, in parentheses.

A potential concern with our results might be that we are not capturing growth and urbanization, but simply the location of harbors. To address this, table 4.4.3 repeats the analysis of table 4.4.2, but omitting coastal cells themselves from the calculation of settlement density. Here we are investigating whether a better connected coast gives rise to more settlements further inland. The results are similar to those from the previous table, indicating that the effects we observe are not driven by coastal locations but also manifest themselves in the immediate hinterland of the coast. This bolsters the case that we are seeing real growth effects of better connections. The same is true when we exclude short connections within 100 km from the connectedness variable in table 4.4.4. This is important as we are primarily interested in the longer range connections which opened up with open sea crossing.

The connectedness variable measures how many coastal points a ship can reach from a given starting destination. Coastal points are only a proxy for market access. A more direct measure would be to measure how many settlements a ship can reach, rather than how many coastal points. In table 4.4.5 we use such a more direct measure of market access by counting the number of sites within distance d. To account for the endogenous location of settlements we instrument this market access with the connectedness variable, both in logs. The first stage F-tests we report show

Table 4.4.3: Results excluding coastal cells from outcome definition

Dependent variable	(1)	(2)	(3)
Pleiades wide 750BC	0.174	0.093	0.182
	(0.064)	(0.047)	(0.063)
Pleiades narrow 750BC	0.130	0.072	0.139
	(0.053)	(0.041)	(0.053)
Observations	8647	7552	6631
Controls:			
Longitude and latitude	X	X	X
Distance to coast and Fertile Crescent	X	X	X
Dropping Aegean		X	
Dropping North Africa			X

Coefficients from regressions on 500 km log connectedness. Standard errors clustered at the level of 200×200 km cells, in parentheses. Coastal cells and their sites are omitted from the sample.

Table 4.4.4: Results excluding short connections

Dependent variable	(1)	(2)	(3)
Pleiades wide 750BC	0.200	0.101	0.196
	(0.052)	(0.042)	(0.053)
Pleiades narrow 750BC	0.151	0.075	0.151
	(0.045)	(0.034)	(0.045)
Observations	11999	10049	9448
Controls:			
Longitude and latitude	X	X	X
Distance to coast and Fertile Crescent	X	X	X
Dropping Aegean		X	
Dropping North Africa			X

Coefficients from regressions on 100 km - 500 km log connectedness. Standard errors clustered at the level of 200x200 km cells, in parentheses.

Table 4.4.5: 2SLS regressions for market instrumenting with connectedness

Dependent variable	(1)	(2)	(3)
Pleiades wide 750BC	0.225	0.099	0.250
	(0.056)	(0.038)	(0.065)
First-stage F statistic	32	17	37
Pleiades narrow 750BC	0.178	0.073	0.213
	(0.050)	(0.031)	(0.060)
First-stage F statistic	30	16	32
Observations	11999	10049	9448
Controls:			
Longitude and latitude	X	X	X
Distance to coast and Fertile Crescent	X	X	X
Dropping Aegean		X	
Dropping North Africa			X

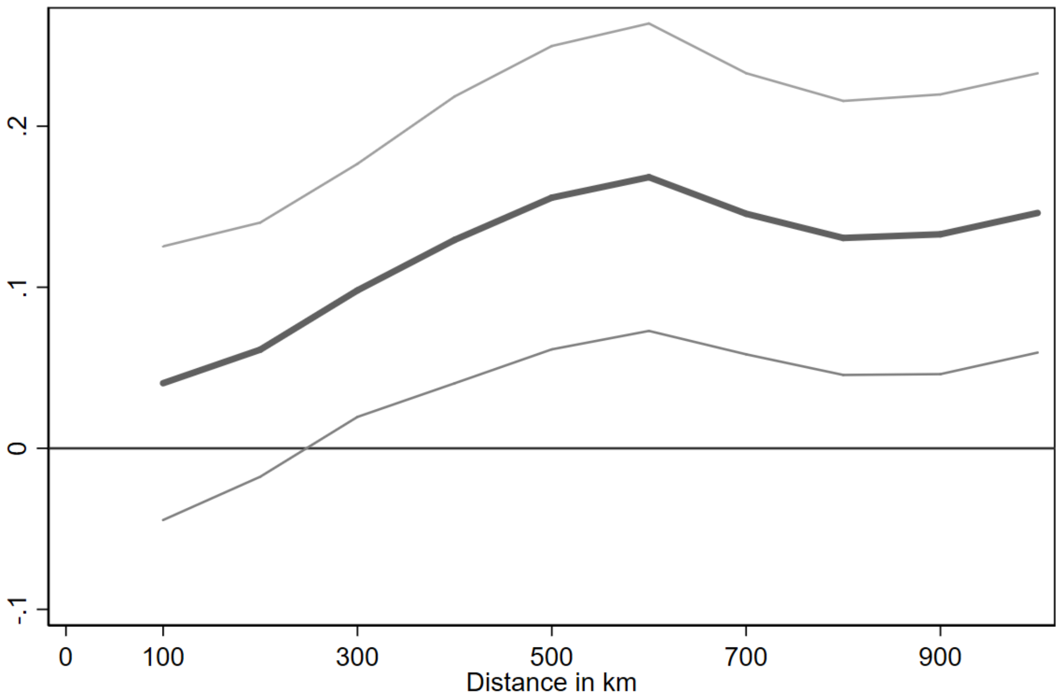
Coefficients from a 2SLS regression of various dependent variables on log market access for 500 km. In the first stage market access is instrumented using 500 km log connectedness. Standard errors clustered at the level of 200x200 km cells, in parentheses.

that connectedness is strongly correlated with market access. The magnitude of the 2SLS effect is similar for all these specifications to the one seen in the connectedness estimation. A one percent increase in market access increases the number of sites by around 0.002.[11] This effect is large compared with existing estimates of the impact of market access. For example, it is about twice as large as the estimate for the land value elasticity in Donaldson and Hornbeck (2016). This may reflect the unusual importance of connections in the Iron Age Mediterranean, where trade served both comparative advantage and insurance functions, as well as facilitating migrations and the spread of ideas. It may also show that in a less technologically advanced economy, market access mattered more relative to other fundamentals.

Table 4.4.6 shows some further robustness checks of our results for different subsamples. Column (1) repeats our baseline results from table 4.4.2. Columns (2) to (4) use only continental cells as starting points, dropping island locations. In column (2), we keep both continent and island locations as potential destinations. Results are similar. Columns (3) and (4) explore whether it is coastal shape or

[11] Table 4.A.1 in Appendix 4.A contrasts these estimates with an OLS estimator. Magnitudes are similar when we exclude the Aegean. Otherwise the 2SLS estimates are larger.

Figure 4.4.1: Coefficients for narrow Pleiades sites by distance, 750BC

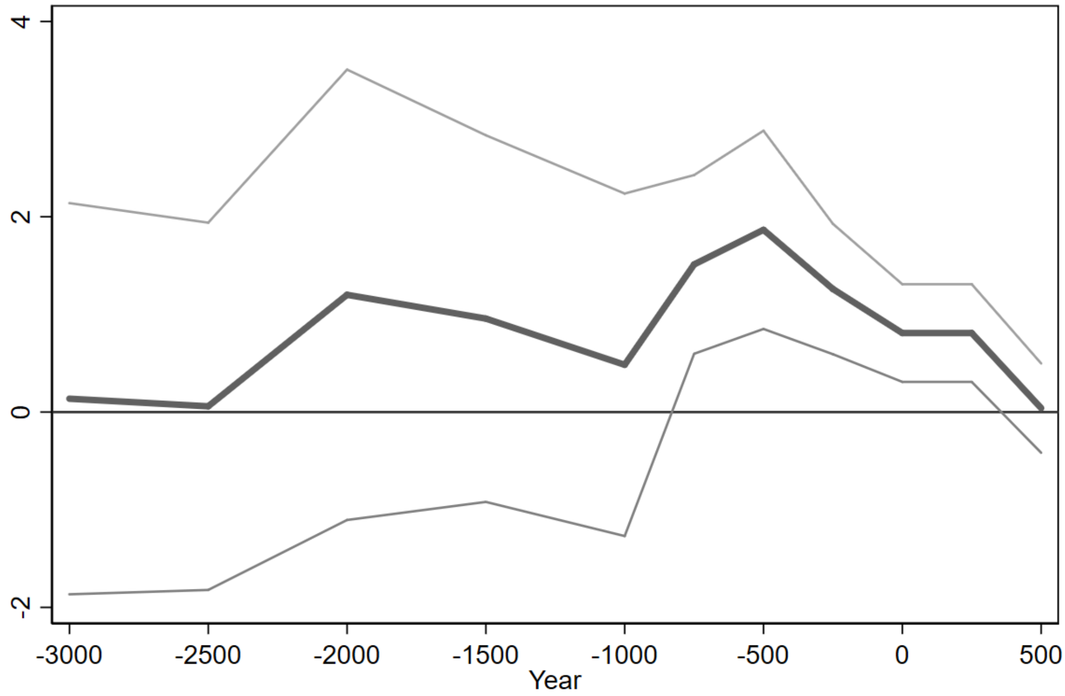


the locations of islands which drive our results. Here, we calculate connectedness using either only island cells as destinations (in column 4) or only continental cells (in column 3). Both matter, but islands are more important for our story. These results suggest that the relationships we find are not driven only by a particular subsample or connection measure.[12]

Our previous results are for connections within a 500 km radius. Figure 4.4.1 displays coefficients for connectivities at different distances, using the basic specification with the narrow Pleiades set of sites in the year 750 BC. It demonstrates that coefficients are fairly similar when we calculate our connectivity measure for other distances. This is likely due to the fact that these measures correlate pretty closely across the various distances. There is a small hump with a peak after 500 km, probably distances which were important during the Iron Age when sailors

[12] We find very similar results using a measure of eigenvector centrality instead of our connectedness variable, which adds weighting to connecting cells, but it is very highly correlated to the original connections measure.

Figure 4.4.2: Scaled coefficients for narrow Pleiades sites over time, 500 km connectedness measure



started to make direct connections between Cyprus and Crete or Crete and Sicily. But we don't want to make too much of this.

Figure 4.4.2 shows results from the narrow Pleiades data over time using the 500 km connectedness measure. The total number of sites differs by year. To enable comparison over time we divide the left hand side by the total number of sites in each year, turning the estimates effectively into elasticities. The figure has various features. Coefficients are positive and sizable but mostly insignificant until 1,000 BC but increase in 750 BC, consistent with the Iron Age expansion of open sea routes. From 500 BC, the effects of connectivity decline and no correlation between sites and connectivity is left by the end of the Roman Empire. In table 4.4.2, we have demonstrated that the large association between connectedness and the presence of sites is replicated across various datasets and specifications for the year 750 BC, so we are fairly confident in that result. Figure 4.4.2 therefore raises two questions: Is the upturn in coefficients between 1,000 BC and 750 BC real or

Table 4.4.6: Results for different connections

	Standard 500 km connectedness			
	(1)	(2)	(3)	(4)
Pleiades wide 750BC	0.207	0.170	0.065	0.078
	(0.056)	(0.076)	(0.071)	(0.026)
Pleiades narrow 750BC	0.156	0.141	0.062	0.062
	(0.048)	(0.062)	(0.057)	(0.021)
Observations	11999	10400	10400	8937
From	All	Continent	Continent	Continent
To	All	All	Continent	Island

Coefficients from a regression on 500 km log connectedness for different subsamples. Robust standard errors, clustered at the level of 200×200 km cells, in parentheses. All regressions control for longitude, latitude, and distance to the coast and the Fertile Crescent.

an artefact of the data? And does the association between sites and connectedness vanish during the course of the Roman Empire? On both counts there are reasons to be suspicious of the Pleiades data. Coverage of sites from before 750 BC is poor in the data while coverage during the Roman period may be too extensive. We explore this last issue in the following subsection.

4.4.2 Persistence

Once geographical conditions have played a role in a site location, do we expect this relationship to be stable into the future? There are two reasons why the answer would be affirmative. Connections should have continued to play a role during the period of the Roman Empire when trade in the Mediterranean reached yet a more substantial level. Even if the relative role of maritime connectivity declined—maybe because sailors got better and distance played less of a role, or other modes of transport, e.g. on Roman roads, also became cheaper—human agglomerations created during the Phoenician period may have persisted. A large literature in urban economics and economic geography has addressed this question and largely found substantial persistence of city locations, sometimes across periods of major historical disruption (Davis and Weinstein 2002, Bleakley and Lin 2012, Bosker and

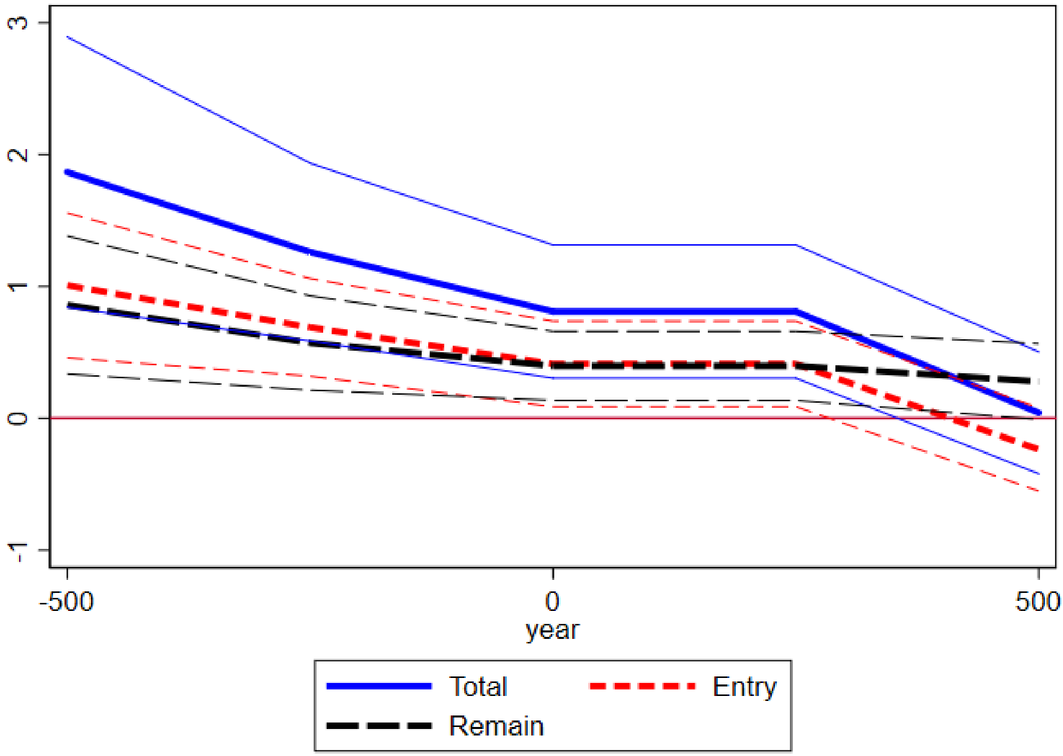
Buringh 2017, Michaels and Rauch 2018, among others). Either explanation is at odds with the declining coefficients over time in figure 4.4.2 after 750 BC.

We suspect that the declining coefficients in the Pleiades data stems from the fact that the site density is becoming too high during the Roman period. In 750 BC there are 1,565 sites in the wide dataset and this number increases to 5,707 in 1 AD at the height of the Roman Empire.[13] There are only 11,999 cells in our dataset. As a result, our grid is quickly becoming saturated with sites after the start of the Iron Age. We suspect that this simply eliminates a lot of useful variation within our dataset: By the height of the Roman Empire many grid points will be the location of archaeological sites. Moreover, existing sites may be concentrated in well-connected locations already and maybe these sites grow further but our data don't provide an extensive margin of settlement size. New settlements after 750 BC, on the other hand, might arise in unoccupied locations, which are actually less well connected.

In order to investigate this, we split the sites in the Pleiades data into those which existed already in 750 BC but remained in the data in subsequent periods and those which first entered at some date after 750 BC. Figure 4.4.3 shows results for the period 500 BC to 500 AD. As in figure 4.4.2, we show coefficients divided by the mean number of sites in the period. The blue, solid line shows the original coefficients for all sites. The black, broken line shows coefficients for sites present in 750 BC which remained in the data while the red, dashed line refers to sites that have newly entered since 750 BC. The coefficients for remaining sites are more stable (and only fall because site density rises), while the relationship between connectedness and the location of entering sites becomes weaker and even turns negative towards the end of the period. Because the new entrants make up an increasing share of the total over time, the total coefficients (solid line) are being dragged down by selective site entry towards the end of the Roman era. This is consistent with the results of Bosker and Buringh (2017) for a later period, who find that having a previously existing city close by decreases a location's chance of becoming a city seed itself.

[13] See table 4.A.2 in the appendix for more details on the numbers of sites in each dataset and time period.

Figure 4.4.3: Scaled coefficients for wide Pleiades sites: Entry, existing, total

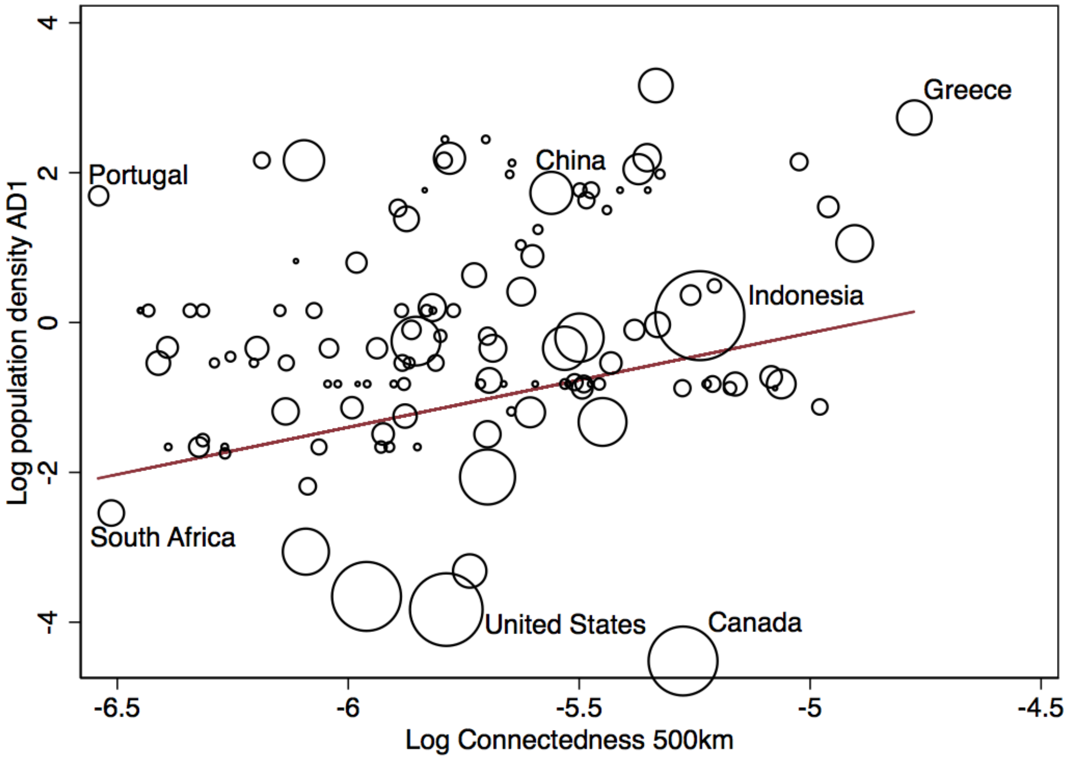


4.4.3 Results for a world scale

Finally, we corroborate our findings for the Mediterranean at a world scale, using population in 1 AD from McEvedy and Jones (1978) as outcome variable. Population density is measured at the level of modern countries, and the sample includes 123 countries. Recall that we compute connectivity for coastal cells based on a grid of 50 x 50 km cells for this exercise.

We aggregate the world data to the level of countries, which is the unit at which the dependent variable is measured anyway. Figure 4.4.4 is a scatter plot of c_{500} against log population density at the country level. The weights in this figure correspond to the number of coastal grid points in each country. The line in the figure comes from a standard bivariate regression and has a slope of 1.24 (0.99). This estimate very similar to the implied elasticity for the Mediterranean

Figure 4.4.4: Connectedness and population density around 1AD at the world scale



Weights reflect length of coasts of countries. For graphical reasons, the figure omits Bermuda, which is an outlier in terms of connectedness. This is inconsequential for our estimates. The weighted slope (standard error) with Bermuda is 1.24 (0.99), as opposed to 1.26 (1.01) without it. When we include a control variable for the absolute latitude the slope becomes 1.67 (0.85) with Bermuda and 1.70 (0.86) without it.

in table 4.4.2, although the nature of the dependent variable is different. Note that many Mediterranean countries can be found in the upper right quadrant of this plot, highlighting how connectivity in the basin may have contributed to the early development of this region.

Additionally, we regress log population density in 1 AD on log 500km connectedness, controlling for absolute values of latitude and again weighting by the number of coastal grid points in each country.[14] This results in a point estimate for connectivity of 1.67 with a standard error of 0.85.

[14] Neither east-west orientation nor distance from the Fertile Crescent seems to make as much sense on a world scale. Unlike for the Mediterranean, there were various centers of early development around the world.

4.5 Conclusion

We argue that connectedness matters for human development. Some geographic locations are advantaged because it is easier to reach a larger number of neighbors. We exploit this idea to study the relationship between connectedness and early development around the Mediterranean. We argue that this association should emerge most potently when sailors first started crossing open seas systematically. This happened during the time when Phoenician, Greek, and Etruscan sailors and settlers expanded throughout the Mediterranean between 800 and 500 BC. Barry Cunliffe (2008) calls this period at the eve of Classical Antiquity "The Three Hundred Years That Changed the World" (p. 270).

This is not to say that sea trade and maritime networks were unimportant earlier. While we find clear evidence of a significant association between connectedness and the presence of archaeological sites for 750 BC our results are more mixed as to whether this relationship began to emerge at that period because the data on earlier sites are more shaky. On the other hand, we find that once these locational advantages emerged the favored locations retain their urban developments over the ensuing centuries. This is in line with a large literature on urban persistence.

Appendix

4.A Additional specifications

4.A.1 OLS vs 2SLS

Table 4.A.1 provides the 2SLS market access results from table 4.4.5, and contrasts them with their corresponding OLS coefficients.

4.A.2 Alternative data sources

The results in the body of this paper rely on the Pleiades dataset. We repeat part of the exercise using two alternative data sources. First we created an additional dataset of sites from the *Archaeological Atlas of the World* (Whitehouse and Whitehouse 1975). The advantage of the *Whitehouse Atlas* is that it focuses heavily on the pre-historic period, and therefore complements the Pleiades data well. We therefore hoped it would help resolve the issue of whether the association between sites and connectedness changed between the Bronze and Iron Ages.

One possible disadvantage of the Whitehouse data is that it is 40 years old. Although there has been much additional excavation in the intervening period, there is little reason to believe that it is unrepresentative for the broad coverage of sites and locations. The interpretation of the archaeological evidence may well have changed but this is of little consequence for our exercise. Another drawback of the *Whitehouse Atlas* is that the maps are much smaller than in the *Barrington Atlas*. As a result, there may have been a tendency by the authors to choose the number of sites so as to fill each map without overcrowding it. This, however, is offset by the tendency to include maps for smaller areas in locations with many sites. For example, there are separate maps for each of Malta, Crete, and Cyprus but only three maps for all of Iberia. Nevertheless, the particular choice of maps may have influenced which sites are recorded in different parts of the Mediterranean.

The number of sites each period is very different in the Pleiades, Whitehouse, and Barrington data (which we discuss below). Table 4.A.2 displays the number of sites we have in each dataset. We repeat the exercise with the Pleiades data from figure 4.4.1 using the Whitehouse data in figure 4.A.1, showing coefficients scaled by the average number of sites per cell for comparability again. We find positive associations between the connectedness measure and sites in the *Whitehouse Atlas*, both for the Bronze and Iron Age. As in the Pleiades data, the association is strongest for the measure around 500km. To account for the possibly artificial difference in site density across space in the *Whitehouse Atlas*, we include map fixed effects, where each fixed effect corresponds to sites visible on one of the Whitehouse maps (a site can be shown on more than one map). Figure 4.A.2 shows that results change a bit and become noisier, which reflects the fact that the maps absorb some geographic variation and the relatively small number of sites in the Whitehouse data. Given the confidence intervals, no clear pattern emerges from 4.A.2.

As a second alternative, we record sites directly from the *Barrington Atlas* (Talbert et al 2000). This atlas provides a unified source of towns and cities in the Greek and Roman period. One advantage of the Barrington maps is that they display the sizes of sites in three broad size classes but these are not recorded in the Barrington gazetteer, on which the Pleiades data are based. We digitize the location of cites on the main overview map of this atlas to have one unified source of cities, and record the size of cities visible on that map. The three different size classes are indicated by different font sizes on the map. Instead of an indicator for a site, we code the dependent variable with weights of 1, 2, and 3 corresponding to small, medium and large cities. We believe that this coding corresponds roughly to log size. The largest cities during this period had populations in the 100,000s (e.g. Rome, Carthage), while the smallest ones would have had populations in the 1,000s. This weighting by size allows us to add an intensive margin to the analysis. We merge the sites from the Barrington map with the Pleiades dataset, which records other attributes of the cities, like the time when the site was active. Our dependent variable is either the size class of the city in a cell or the sum of the size classes if multiple cities are present in a cell. We scale the dependent variable by dividing by its mean in the period again to facilitate comparisons over time.

Figure 4.A.3 displays the scaled regression coefficients over the period 750 BC to 500 AD. It shows a similar downward trend of coefficients as we found in the Pleiades dataset in figure 4.4.2. Whether we weight cities by their size or not has very little influence on the results. This suggests that connectedness did not lead sites in better connected places to grow; rather the effects we find must be explained by entry. We should note that the Barrington size classification is not ideal as we only have one single size indicator. Presumably the *Barrington Atlas* records the peak size of the city but it does not provide any information of size over time. We also note that the Barrington results are very noisy, which reflects the relatively small number of sites on the map we coded.

Table 4.A.1: Market access regressions: 2SLS & OLS

	2SLS			OLS		
Dependent variable	(1)	(2)	(3)	(4)	(5)	(6)
Pleiades wide 750BC	0.225	0.099	0.250	0.124	0.091	0.147
	(0.056)	(0.038)	(0.065)	(0.023)	(0.021)	(0.031)
First-stage F statistic	32	17	37			
Pleiades narrow 750BC	0.178	0.073	0.213	0.091	0.065	0.121
	(0.050)	(0.031)	(0.060)	(0.018)	(0.016)	(0.026)
First-stage F statistic	30	16	32			
Observations	11999	10049	9448	11999	10049	9448
Controls:						
Longitude and latitude	X	X	X	X	X	X
Distance to coast and Fertile Crescent	X	X	X	X	X	X
Dropping Aegean		X			X	
Dropping North Africa			X			X

Coefficients from 2SLS and OLS regressions using 500km market access. Standard errors clustered at the level of 200x200 km cells, in parentheses.

Table 4.A.2: Number of sites in the different datasets

Time period	Pleiades narrow	Pleiades wide	Whitehouse	Barrington
-3000	28	37		
-2000	85	119		
-1500	105	142	243	
-1000	100	116		
-750	1,235	1,565	322	75
-500	2,126	2,772		97
0	3,617	5,707		120
500	2,265	3,667		107

Figure 4.A.1: Scaled Whitehouse results by distance, different periods

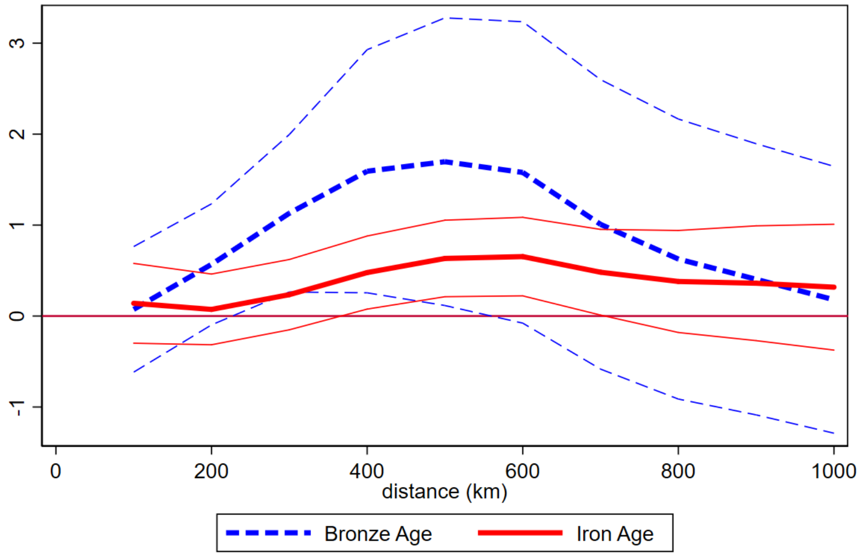


Figure 4.A.2: Scaled Whitehouse results by distance, different periods with map fixed effects

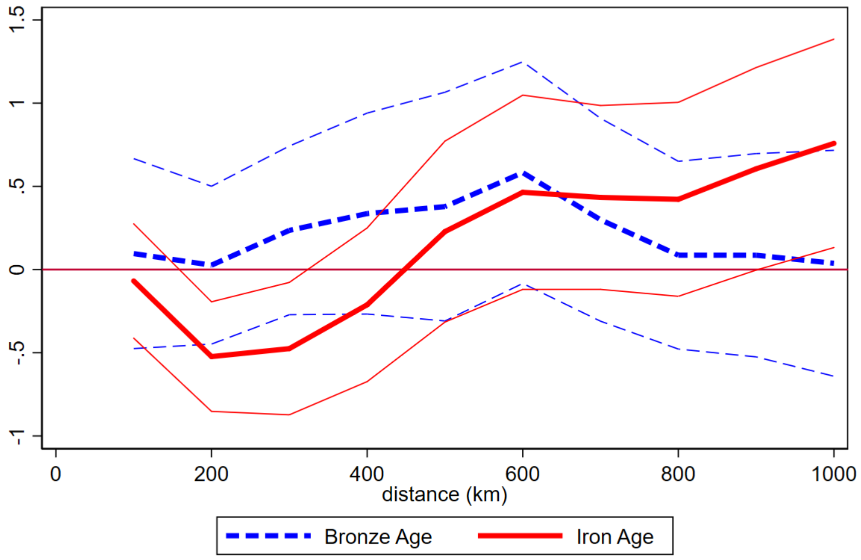
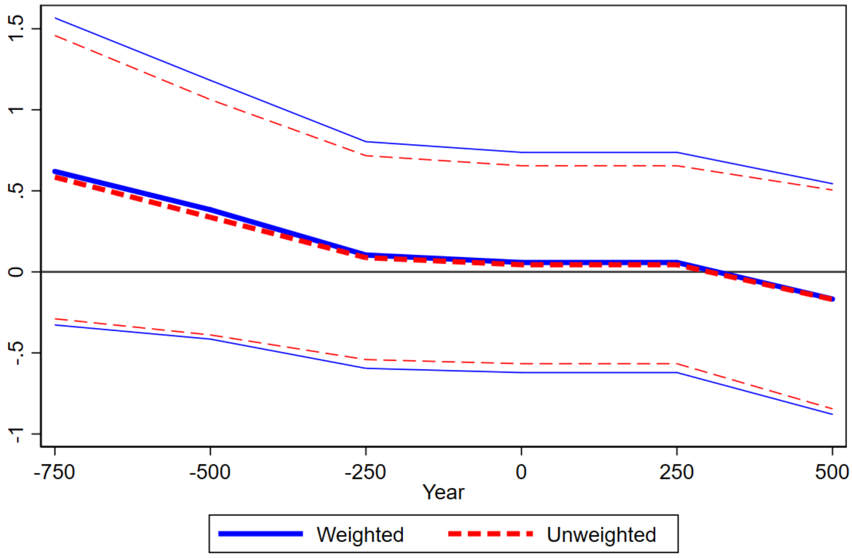


Figure 4.A.3: Scaled Barrington results over time, 500km connectedness measure



4.B Coding of Whitehouse sites

To create the Whitehouse dataset, we geo-referenced all entries within 50km of the coasts on 28 maps covering the Mediterranean and Black Sea in the *Whitehouse Atlas* ourselves. Using the information in the map titles and accompanying text, we classified each map as belonging to one of three periods: the Neolithic, the Bronze Age, or the Iron Age and later. Some maps contain sites from multiple periods but give a classification of sites, which we use. Other maps straddle periods without more detailed timing information. In this case, we classified sites into the three broad periods ourselves using resources on the internet. In a few cases, it is not possible to classify sites clearly as either Neolithic or Bronze Age in which case we classified them as both (see below for details).

Table 4.B.1 provides details of our classification of the maps. The maps on pages 72, 76, 90, and 96 straddle both the Neolithic and Bronze Age period, while the map on page 102 could refer to either the Bronze or Iron Age. For these maps, we narrowed down the dating of sites based on resources we could find on the Internet about the respective site. Table 4.B.2 provides details of our dating.

Table 4.B.1: Classification of maps in the *Whitehouse Atlas*

Pages	Map title/details	Time period
72f.	Neolithic to Bronze Age sites in Anatolia	Bronze Age or earlier
74f.	Hittites and their successors	Bronze Age
76f.	Late prehistoric and proto-historic sites in Near East	Bronze Age or earlier
90f.	Neolithic to Bronze Age sites in Western Anatolia and the Cyclades	Bronze Age or earlier
92f.	Neolithic sites in Greece	Neolithic
94f.	Cyprus	various
96f.	Crete	Bronze Age or earlier
98f.	Mycenaean and other Bronze Age sites in Greece	Bronze Age
100f.	The Mycenaeans abroad	Bronze Age
102f.	The Phoenicians at home	Bronze Age or Iron Age
104f.	The Phoenicians abroad	Iron Age or later
106f.	Archaic and Classical Greece	Iron Age or later
108f.	The Greeks overseas	Iron Age or later
110f.	Neolithic sites in the central Mediterranean	Neolithic
112f.	Copper and Bronze Age sites in Italy	Bronze Age
114f.	Copper and Bronze Age sites in Sicily and the Aeolian Islands	Bronze Age
116f.	Copper and Bronze Age sites in Corsica and Sardinia	Bronze Age
118f.	Early Iron Age sites in the central Mediterranean	Iron Age or later
120f.	The central Mediterranean: Carthaginians, Greeks and Etruscans	Iron Age or later
122	Malta	Bronze Age or earlier
123ff.	Neolithic sites in Iberia	Neolithic
126ff.	Copper and Bronze Age sites in Iberia	Bronze Age
129ff.	Early Iron Age sites in Iberia	Iron Age or later
140f.	Neolithic and Copper age sites in France and Switzerland	Neolithic
164f.	Bronze Age sites in France and Belgium	Bronze Age
172f.	The spread of Urnfield Cultures in Europe	Iron Age or later
174f.	The Hallstatt and La Tene Iron Ages	Iron Age or later
176f.	Iron Age sites in Europe	Iron Age or later

Sources and notes for site classification

Dundartepe: The Cambridge Ancient History, 3rd ed. Vol. 1, Part 2, Early History of the Middle East, eds. I. E. S. Edwards, C. J. Gadd, N. G. L. Hammond, 1971, p. 400 and Ancient West and East, Vol 1, Number 2, 2002, ed. Gocha R. Tsetskhladze, p.245

TAY Project: `http://www.tayproject.org/veritabeng.html` under the site name

Wikipedia: `https://en.wikipedia.org` under the site name

Beisamoun: Israel Antiquities Authority, Beisamoun (Mallaha), `http://www.hadashot-esi.org.il/report_detail_eng.aspx?id=809`

Table 4.B.2: Classification of specific sites in the *Whitehouse Atlas*

Map page	Site name	Neolithic	Bronze Age	Iron Age	Source
72	Dundartepe	1	1	0	see notes
72	Fikirtepe	1	1	0	Whitehouse
72	Gedikli	1	1	1	TAY Project
72	Karatas	0	1	1	Wikipedia
72	Kayislar	1	1	0	TAY Project
72	Kizilkaya	0	1	1	Wikipedia (Kizilkaya/Burdur)
72	Kumtepe	1	0	0	Wikipedia
72	Maltepe	1	1	1	TAY Project
72	Mentese	1	0	0	TAY Project
72	Mersin	1	1	1	Wikipedia
72	Silifke	0	1	1	Wikipedia
72	Tarsus	1	1	1	Wikipedia
72	Tilmen Huyuk	1	1	1	TAY Project
72	Troy	0	1	1	Wikipedia
76	Amrit/Marathus	0	1	0	Wikipedia
76	Amuq	1	1	0	Whitehouse
76	Aradus	0	1	1	Wikipedia (Arwad)
76	Atchana/Alalakh	0	1	0	Wikipedia
76	Beisamoun	1	0	0	see notes
76	Byblos	1	1	1	Wikipedia
76	Gaza	0	1	1	Wikipedia
76	Gezer	0	1	1	Wikipedia
76	Hazorea	1	1	0	Whitehouse
76	Kadesh	1	1	0	Wikipedia (Kadesh (Syria))
76	Megiddo	1	1	1	Wikipedia
76	Mersin	1	1	1	Wikipedia
76	Samaria	1	1	1	New World Encyclopedia
76	Sidon	1	1	1	Wikipedia
76	Tainat	1	1	0	Whitehouse
76	Tell Beit Mirsim	0	1	1	see notes
76	Tyre	0	1	1	Wikipedia
76	Ugarit/Ras Shamra	1	1	0	Wikipedia
90	Akrotiraki	1	1	0	see notes
90	Chalandriani	0	0	0	Wikipedia
90	Dhaskalio	0	1	0	Wikipedia
90	Dokathismata	0	1	1	Wikipedia (see notes)
90	Emborio	1	1	0	see notes
90	Fikirtepe	1	1	0	Whitehouse
90	Glykoperama	1	1	0	Whitehouse
90	Grotta	0	1	0	see notes
90	Heraion	1	1	0	Whitehouse
90	Kephala	1	1	0	Whitehouse
90	Kumtepe	1	0	0	Wikipedia
90	Mavrispilia	1	1	0	Whitehouse
90	Paroikia	1	1	0	Whitehouse
90	Pelos	1	1	0	Whitehouse
90	Phylakopi	0	1	0	Wikipedia
90	Poliochni	1	1	0	Wikipedia (see notes)
90	Protesilaos	1	1	0	Whitehouse
90	Pyrgos	1	1	0	Whitehouse

Table 4.B.2: Classification of specific sites in the *Whitehouse Atlas*, continued

Map page	Site name	Neolithic	Bronze Age	Iron Age	Source
90	Saliagos	1	0	0	Wikipedia
90	Spedos	0	1	0	Wikipedia
90	Thermi	0	1	0	Wikipedia (Lesbos)
90	Tigani	1	1	0	Whitehouse
90	Troy	0	1	1	Wikipedia
90	Vathy	1	1	0	Whitehouse
90	Vryokastro	0	1	0	see notes
94	Alambra	0	1	0	Whitehouse
94	Amathous	0	0	1	Whitehouse
94	Anoyira	0	1	0	Whitehouse
94	Arpera	0	1	0	Whitehouse
94	Athienou/Golgoi	0	0	1	Whitehouse
94	Ayia Irini	0	1	0	Whitehouse
94	Ayios Iakovos	0	1	0	Whitehouse
94	Ayios Sozomenos	0	1	0	Whitehouse
94	Dhenia	0	1	0	Whitehouse
94	Enkomi	0	1	0	Whitehouse
94	Erimi	1	0	0	Whitehouse
94	Idalion	1	1	0	Whitehouse
94	Kalavassos	1	0	0	Whitehouse
94	Kalopsidha	0	1	0	Whitehouse
94	Karmi	0	1	0	Whitehouse
94	Karpasia	0	0	1	Whitehouse
94	Kato Paphos	1	1	0	Whitehouse
94	Khirokitia	1	0	0	Whitehouse
94	Kition	0	0	1	Whitehouse
94	Kouklia/ Old Paphos	0	1	0	Whitehouse
94	Kourion	1	1	1	Whitehouse
94	Krini	0	1	0	Whitehouse
94	Ktima	0	0	1	Whitehouse
94	Kyrenia	0	0	1	Whitehouse
94	Kythrea	1	0	0	Whitehouse
94	Lapithos	1	0	0	Whitehouse
94	Myrtou	0	1	0	Whitehouse
94	Nikosia	0	1	1	Whitehouse
94	Nitovikla	0	1	0	Whitehouse
94	Palaiokastro	0	1	0	Whitehouse
94	Palaioskoutella	0	1	0	Whitehouse
94	Petra tou Limniti	1	0	0	Whitehouse
94	Philia	0	1	0	Whitehouse
94	Pyla-Kokkinokremmos	0	1	0	Whitehouse
94	Salamis	0	1	1	Whitehouse
94	Sinda	0	1	0	Whitehouse
94	Soli/Ambelikou	1	0	0	Whitehouse
94	Sotira	1	0	0	Whitehouse
94	Troulli	1	0	0	Whitehouse
94	Vasilia	0	1	0	Whitehouse
94	Vouni	1	1	0	Whitehouse
94	Vounous	0	1	0	Whitehouse

Table 4.B.2: Classification of specific sites in the *Whitehouse Atlas*, continued

Map page	Site name	Neolithic	Bronze Age	Iron Age	Source
96	Amnisos	0	1	0	Wikipedia
96	Apesokari	1	1	0	Wikipedia
96	Apodhoulou	1	1	0	Whitehouse
96	Arkhanes	0	1	0	Wikipedia
96	Armenoi	1	1	0	Minoan Crete
96	Ayia Triadha	0	1	1	Wikipedia (Hagia Triadna)
96	Diktaean Cave	1	1	0	Wikipedia (Psychro Cave)
96	Erganos	1	1	0	Whitehouse
96	Fournou Korifi	0	1	0	Minoan Crete
96	Gournes	1	1	0	Whitehouse
96	Gournia	0	1	0	Minoan Crete
96	Idaean Cave	1	1	0	Wikipedia
96	Kamares Cave	1	1	0	Wikipedia
96	Karfi	0	1	0	Wikipedia
96	Katsamba	1	1	0	Whitehouse
96	Khania	1	1	1	Wikipedia
96	Knossos	1	1	1	see notes
96	Krasi	1	1	0	Wikipedia (Malia, Crete)
96	Mallia	0	1	0	see notes
96	Mirsini	1	1	0	Whitehouse
96	Mirtos	1	1	0	Minoan Crete
96	Mitropolis	1	1	0	Whitehouse
96	Mochlos	0	1	0	Minoan Crete
96	Monastiraki	0	1	0	Wikipedia
96	Mouliana	1	1	0	see notes
96	Palaikastro	0	1	0	Minoan Crete
96	Petras	0	1	0	Wikipedia
96	Phaistos	1	1	1	Wikipedia
96	Pirgos (Nirou Khani)	0	1	0	Wikipedia
96	Platanos	1	1	0	Whitehouse
96	Plati	1	1	0	Whitehouse
96	Praisos	1	1	1	Wikipedia
96	Pseira	1	1	0	Wikipedia
96	Rousses	1	1	0	Whitehouse
96	Sklavokampos	0	1	0	Wikipedia
96	Stavromenos	0	1	0	see notes
96	Tylissos	0	1	0	Wikipedia
96	Vasiliki	0	1	0	Wikipedia
96	Vathypetro	0	1	0	Minoan Crete
96	Zakro	0	1	0	Wikipedia
96	Zou	1	1	0	Minoan Crete

Table 4.B.2: Classification of specific sites in the *Whitehouse Atlas*, continued

Map page	Site name	Neolithic	Bronze Age	Iron Age	Source
102	Adana (Ataniya)	1	1	1	Wikipedia
102	Al Mina	0	0	1	Wikipedia
102	Amrit/Marathus	0	1	0	Wikipedia
102	Antioch	0	0	1	Wikipedia
102	Aradus	0	1	1	Wikipedia
102	Askalon	1	1	1	Wikipedia
102	Atchana/Alalakh	0	1	0	Wikipedia
102	Atlit	0	1	1	Wikipedia
102	Beersheba	1	1	1	Wikipedia
102	Berytus	0	0	1	Wikipedia
102	Byblos	1	1	1	Wikipedia
102	Enkomi	0	1	0	Wikipedia
102	Gaza	0	1	1	Wikipedia
102	Hazor	0	1	1	Wikipedia
102	Jaffa	1	1	1	Wikipedia
102	Kadesh	1	1	0	Wikipedia
102	Kourion	1	1	1	Wikipedia
102	Megiddo	1	1	1	Wikipedia
102	Minet el-Beida	0	1	1	see notes
102	Nikosia	0	1	1	Wikipedia
102	Salamis	0	1	1	Wikipedia
102	Samaria	1	1	1	New World Encyclopedia
102	Sarepta	0	1	1	Wikipedia
102	Shechem	1	1	1	Wikipedia
102	Sidon	1	1	1	Wikipedia
102	Simyra	0	1	1	Wikipedia
102	Tarsus	1	1	1	Wikipedia
102	Tripolis	0	0	1	Wikipedia
102	Tyre	0	1	1	Wikipedia
102	Ugarit/Ras Shamra	1	1	0	Wikipedia
122	Bahrija	0	1	0	Whitehouse
122	Borg in Nadur	0	1	0	Whitehouse
122	Ghar Dalam	1	1	0	Whitehouse
122	Skorba	1	0	0	Whitehouse
122	Tarxien	1	1	0	Whitehouse

New World Encyclopedia: `http://www.newworldencyclopedia.org` under the site name

Tell Beit Mirsim: Biblewalks, `http://www.biblewalks.com/Sites/BeitMirsim.html`

Akrotiraki: `http://www.aegeanislands.gr/discover-aigaio/archaeology-aigiao/archaeology-aigaio.html`

Dokathismata: Entry under Amnorgos, end date unclear but clearly settled during the Classical period

Emborio: www.archaeology.wiki/blog/2016/03/07 /history-chios-seen-exhibits-archaeological-museum/

Grotta: http://www.naxos.gr/en/naxos/sights-and-sightseeing/archaeological-sit article/?aid=19

Poliochni: End date is unclear

Vryokastro: http://www.tinosecret.gr/tour/museums/512-vryokastro.htm

Minoan Crete: http://www.minoancrete.comusingpull-downmenus

Knossos: Wikipedia lists Knossos as abandoned around 1100 BC but the Whitehouse Atlas has it appear again on Iron Age map 106

Mallia: http://www.perseus.tufts.edu/hopper/artifact?name=Mallia&object=Site

Mouliana: https://moulianaproject.org

Stavromenos: https://greece.terrabook.com/rethymno/page/archaelogical-site-of-stavromenos

Minet el-Beida: Wikipedia. No independent dating info for Minet el-Beida. It is routinely referred to as the harbor of Ugarit. Hence dating the same as Ugar

5 Conclusion

This thesis studies the impact of trade and migration shocks on the spatial distribution of economic activity. It provides evidence that both migration and trade cost shocks have permanent effect on the distribution of economic activity. I also find that cities themselves affect the participation of firms in international trade and hence that geography in terms of city density matters for both the aggregate and distributional gains from trade.

Chapter 2 documents differential export intensity of firms and sectors across different city densities, with denser places being more export intensive. It shows that this is driven by firm productivity, sectoral factor intensities, and variable trade costs, that decrease with city size, while fixed cost of exporting that increase with city density provide a counterveiling force. I also document that a reduction in trade cost leads to an increase in concentration of economic activity in denser places. One corrolorary of this finding that I aim to explore in further work is that we overestimate the gains from trade. Since trade liberalization leads to a concentration of economic activity in denser places it also increases the overall amount of congestion in the economy which is not accounted for in standard trade models, that measure the productivity increases trade liberalization which includes additional gains from agglomeration.

Chapter 3 documents the effect of an (internal) migration shock on cities of various sizes and the city size distribution. It shows that on average an exogenous population shock does not propagate through the system, suggesting that the distribution of population in space is very path dependent and temporary policy measures can have a permanent effect on the distribution of population. I do find heterogeneity across locations. Rural areas tend to experience an outflow of incumbents following a positive population shock while cities experience additional inflows suggestive of agglomeration benefits from additional population.

Chapter 4 provides evidence on the importance of trade for growth and the location of settlements during the Iron Age. It documents that areas that were better connected via sea along the Mediterranean featured more historical sites. This association is particularly strong around 750 BC when the Phoenicians started to systematically engage in crossing the open sea as opposed to sailing along the coasts. This finding also holds at the global level, where there is a positive correlation between maritime connectedness and population density in 1 AD.

Overall, this thesis provides evidence that trading opportun ities and migration shocks affect the lon-run distribution of city sizes and that the agglomeration-congestion cost trade-off matters for the gains from trade.

Bibliography

ABULAFIA, D. (2011): *The Great Sea: A Human History of the Mediterranean.* Oxford University Press, New York; Penguin, London.

ACEMOGLU, D., D. AUTOR, D. DORN, G. H. HANSON, AND B. PRICE (2016): "Import competition and the great US employment sag of the 2000s," *Journal of Labor Economics*, 34(S1), S141–S198.

ACEMOGLU, D., S. JOHNSON, AND J. ROBINSON (2005): "The rise of Europe: Atlantic trade, institutional change, and economic growth," *American economic review*, 95(3), 546–579.

ACKERBERG, D. A., K. CAVES, AND G. FRAZER (2015): "Identification properties of recent production function estimators," *Econometrica*, 83(6), 2411–2451.

AGOURIDIS, C. (1997): "Sea routes and navigation in the third millennium Aegean," *Oxford Journal of Archaeology*, 16(1), 1–24.

ALGAZE, G. (2009): *Ancient Mesopotamia at the dawn of civilization: the evolution of an urban landscape.* University of Chicago Press.

ALLEN, T. (2014): "Information frictions in trade," *Econometrica*, 82(6), 2041–2083.

AMERICAN BIBLE SOCIETY (ed.) (1999): *The Holy Bible, King James Version.* American Bible Society.

ANGRIST, J. D., AND J. S. PISCHKE (2009): *Mostly Harmless Econometrics.* Princeton, NJ: Princeton University Press.

ARKOLAKIS, C., A. COSTINOT, AND A. RODRIGUEZ-CLARE (2012): "New trade models, same old gains?," *American Economic Review*, 102(1), 94–130.

ASHRAF, Q., AND O. GALOR (2011): "Cultural diversity, geographical isolation, and the origin of the wealth of nations," Discussion paper, National Bureau of Economic Research.

AUERBACH, F. (1913): "Das Gesetz der Bevölkerungskonzentration," *Petermanns Geogr Mitt*, 59, 74–76.

AUTOR, D., D. DORN, AND G. H. HANSON (2013): "The China syndrome: Local labor market effects of import competition in the United States," *The American Economic Review*, 103(6), 2121–2168.

BAGNALL, R., AND R. TALBERT (2014): "Pleiades: A Gazetteer of Past Places," .

BALLARD, R. D., L. E. STAGER, D. MASTER, D. YOERGER, D. MINDELL, L. L. WHITCOMB, H. SINGH, AND D. PIECHOTA (2002): "Iron age shipwrecks in deep water off Ashkelon, Israel," *American Journal of Archaeology*, pp. 151–168.

BARJAMOVIC, G., T. CHANEY, K. COŞAR, AND A. HORTAÇSU (2019): "Trade, merchants, and the lost cities of the bronze age," *The Quarterly Journal of Economics*, 134(3), 1455–1503.

BEINART, W. (2001): *Twentieth-Century South Africa.* Oxford University Press.

Bernard, A. B., S. J. Redding, and P. K. Schott (2007): “Comparative advantage and heterogeneous firms,” *The Review of Economic Studies*, 74(1), 31–66.

Black, D., and V. Henderson (2003): “Urban evolution in the USA,” *Journal of Economic Geography*, 3(4), 343–372.

Bleakley, H., and J. Lin (2012): “Portage and path dependence,” *The quarterly journal of economics*, 127(2), 587–644.

——— (2015): “History and the Sizes of Cities,” *American Economic Review*, 105(5), 558–63.

Borjas, G. J. (1999): “The economic analysis of immigration,” *Handbook of Labor Economics*, 3, 1697–1760.

Bosker, M., and E. Buringh (2017): “City seeds: Geography and the origins of the European city system,” *Journal of Urban Economics*, 98, 139–157.

Brakman, S., H. Garretsen, and M. Schramm (2004): “The strategic bombing of German cities during World War II and its impact on city growth,” *Journal of Economic Geography*, 4(2), 201–218.

Broodbank, C. (2006): “The origins and early development of Mediterranean maritime activity.,” *Journal of Mediterranean Archaeology*, 19(2).

——— (2013): *The making of the Middle Sea*. Thames and Hudson Limited.

Brülhart, M., C. Carrère, and F. Robert-Nicoud (fortchoming): “Trade and Towns: Heterogeneous Adjustment to a Border Shock,” *Journal of Urban Economics*.

Christopher, A. J. (2001): *The Atlas of Changing South Africa*. Routledge.

Combes, P.-P., G. Duranton, L. Gobillon, D. Puga, and S. Roux (2012): “The productivity advantages of large cities: Distinguishing agglomeration from firm selection,” *Econometrica*, 80(6), 2543–2594.

Conley, T. G. (1999): “GMM estimation with cross sectional dependence,” *Journal of Econometrics*, 92(1), 1–45.

Coşar, A. K., and P. D. Fajgelbaum (2016): “Internal geography, international trade, and regional specialization,” *American Economic Journal: Microeconomics*, 8(1), 24–56.

Cucchi, T., J.-D. Vigne, and J.-C. Auffray (2005): “First occurrence of the house mouse (Mus musculus domesticus Schwarz & Schwarz, 1943) in the Western Mediterranean: a zooarchaeological revision of subfossil occurrences,” *Biological Journal of the Linnean Society*, 84(3), 429–445.

Cunliffe, B. W. (2008): *Europe between the Oceans*. Yale univ. press.

Czaika, M., and K. Kis-Katos (2009): “Civil Conflict and Displacement: Village-Level Determinants of Forced Migration in Aceh,” *Journal of Peace Research*, 46(3), 399–418.

Davis, D. R., and J. I. Dingel (2015): "The comparative advantage of cities," Discussion paper, National Bureau of Economic Research.

Davis, D. R., and D. E. Weinstein (2002): "Bones, bombs, and break points: the geography of economic activity," *American Economic Review*, 92(5), 1269–1289.

de Kadt, D., and H. A. Larreguy (2018): "Agents of the Regime? Traditional Leaders and Electoral Politics in South Africa," *The Journal of Politics*, 80(2), 382–399.

de Kadt, D., and M. Sands (2016): "Segregation drives racial voting: New evidence from South Africa," *working paper.*

Dharmadhikari, S., and K. Joag-dev (1983): "Mean, median, mode III," *Statistica neerlandica*, 37(4), 165–168.

Dinkelman, T. (2011): "The Effects of Rural Electrification on Employment: New Evidence from South Africa," *American Economic Review*, 101(7), 3078–3108.

——— (2017): "Long-run Health Repercussions of Drought Shocks: Evidence from South African Homelands," *The Economic Journal*, 127(604), 1906–1939.

Dixon, J., J. Cann, and C. Renfrew (1968): "Obsidian and the origins of trade," *Scientific American*, 218(3), 38–46.

Donaldson, D. (2018): "Railroads of the Raj: Estimating the impact of transportation infrastructure," *American Economic Review*, 108(4-5), 899–934.

Donaldson, D., and R. Hornbeck (2016): "Railroads and American economic growth: A "market access" approach," *The Quarterly Journal of Economics*, 131(2), 799–858.

Duranton, G. (2007): "Urban evolutions: The fast, the slow, and the still," *The American Economic Review*, pp. 197–221.

Eeckhout, J. (2004): "Gibrat's law for (all) cities," *American Economic Review*, pp. 1429–1451.

Fajgelbaum, P., and S. J. Redding (2014): "External integration, structural transformation and economic development: Evidence from argentina 1870-1914," Discussion paper, National Bureau of Economic Research.

Feenstra, R. C., H. Ma, and Y. Xu (2017): "US Exports and Employment," Working Paper 24056, National Bureau of Economic Research.

Feinstein, C. (2005): *An Economic History of South Africa.* Cambridge University Press.

Feyrer, J. (2009a): "Distance, trade, and income-the 1967 to 1975 closing of the suez canal as a natural experiment," Discussion paper, National Bureau of Economic Research.

——— (2009b): "Trade and Income–Exploiting Time Series in Geography," Discussion paper, National Bureau of Economic Research.

Frankel, J. A., and D. H. Romer (1999): "Does trade cause growth?," *American economic review*, 89(3), 379–399.

Gabaix, X. (1999): “Zipf’s Law for Cities: An Explanation,” *The Quarterly Journal of Economics*, 114(3), 739–767.

Galor, O., and Ö. Özak (2016): “The agricultural origins of time preference,” *American Economic Review*, 106(10), 3064–3103.

Garcia, l., A. Potlogea, N. Voigtlaender, and Y. Yang (2018): “Cities, Trade and Productivity,” *mimeo.*

Gaubert, C. (2018): “Firm Sorting and Agglomeration,” *American Economic Review*, 108(11), 3117–53.

Gaulier, G., and S. Zignago (2010): “BACI: International Trade Database at the Product-Level. The 1994-2007 Version,” Working Papers 2010-23, CEPII.

Giannone, E. (2017): “Skill-biased technical change and regional convergence,” *mimeo, Princeton University.*

Giraut, F., and C. Vacchiani-Marcuzzo (2013): “Territories and urbanisation in South Africa. Atlas and geo-historical information system (Dysturb),” *IRD Editions*, (117).

Gollin, D., R. Jedwab, and D. Vollrath (2016): “Urbanization with and without Industrialization,” *Journal of Economic Growth*, 21(1), 35–70.

González-Val, R., L. Lanaspa, and F. Sanz-Gracia (2013): “New evidence on Gibrat’s law for cities,” *Urban Studies*, 51(1), pp. 93–115.

Hanlon, W. W. (2017): “Temporary shocks and persistent effects in urban economies: Evidence from British cities after the US Civil War,” *Review of Economics and Statistics*, 99(1), 67–79.

Harris, J. R., and M. P. Todaro (1970): “Migration, Unemployment and Development: A Two-Sector Analysis,” *The American Economic Review*, 60(1), 126–142.

Head, K., and T. Mayer (2014): “Gravity equations: Workhorse, toolkit, and cookbook,” in *Handbook of international economics*, vol. 4, pp. 131–195. Elsevier.

Helpman, E. (2016): “Globalization and wage inequality,” Discussion paper, National Bureau of Economic Research.

Henderson, J. V. (1974): “The Sizes and Types of Cities,” *The American Economic Review*, 64(4), pp. 640–656.

Henderson, J. V. (2005): “Urbanization and growth,” *Handbook of economic growth*, 1, 1543–1591.

Hering, L., and S. Poncet (2010): “Market access and individual wages: Evidence from China,” *The Review of Economics and Statistics*, 92(1), 145–159.

Holmes, T. J., and S. Lee (2010): “Cities as six-by-six-mile squares: Zipf’s law?,” in *Agglomeration Economics*, pp. 105–131. University of Chicago Press.

Imbert, C., and J. Papp (2018): “Short-term Migration and Rural Workfare Programs: Evidence from India,” *Working paper, University of Warwick.*

Jared, D. (1997): "Guns, germs, and steel: the fates of human societies," *NY: WW Norton & Company*, 14.

Knappett, C., T. Evans, and R. Rivers (2008): "Modelling maritime interaction in the Aegean Bronze Age," *Antiquity*, 82(318), 1009–1024.

Kovak, B. K. (2013): "Regional effects of trade reform: What is the correct measure of liberalization?," *The American Economic Review*, 103(5), 1960–1976.

Kremer, M., and S. Watt (2006): "The globalization of household production," *Weatherhead Center For International Affairs, Harvard University.*

Krugman, P. (1980): "Scale economies, product differentiation, and the pattern of trade," *The American Economic Review*, 70(5), 950–959.

Krugman, P., and R. L. Elizondo (1996): "Trade policy and the Third World metropolis," *Journal of Development Economics*, 49(1), 137 – 150.

Lapping, B. (1986): *Apartheid: A History.* Grafton Books.

Lewis, W. A. (1954): "Economic Development with Unlimited Supplies of Labour," *The Manchester School*, 22(2), 139–191.

MacEvedy, C. (1967): "The Penguin atlas of ancient history," .

Manning, J. G. (2018): *The open sea: the economic life of the ancient Mediterranean world from the Iron Age to the rise of Rome.* Princeton University Press.

McEvedy, C., and R. Jones (1978): *Atlas of world population history.* Penguin Books Ltd, Harmondsworth, Middlesex, England.

Melitz, M. J. (2003): "The impact of trade on intra-industry reallocations and aggregate industry productivity," *Econometrica*, 71(6), 1695–1725.

Michaels, G., and F. Rauch (2018): "Resetting the urban network: 117–2012," *The Economic Journal*, 128(608), 378–412.

Michaels, G., F. Rauch, and S. Redding (2012): "Urbanization and structural transformation," *Quarterly Journal of Economics*, 127(2), 535–586.

Miguel, E., and G. Roland (2011): "The long-run impact of bombing Vietnam," *Journal of Development Economics*, 96(1), 1 – 15.

Neame, L. E. (1962): *The History of Apartheid.* Pall Mall Press.

Nunn, N., and D. Puga (2012): "Ruggedness: The blessing of bad geography in Africa," *Review of Economics and Statistics*, 94(1), 20–36.

Ogura, M. (1996): "Urbanization and apartheid in South Africa: Influx controls and their abolition," *The Developing Economies*, 34(4), 402–423.

Olley, G. S., and A. Pakes (1996): "The Dynamics of Productivity in the Telecommunications Equipment Industry," *Econometrica*, 64(6), 1263–1297.

Oxford Roman Economy Project (2017): "OXREP Mines Database," Discussion paper.

Pascali, L. (2017): “The wind of change: Maritime technology, trade, and economic development,” *American Economic Review*, 107(9), 2821–54.

Pavcnik, N. (2002): “Trade liberalization, exit, and productivity improvements: Evidence from Chilean plants,” *The Review of Economic Studies*, 69(1), 245–276.

Peri, G. (2012): “The Effect Of Immigration On Productivity: Evidence From U.S. States,” *The Review of Economics and Statistics*, 94(1), 348–358.

Peters, M. (2017): “Refugees and Local Agglomeration - Evidence from Germany’s Post-War Population Expulsions,” *Yale University, mimeo.*

Potts, D. (2012): *Whatever Happened to Africa’s Rapid Urbanisation?* Africa Research Institute (ARI).

Purcell, N., and P. Horden (1999): *The Corrupting Sea: A Study of Mediterranean History.* Blackwell Publishers.

Ranis, G., and J. C. Fei (1961): “A theory of economic development,” *The American economic review*, pp. 533–565.

Rauch, F. (2013): “Cities as spatial clusters,” *Journal of Economic Geography*, 4(14), 759–773.

——— (2016): “The geometry of the distance coefficient in gravity equations in international trade,” *Review of International Economics*, 24(5), 1167–1177.

Redding, S., and A. J. Venables (2004): “Economic geography and international inequality,” *Journal of international Economics*, 62(1), 53–82.

Redding, S. J., and D. M. Sturm (2008): “The costs of remoteness: Evidence from German division and reunification,” *American Economic Review*, 98(5), 1766–97.

Romalis, J. (2004): “Factor Proportions and the Structure of Commodity Trade,” *American Economic Review*, 94(1), 67–97.

Rossi-Hansberg, E., and M. L. Wright (2007): “Urban structure and growth,” *The Review of Economic Studies*, 74(2), 597–624.

Ruggles, S., K. Genadek, R. Goeken, J. Grover, and M. Sobek (2017): “Integrated Public Use Microdata Series: Version 7.0,” *Minneapolis: University of Minnesota.*

Schumann, A. (2014): “Persistence of Population Shocks: Evidence from the Occupation of West Germany after World War II,” *American Economic Journal: Applied Economics*, 6(3), 189–205.

Sherratt, S., and A. Sherratt (1993): “The growth of the Mediterranean economy in the early first millennium BC,” *World Archaeology*, 24(3), 361–378.

Snodgrass, A. M. (1992): *An Archaeology of Greece: the present state and future scope of a discipline*, no. 53. Univ of California Press.

Soo, K. T. (2007): “Zipf’s Law and urban growth in Malaysia,” *Urban Studies*, 44(1), 1–14.

Surplus People Project (1985): *The Surplus People.* Ravan Press, Johannesburg.

Talbert, R. J. (2000): *Barrington Atlas of the Greek and Roman World: Map-by-map Directory*, vol. 1. Princeton University Press.

Temin, P. (2006): "Mediterranean trade in Biblical times," *Eli Heckscher, International Trade, and Economic History (Cambridge, Mass., 2006)*, pp. 141–56.

——— (2017): *The Roman market economy*, vol. 44. Princeton University Press.

Turok, I. (2012): *Urbanisation and Development in South Africa: Economic Imperatives, Spatial Distortions and Strategic Responses.* International Institute for Environment and Development.

United Nations (2014): *World Urbanization Prospects: The 2014 Revision, Highlights.* United Nations, Department of Economic and Social Affairs, Population Division.

Whitehouse, D., R. Whitehouse, J. Woodcock, and S. Schotten (1975): *Archaeological atlas of the world.*

Zipf, G. K. (1949): *Human behavior and the principle of least effort.* Addison-Wesley Press.