The World Bank Economic Review, 34(2), 2020, 509–532
doi: 10.1093/wber/lhy030
Article

OXFORD

Migration and Urbanization in Post-Apartheid South Africa

Jan David Bakker, Christopher Parsons, and Ferdinand Rauch

Abstract

Although Africa has experienced rapid urbanization in recent decades, little is known about the process of urbanization across the continent. This paper exploits 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, this study examines the effect of migration on urbanization in South Africa. While it is found that on average there is no endogenous adjustment of population location to a positive population shock, there is heterogeneity in the 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, the evidence suggests that exogenous migration shocks can foster urbanization in the medium run.

JEL classification: R12, R23, N97, O18

Keywords: economic geography, migration, urbanization, natural experiment

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

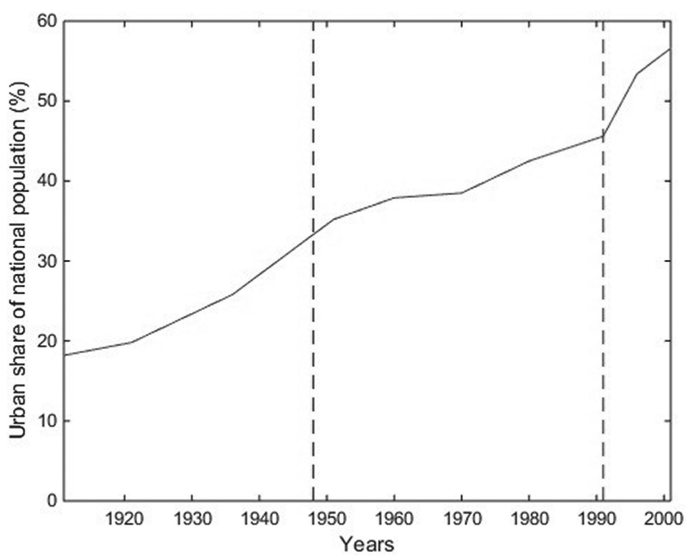
Jan David Bakker is a PhD student in economics at the University of Oxford in economics; his email address is jan.bakker@economics.ox.ac.uk. Christopher Parsons is an associate professor at the University of Western Australia; his email address is christopher.parsons@uwa.edu.au and Ferdinand Rauch (corresponding author) is an associate professor in economics at the University of Oxford; his email address is ferdinand.rauch@economics.ox.ac.uk. The authors gratefully acknowledge the support of the Africa Research Program on Spatial Development of Cities at London School of Economics (LSE) and Oxford funded by the Multi Donor Trust Fund on Sustainable Urbanization of the World Bank and supported by the UK Department for International Development. The authors thank Daniel de Kadt and Melissa Sands who shared census data from 1991. They also thank Frederic Giraut and Celine Vacchiani-Marcuzzo, who shared the Dysturb dataset. Thanks go to Matteo Escudé, James Fenske, Doug Gollin, Leander Heldring, Vernon Henderson, Daniel Kaliski, Daniel de Kadt, Lu Liu, Chris Roth, Ludvig Sinander, Andrea Szabo, Tony Venables, Helene Verhoef, Johannes Wohlfart, and seminar participants at Oxford, 2017 Economic Geography and International Trade research meeting (EGIT) and the 2016 Annual Bank Conference on Africa (ABCA) for useful comments and discussion. A supplementary online appendix is available with this article at *The World Bank Economic Review* website.

a decade (United Nations 2014). Managing the challenges of this rapid transformation is 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, that is, 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. Second, 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. Third, 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.

This paper studies 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 fig. 1 below). This study uses the fact that the locations of the homelands resulted from a long historical process beginning in the eighteenth 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, the study is 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

Figure 1. Urban Share of the National Population (percent), 1911–2001



Source: Authors' own work based on data from Turok (2012).
Note: Vertical dashed lines mark the apartheid regime of the National Party (1948–1991).

a larger inflow of previously mobility-restricted black migrants.[1] Hence, while the homelands are crucial to the empirical design, this article does not study the development of population within the homelands.

Our main findings are threefold. First, the study shows that the distance to homelands is a strong predictor of black population growth in the years following the end of apartheid (i.e. the study's "first stage"). Second, the study shows 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, heterogeneous responses are found to the exogenous population shocks across rural and urban areas. Population levels in areas with initially high population densities experience further agglomeration; that is, exogenous immigration leads to population growth. Only in rural areas does the study continue to find path dependence.[2] 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. The study further investigates this heterogeneity by examining how the effect varies with both initial population density and the reduced form magnitude 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 this 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 the 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 the data on the regional level is that they do not identify internal migrants explicitly, so it is necessary 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 the purposes of this study, information from outside homelands is sufficient. Hence it is not possible to pinpoint the underlying micro-mechanisms driving the results.

The remainder of this paper is organized as follows. Section 2 details the historical development of South Africa thereby providing evidence for the quasi-random location of the homelands. Section 3 discusses the related literature and introduces the theoretical thought experiment that serves as a framework for the empirical analysis presented in section 4. The heterogeneous responses to a positive population shock are discussed in section 5, and section 6 concludes.

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 twentieth century, urbanization in South Africa was shaped by the apartheid policy of the National Party government (1948–1994). Apartheid–literally meaning "apartness"–was by its very nature a spatial concept (Christopher 2001). The government aimed to completely separate the black and nonblack populations.[3] Policies ranged from installing two town

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

2 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).

3 This study uses 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 "Nonblack" category.

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 eighteenth 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). 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). 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 "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.[4] If black Africans 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),[5] 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). Table 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 percent, 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 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

4 It built on preapartheid 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.

5 The Surplus People Project was a nongovernmental organization that documented forced removals through the apartheid government.

Table 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

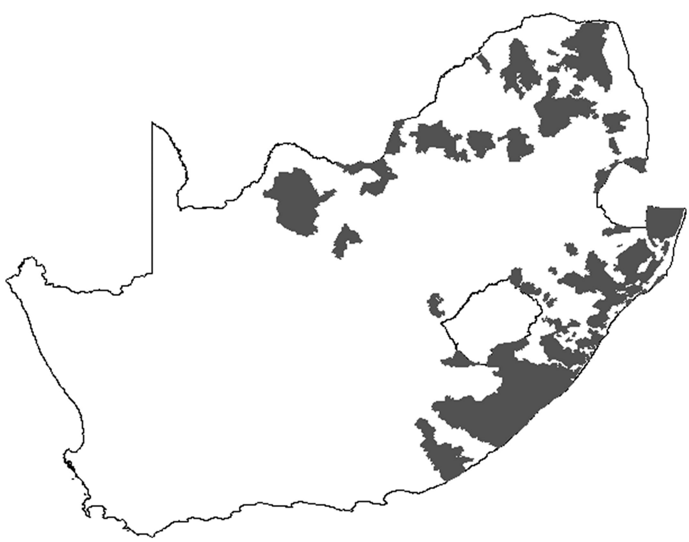
Source: Surplus People Project (1985).

increasing urbanization driven by economic growth and instead channel urbanization dynamics away from (white) cities and towards the homelands.

Table 1 shows the share of the population living in urban areas during apartheid by population group. The three nonblack population groups were already far more urbanized in 1951, and by 1991 around 90 percent of the nonblack 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 fig. 1). Since the nonblack 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 nonrandom and that these could have, for instance, been located nearer to large industrial centers to serve as labor reservoirs; second, that the constraint on internal mobility was binding.

Figure 2. Homelands (Bantustans) Established under Apartheid



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

The homelands established under apartheid (see fig. 2) were confined to areas designated as native reserves under the Native Land Act in 1913. This land comprised 7 percent 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). 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 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). 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). Second, while more than 90 percent of whites and Indians lived in urban areas in 1986, less than 60 percent of blacks did, and the study observes a large jump in urbanization starting in the 1990s. Third, 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).

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, Jedwab, and Vollrath (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.

This paper adds to the literature by studying how the urban system in South Africa reacts to an exogenous populaton shock. The main policy question that is addressed 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 $N_0 + \Delta$. Second, 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 $N_0 + \Delta$. Third, 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). 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).[6] 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 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, characteristics of the shape of the other function can be inferred 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 nonmonotonic, 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 $C(N)$ by contrast, 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 ($lim_{N \to \bar{N}} C(N) = \infty$). 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.[7] 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.

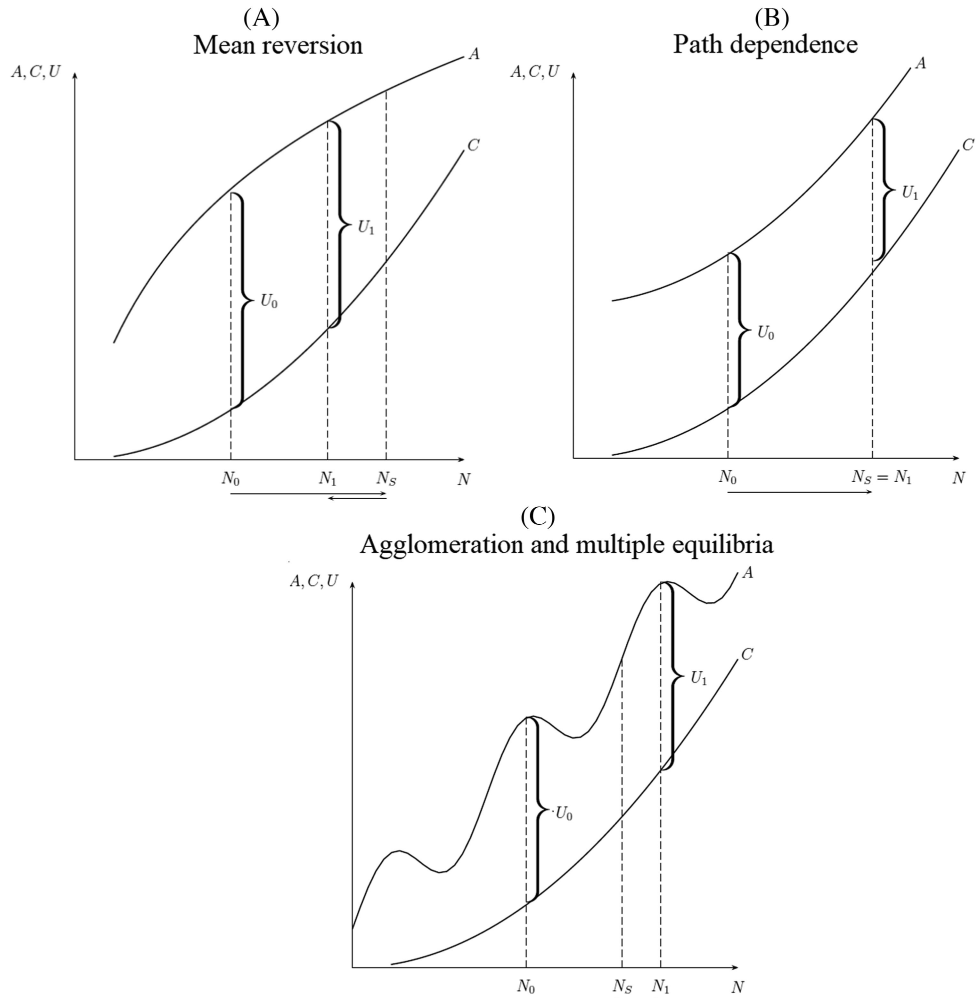
The population level mean reverts (panel A in fig. 3)[8] if the utility at the new population level $U(N_S)$ is below the utility in the initial equilibrium $U(N_0)$ that can be attained in the untreated areas. Agents move from treated areas to the control areas until the utilities are equalized across both types of locations

6 In the context of this stylized model, the study uses changes in the population level and changes in population density interchangeably, since it is considering a fixed amount of space.

7 It is implicitly assumed 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 nonwhite 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 this study abstracts from them in the model for simplicity.

8 Note that the graphs only display the evolution of population in treated areas.

Figure 3. Modified Henderson Model with Gains from Agglomeration and Congestion Costs



Source: Authors' own work.

Note: This figure illustrates three possible urban equilibria in the Henderson model. The x-axis shows population size, the y-axis uility. In Panel A, the equilibrium is unique, and a temporary shock will lead to mean reversion. In Panel B, a temporary shock will lead to long-term transition from one to another of multiple equilibria. In Panel C, a small shock will lead to mean reversion, while a large shock will lead to a new equilibrium size.

$U(N_1^T) = U(N_1^C)$. 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 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 nonmonotonic, implying the existence of multiple equilibria for the spatial distribution of population. In order for the utility function to be nonmonotonic, the slope of the agglomeration function has to be nonmonotonic.[9] This model demonstrates the three possible cases between which this paper aims to distinguish. 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 researchers 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.

This setup relates to a large empirical literature that studies how exogenous shocks to cities affect 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, Garretsen, and Schramm (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.

This paper contributes to this literature in several ways. First, this 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 this study analyzes a positive population shock in which incoming migrants have neither social networks nor private sunk investments, it is able to isolate the effect of gains from agglomeration. Second, it provides evidence from a credible natural experiment that is well-identified and is able to draw upon a much larger sample in comparison with most studies in this literature. Third, this study is the first to provide evidence from Africa, a region that is among the most rapidly urbanizing regions in the world, the continent in which such policy-related questions matter most. Fourth, many of the previous studies focus solely upon urban areas, whereas this study is 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). This study, however, is 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

9 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.

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, Lanaspa, and Sanz-Gracia 2013; and Holmes and Lee 2010). Michaels, Rauch, and Redding (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, that is, subsequent population growth was uncorrelated with initial population density. The research in this paper contributes to this literature by demonstrating the heterogeneous reaction of towns of different size to a population shock.

4. Empirical Analysis

Data

In order to empirically test the three hypotheses, this paper makes 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 makes it possible to distinguish between the short-, medium- and long-run effects of the exogenous population shock. Second, 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 percent 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.

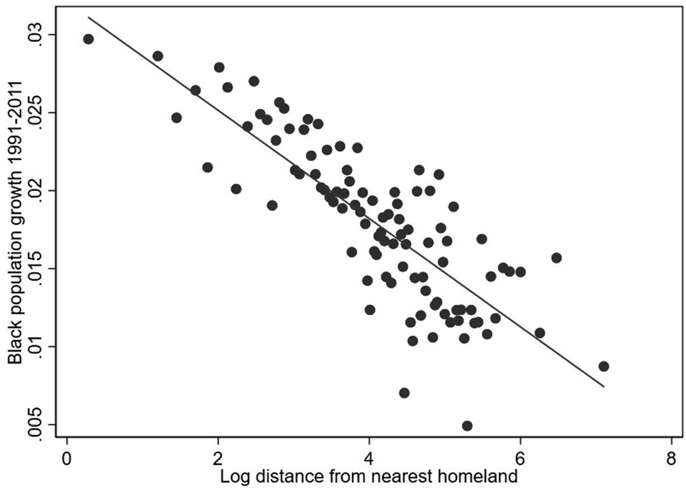
While the Pass Laws Act had already been repealed in 1986 and although identification would be cleaner if data from before 1986 were available, this timeline does not pose a major threat to the identification strategy. This is because the Group Areas Act and the Population Registration Act were still in place, and as such the black population was still severely constrained in its choice of residence until June 1991.[10]

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 4 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, the paper pools neighboring observations into discrete bins to improve clarity. The study specifies 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 2. The assumption may be violated for areas adjacent to the homelands however, as they are likely to be affected by economic spillovers from

10 The data from 1991 do 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 this paper only looks at areas outside the homelands. Another, more relevant, drawback is that there are a few areas that are not covered within South Africa (see fig. S1.1 in the supplementary online appendix S1, available with this article at *The World Bank Economic Review* website). 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. Second, 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 the present analysis since this study excludes areas with high racial tension, which could otherwise bias the results. So, while this reduces the number of observations and therefore the statistical power in the empirical analysis, the parameter estimates will remain consistent.

Figure 4. First Stage



Source: Authors' analysis based on South African census data.
Note: Relationship between distance to the homelands and black population growth, conditional on controls for education, income, population group, population density and employment in 1991 and province fixed effects. Data are collapsed into 100 bins, representing roughly 20 wards each.

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, the present study excludes areas within 10 km from the homelands as a robustness check to ensure that the estimates are not driven by local economic spillovers.

For this instrument to be informative, the cost of migration has to increase 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.[11] 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 S1.1 in the supplementary online appendix S1, the informativeness of the instrument decreases almost monotonically in the granularity of the fixed effects.[12] A trade-off therefore exists between accounting for local unobservables and retaining sufficient identifying variation, in order to ensure that the instrument remains informative. This paper includes province-level fixed effects in order to account for different trends and policies across provinces while allowing for sufficient spatial variation.[13]

Estimation

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

$$\Delta N^{B}_{i,t-1991} = \alpha_2 + log(distance_i)\pi + \boldsymbol{X}_{i,1991}\gamma_2 + \delta_p + \upsilon_m \quad (1)$$

$$\Delta N_{i,t-1991} = \alpha_1 + \widehat{\Delta N^{B}_{i,t-1991}}\beta + \boldsymbol{X}_{i,1991}\gamma_1 + \delta_p + \epsilon_m \quad (2)$$

11 For example, Peri (2012) uses distance to the Mexican border as an instrument for the intensity of migration to different U.S. states.

12 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.

13 Provinces are equivalent to states in the Unites States.

where $\Delta N_{i,\,t-1991}$ denotes overall population growth in ward i between 1991 and t and $\Delta N^{B}_{i,t-1991}$ denotes black population growth.[14] The paper's 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 2). 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 important, there is a lot of circular migration from the Eastern and the Northern Capes into Cape Town. These migration dynamics potentially distort this paper's identification strategy such that it includes a dummy for Cape Town, which significantly increases the predictive power of the 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, 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 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, it is necessary 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 the area, which makes the interpretation this study is looking for impossible. Therefore this paper instead defines the variables of interest as absolute growth rates relative to the overall population in period t where t corresponds to 1996, 2001, and 2011. This study incorporates 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 this study expects 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. The dataset used in this paper does not cover any year other than 1991 in the period pre-1994, and so it can't be used for this purpose. Instead this paper uses 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). The study uses the UA dataset because unlike the MD the units used here are defined consistently over time. In table 3 this study regresses population growth on the distance

14 Δ black population is defined as absolute growth of the black population from 1991 to year t divided by the overall population in t, where t corresponds to 1996, 2001 or 2011. Δ overall population growth is defined similarly using overall instead of black absolute population growth.

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 2. Summary Statistics of Included Variables

Variables	N (1)	mean (2)	sd (3)	min (4)	max (5)
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
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: Authors' analysis based on South African census data.

Note: Log distance is the distance to the nearest homeland in logs. All population growth variables are defined as absolute population growth in the relevant time period divided by overall population. Male share is the share of males in the overall population and population group ratio the share of white population in total population. All other demographic variables are the log of the number of people falling within a given category (e.g. "Finished primary school" is the log of the size of the population that has finished primary school and no further education).

Table 3. Pre-Trend Regressions

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

Source: Authors' analysis based on South African census data.

Note: 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 percent level of confidence are marked with *; at the 95 percent level **; and at the 99 percent level ***. *dist* measures the distance to the nearest homeland in km, *pop* measures population and Δ *pop* measures population growth.

to homelands variable. In columns (4), (5), and (6), the study controls for initial population; in the first three columns it does not. In columns (1) and (4) the study uses all units with nonmissing data in 1980 and 1991. In columns (2) and (5), the study uses all units with nonmissing data in 1991 and 2001. In columns (3) and (6), the study uses all units with nonmissing 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 the study finds the expected negative correlation between distance to the homeland and population growth for 1991–2001, but not for 1980–1991.

5. Results

Baseline Results

Table 4 summarizes the main results and table 5 provides further results from different subsamples as robustness checks. Each cell of the tables summarizes one regression.[17] Before moving on to interpreting the estimated coefficients of interest, this paper will discuss a number of results. 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 medium and long horizon.[18] Weak instrument problems only arise for the short period between 1991 and 1996, and this paper 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 4, column (1)), it is not possible to 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 effect from exogenous black migration on aggregate migration decisions of nonblack incumbents. This is evidence that an exogenous population

17 That is, each cell in the first row of table 4 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.

18 The corresponding first-stage regressions for tables 4 and 5 as well as the other main tables shown here are reported in the supplementary online appendix S1.

Table 4. OLS and 2SLS Baseline Regressions

	OLS (1)	2SLS (2)
	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

Source: Authors' analysis based on South African census data.

Note: This table displays estimates of equation (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 are 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 the supplementary online appendix S1. Coefficients that are significantly different from *one* at the 90 percent level of confidence are marked with *; at the 95 percent level **; and at the 99 percent level ***. Standard errors are in parentheses. Δ black population is defined as absolute growth of the black population from 1991 to year *t* divided by the overall population in *t*, where *t* corresponds to 1996, 2001 or 2011.

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). The study notes 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.

In addition to the baseline regressions, this study reports several regressions based on different subsamples as robustness checks (table 5). The study includes 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) the study removes the dummy variable for Cape Town that is usually included, which does not significantly affect results. In column (3) the study includes separate fixed effects for all the metropolitan areas in the sample, which again does not seem to change the results significantly. As outlined above, the study excludes 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)). The study also excludes 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, the study excludes 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) the study reports results using district instead of province fixed effects. The study also aggregates wards up to the municipality level and runs a separate regression to test whether the results are robust to using a different level of aggregation (column (9)). These robustness tests using different subsamples as reported in table 5 corroborate the baseline findings since none of the

Table 5. 2SLS Regressions Using Different Subsamples

	Dummy for Johannesburg (1)	No dummy for Cape Town (2)	Dummies for all metro areas (3)	Drop within 10 km (4)	Drop <5 percent white (5)	Drop <10 percent white (6)	Drop distance ≥6 (7)	District FE (8)	Municipality level (9)
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

Source: Authors' analysis based on South African census data.

Note: This table displays estimates of equation (2) in the text our second-stage regression, for different subsamples. Column headings denote subsample used in each specification. Each cell presents estimates from a separate regression. The standard errors are clustered the municipality level and presented in parentheses. 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 the supplementary online appendix S1. Coefficients that are significantly different from *one* at the 90 percent level of confidence are marked with *; at the 95 percent level **; and at the 99 percent level ***. Δ black population is defined as the absolute growth rate of the black population from 1991 to year *t* divided by the overall population in year *t*, where *t* corresponds to 1996, 2001 or 2011. displays estimates.

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, the study repeats the entire exercise in table 5 for people of working-age population only. These results are in table 6. Here the study defines working age as the population that is aged between 15 and 64. All coefficients are similar to their counterpart in table 5.

In the main regressions, the study measures 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 the results, the study tests for them in an alternative specification. Table 7 reports the result of a specification where the study runs 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 7 is significantly different from 0 indicates that the dynamic response is also consistent with path dependence.

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 nonmonotonic 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 United States 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, the study is 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, Rauch, and Redding (2012) show that long-run population growth in the United States 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, this study defines dummy variables for high initial population densities and for high initial shares of urbanized households. The results reported in table 8 show that there is a positive and significant interaction between high 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 loses significance in the long run for the share of urbanized households. The results of both specifications indicate that the population dynamics induced by

Table 6. 2SLS Regressions Using Working-Age Population Only

	Baseline (1)	Dummy for Johannesburg (2)	No Cape Town dummy (3)	Dummies for metro areas (4)	Drop within 10 km (5)	Drop <5 percent white (6)	Drop <10 percent white (7)	Drop dist. ≥6 (8)	District FE (9)	Municipality level (10)
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

Source: Authors' analysis based on South African census data.

Note: This table displays estimates of equation (2) our second-stage regression, in the text using working-age population rather than overall population for different subsamples. This includes everyone aged 15 to 64. Column headings denote subsample used in each specification. Each cell presents estimates from a separate regression. The standard errors are clustered at the municipality level and presented in parentheses. 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 the supplementary online appendix S1. Coefficients that are significantly different from *one* at the 90 percent level of confidence are marked with *; at the 95 percent level **; and at the 99 percent level ***. Δ black population is defined as absolute growth of the black population from 1991 to year *t* divided by the overall population in *t*, where *t* corresponds to 1996, 2001 or 2011.

Table 7. Alternative Specification Using Different Time Periods

	Full sample		Working-age population	
	Δ Overall population (1)	Δ Nonblack population (2)	Δ Overall population (3)	Δ Nonblack population (4)
ΔBlack Population	0.167	0.0319	0.246	−0.035
(1991–2001)	(0.139)	(0.484)	(0.153)	(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

Source: Authors' analysis based on South African census data.
Note: This table reports 2SLS results. In the second stage, the study regresses different population growth measures for the period 2001–2011 on predicted black population growth in the previous period (1991–2001). In the first stage the study instruments 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 percent level of confidence are marked with *; at the 95 percent level **; and at the 99 percent level ***. Standard errors are clustered at the municipality level and presented in parentheses. Δ black population is defined as absolute growth of the black population from 1991 to year *t* divided by the overall population in *t*, where *t* corresponds to 1996, 2001 or 2011.

a positive population shock differ between less densely populated rural areas and highly populated urban areas. While the study 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.

The study demonstrates the effects graphically in fig. 5, which corresponds closely to panel C in table 8. The study aggregates observations into bins to increase clarity, it specifies 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 evidence of agglomeration economies appears: Inflows of people lead the rest of the population to positively react.

The study next considers the size of the population shock as an additional dimension of heterogeneity. So the study estimates how overall population growth varies with the size of the shock and initial population density. In order to do so, the study combines the deciles of the two distributions and estimates 100 distinct conditional means:

$$\Delta N_{i,2011} = \sum_{j=1}^{10}\sum_{k=1}^{10}\beta_{j,k}\Big[1[\text{if } Popden_{i,1991} \text{ in decile } j] \times 1[\text{if } distance_i \text{ in decile } k]\Big] + \gamma' X_{i,1991} + \delta_p + \epsilon_m \tag{3}$$

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 form, that is, 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 fig. 6 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 overproportionally from a positive population shock.

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).

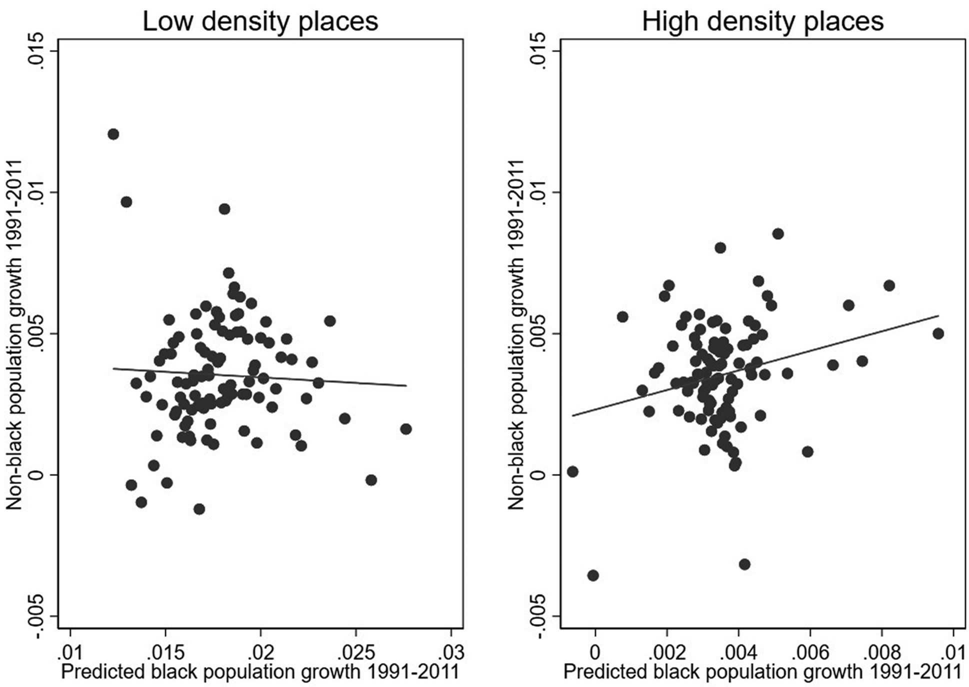
Table 8. Heterogeneity with Respect to the Initial Population Density and Level of Urbanization

	High population density dummy (1)	High urban share dummy (2)
Panel A: Population growth rates (1991–1996)		
ΔBlack Population	1.026	−6.333
	(1.425)	(11.35)
High initial urban share dummy × ΔBlack Population		7.589
		(11.56)
High initial population density dummy × ΔBlack Population	0.509	
	(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 × ΔBlack Population		0.178**
		(0.084)
High initial population density dummy × ΔBlack Population	0.681**	
	(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 × ΔBlack Population		0.040
		(0.062)
High initial population density dummy × ΔBlack Population	0.348**	
	(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

Source: Authors' analysis based on South African census data.

Note: This table displays estimates of equation (2) our second-stage regression, in the text with an additional interaction term. Each column displays one specification. The standard errors are clustered at the municipality level and presented in parentheses. 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 percent most dense areas. An area is defined as having a high urban share if it is among the areas with the 75 percent 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 the supplementary online appendix S1. Coefficients on the interaction terms that are significantly different from zero at the 90 percent level of confidence are marked with *; at the 95 percent level **; and at the 99 percent level ***. Δ black population is defined as absolute growth of the black population from 1991 to year *t* divided by the overall population in *t*, where *t* corresponds to 1996, 2001 or 2011.

Figure 5. Population Reaction for High- and Low-Density Places

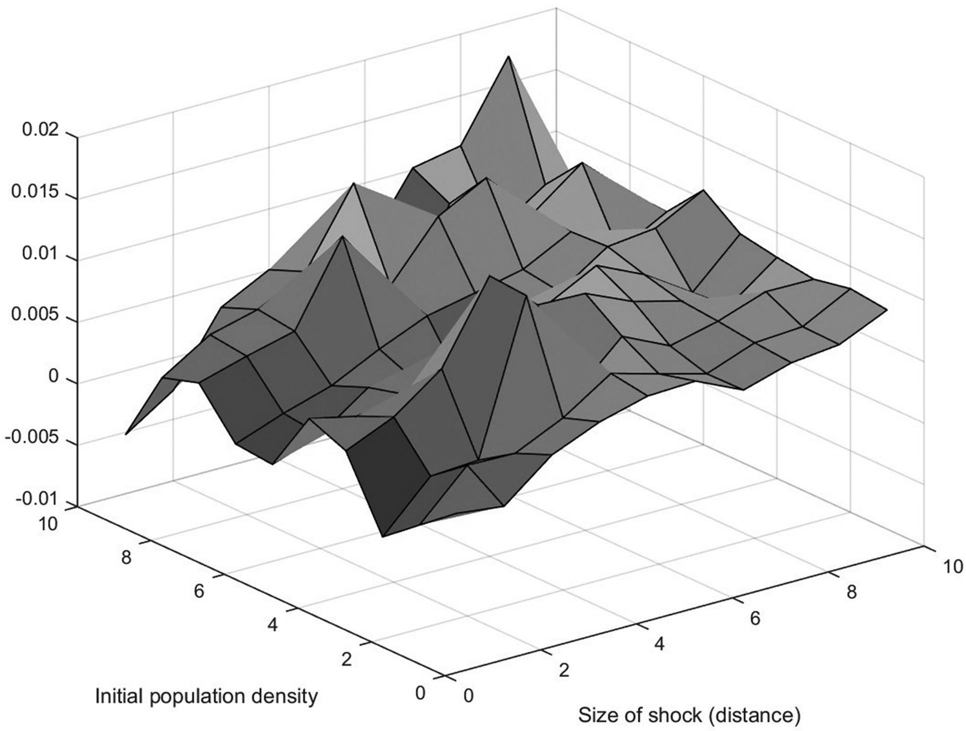


Source: Authors' analysis based on South African census data.
Note: The figure displays the relationship between the predicted growth in the black population between 1991 and 2011 and its impact on non-black population growth over the same period, distinguishing between initially low and high density locales. These results are analogous to the results in Panel C of table 8.

In the context of the modified Henderson model presented in section 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 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 are several 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.

Figure 6. The Effect of Initial Density and Size of Shock for Population Growth



Source: Authors' analysis based on South African census data.
Note: This figure displays the $\beta_{j,k}$ coefficients resulting from estimating equation (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.

A similar result emerges in a standard model used in the migration literature (e.g., Borjas 1999; 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 the study analyzes 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). 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.

6. Conclusion

This study examines 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,that is, 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, that study finds 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. This study provides 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 its slope is nonmonotonic. 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.

References

American Bible Society, ed. (1999). *The Holy Bible, King James Version*. New York City: American Bible Society.

Angrist, J. D., and J. S. Pischke. 2009. *Mostly Harmless Econometrics*. Princeton, NJ: Princeton University Press.

Auerbach, F. 1913. "Das Gesetz der Bevölkerungskonzentration." *Petermanns Geographische Mitteilungen* 59 (1): 74–76.

Beinart, W. 2001. *Twentieth-Century South Africa*. Oxford, United Kingdom: Oxford University Press.

Black, D., and V. Henderson. 2003. "Urban Evolution in the USA." *Journal of Economic Geography* 3 (4): 343–72.

Bleakley, H., and J. Lin. 2012. "Portage and Path Dependence." *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." In *Handbook of Labor Economics* vol. 3A, edited by O. Ashenfelter and D. Card, 1697–760. Amsterdam: North Holland.

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): 201–18.

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

Conley, T. G. 1999. "GMM Estimation with Cross Sectional Dependence." *Journal of Econometrics* 92 (1): 1–45.

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 D. E. Weinstein. 2002. "Bones, Bombs, and Break Points: The Geography of Economic Activity." *American Economic Review* 92 (5): 1269–289.

de Kadt, D., and H. A. Larreguy. 2018. "Agents of the Regime? Traditional Leaders and Electoral Politics in South Africa." *Journal of Politics* 80 (2): 382–99.

de Kadt, D., and M. Sands. 2016. "Segregation Drives Racial Voting: New Evidence from South Africa." Working Paper, Department of Political Science, MIT, Department of Government, Harvard University, Cambridge, MA.

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." *Economic Journal* 127 (604): 1906–39.

Duranton, G. 2007. "Urban Evolutions: The Fast, the Slow, and the Still." *American Economic Review* 97 (1): 197–221.

Eeckhout, J. 2004. "Gibrat's Law for (All) Cities." *American Economic Review* 94 (5): 1429–51.

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

Gabaix, X. 1999. "Zipf's Law for Cities: An Explanation." *Quarterly Journal of Economics* 114 (3): 739–67.

Giraut, F., and C. Vacchiani-Marcuzzo. 2013. *Territories and Urbanisation in South Africa: Atlas and Geo-Historical Information System (DYSTURB)*. Marseille, France: IRD.

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): 93–115.

Harris, J. R., and M. P. Todaro. 1970. "Migration, Unemployment and Development: A Two-Sector Analysis." *American Economic Review* 60 (1): 126–42.

Henderson, J. V. 1974. "The Sizes and Types of Cities." *American Economic Review* 64 (4): 640–56.

———. 2005. "Urbanization and Growth." In *Handbook of Economic Growth*, vol. 1B, edited by P. Aghion and S. Durlau, 1543–91. Amsterdam: North-Holland.

Holmes, T. J., and S. Lee. 2010. "Cities as Six-by-Six-Mile Squares: Zipf's Law?" In *Agglomeration Economics*, edited by E. L. Glaeser, 105–131. Chicago: University of Chicago Press.

Imbert, C., and J. Papp. 2018. "Short-Term Migration and Rural Workfare Programs: Evidence from India." *Journal of the European Economic Association*. https://doi.org/10.1093/jeea/jvz009.

Kremer, M., and S. Watt. 2006. "The Globalization of Household Production." Prepared for the Blue Sky Conference, *Weatherhead Center For International Affairs*, Harvard University, Cambridge, MA.

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

Lewis, W. A. 1954. "Economic Development with Unlimited Supplies of Labour." *Manchester School* 22 (2): 139–91.

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

Michaels, G., F. Rauch, and S. Redding. 2012. "Urbanization and Structural Transformation." *Quarterly Journal of Economics* 127 (2): 535–86.

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*. London: Pall Mall Press.

Ogura, M. 1996. "Urbanization and Apartheid in South Africa: Influx Controls and Their Abolition." *Developing Economies* 34 (4): 402–23.

Peri, G. 2012. "The Effect of Immigration on Productivity: Evidence From U.S. States." *Review of Economics and Statistics* 94 (1): 348–58.

Peters, M. 2017. "Refugees and Local Agglomeration: Evidence from Germany's Post-War Population Expulsions." Unpublished Manuscript.

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

Ranis, G., and J. C. Fei. 1961. "A Theory of Economic Development." *American Economic Review* 51 (4): 533–65.

Rauch, F. 2013. "Cities as Spatial Clusters." *Journal of Economic Geography* 4 (14): 759–73.

———. 2016. "The Geometry of the Distance Coefficient in Gravity Equations in International Trade." *Review of International Economics* 5 (24): 1167–77.

Rossi-Hansberg, E., and M. L. Wright. 2007. "Urban Structure and Growth." *Review of Economic Studies* 74 (2): 597–624.

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.

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*. Johannesburg: Ravan Press.

Turok, I. 2012. "Urbanisation and Development in South Africa: Economic Imperatives, Spatial Distortions and Strategic Responses." Urbanization and Emerging Population Issues Working Paper No. 8, International Institute for Environment and Development, London, UK.

United Nations. 2014. *World Urbanization Prospects: The 2014 Revision, Highlights*. New York: United Nations, Department of Economic and Social Affairs, Population Division.

Zipf, G. K. 1949. *Human Behavior and the Principle of Least Effort*. Cambridge: Addison-Wesley Press: Cambridge Massachusetts.