The persistence of apartheid regional wage disparities in South Africa*

Gibson Mudiriza[a] and Lawrence Edwards[b]

September 2019

Abstract

Despite the ending of apartheid, regional wage disparities remain prevalent in South Africa with the former homelands characterised by persistently low wages and incomes. In this paper, we use a new economic geography (NEG) model to estimate the extent to which the persistence in apartheid regional wage disparities are an outcome of economic forces such as access to markets. We estimate a structural wage equation derived directly from the NEG theory for 354 regions over the period 1996 to 2011. We find strong support for the NEG model in explaining regional wage disparities, but only after we augment the model to include regional specific factors such as human capital, mineral resources, local climatic conditions, industrial structure and unemployment. We also find persistently adverse wage effects associated with the apartheid policies, with wages substantially lower in the former homeland areas, even after controlling for NEG and other region-specific characteristics. Average wages of workers in homeland areas were 22% lower than predicted in 1996, with this gap rising to 39% in 2011. These findings show that the reintegration of homeland areas into South Africa and continuous implementation of regional policies since the end of apartheid were not sufficient to reduce the homeland wage gap.

Keywords: Economic geography; Labour market, Wage differentials, Regional economic activity, Economic development
JEL Codes: F12, F16, J31, R11, 010

1 Introduction

Racial discrimination was at the heart of the apartheid system that was instituted in 1948. The system implemented a number of policies aimed at promoting segregation and separate development along racial lines (Simkins 1983, 2011). One of these policies was the homeland policy that dispossessed about 3.5 million black South Africans of their land and forcefully relocated them into ten "homeland" areas, according to their ethnic groups (The Presidency,

* We thank participants of the CSAE 2018 Conference for comments on an earlier draft. We are also grateful for the funding provided by African Economic Research Consortium (AERC), the South African National Research Foundation (CSUR 14080989186, Unique Grant No: 93648) and the Carnegie Foundation to carry out the research. All errors and omissions remain our own.

[a] Corresponding author: Post-doctoral Research Fellow, School of Economics, University of Cape Town, South Africa, gmudiriza@gmail.com.

[b] Professor, School of Economics, University of Cape Town, South Africa, Lawrence.edwards@uct.ac.za.

2014; Abel, 2015)[1]. Apart from overpopulation, a key feature of these homelands was their geographical remoteness, far from major markets, road networks, international airports and harbours. In addition, these homeland areas were allocated limited resources leading to provision of inferior services compared to other areas were whites lived.

Following the advent of democracy in 1994, several redressal measures including black economic empowerment, provision of social grants, better housing and education were also implemented, to uplift the social and economic outcomes for previously disadvantaged race groups (Jensen and Zenker, 2015). In addition, while former homeland areas we reintegration into South Africa, regional policies to level out regional imbalances were also implemented. Among these policies is the National Spatial Development Framework (NSDF) of 1995, the Spatial Development Initiatives (SDIs) of 1996, the Regional Industrial Development Strategy (DTI, 2006), as well as the National Spatial Development Perspective (NSDP, 2003; 2006). In addition, changes to labour market policies were also implemented through the Labour Relations Act (LRA) of 1995 and the Basic Conditions of Employment Act (BCEA) of 1997 to advance the interests of black workers, who had been marginalised for many years.

Despite these efforts, former homeland areas remain underdeveloped with high levels of unemployment, poverty, deprivation, low wages and slow industrialisation (Noble and Wright, 2013; Noble, Zembe and Wright, 2014; Frame et al., 2016; von Fintel, 2017, 2018). Consequently, more than two decades after the ending of apartheid, apartheid-era spatial inequalities (disparities) continue to be reproduced (von Fintel & Moses, 2017; Burger et al. 2017). This is evident in Figure 1, that displays the spatial distribution of average monthly wages across regions in South Africa, where the darker colours are regions with high average wages and lighter colours low average wages in 1996 and 2011. The figure clearly shows that wages vary significantly across regions in South Africa and this variation tends to persist over time. The figure further indicates that wages are persistently low in former homeland areas (areas with blue colour boundaries).

This evidence is also confirmed in Table 1 that provides the summary statistics of average wages for homeland and non-homeland areas in 1996 and 2011. The table shows that, while average monthly wages have increased for both homeland and non-homeland areas, average wages of workers in homeland areas were 1.1 lower in 1996 compared to non-homeland areas and the gap rose to 1.3 in 2011. Based on this evidence, regional wage disparities, in particular, the homeland wage gap has not only persisted but has even increased over time. In spite of this evidence, there is remarkably little research that tries to identify and unpack the causes of the persistence of apartheid regional wage disparities.

Currently, the bulk of the existing empirical literature on wage disparities (inequality) in South Africa focus at individual-level analysis (Bhorat and Goga, 2013; Ntuli and Kwenda, 2014; Kwenda and Ntuli, 2018; Wittenberg, 2017a; b; Finn and Leibbrandt, 2018; Mosomi, 2019). A few studies focus on wage disparities at the regional level (Kingdon and Knight, 2006; Magruder, 2012 and von Fintel, 2018). While this research has certainly advanced our understanding of wage disparities in the country, its major drawback is that it is primarily descriptive, with few studies properly controlling for location and the spatial determinants of

[1] Apart from land dispossessed, blacks were also dispossessed of other means of livelihoods such as livestock. The rights to own, rent or transfer property for blacks was limited to these ten homeland areas only.

wages, particularly the importance of access to markets. Location consists of first and second nature geography, where first nature geography is concerned with physical geography factors of minerals resources, climate, agricultural land, presence of natural harbours (Roos, 2005) and in some studies human capital, unemployment and industrial structure (Bosker, 2008). Second nature geography deals with factors related to the location of economic agents relative to one another in space (Redding, 2009)[2].

At best, studies in South Africa control for location by including provincial dummies (or rural-urban dummy), for example, in the individual-level studies estimating mincer wage regressions (Bhorat and Goga, 2013; Ntuli and Kwenda, 2014), while the regional-level studies (Kingdon and Knight, 2006; Magruder, 2012 and von Fintel, 2018) control only for first nature geography factors to the neglect of second nature geography factors. More importantly, the above-mentioned regional level studies are based on reduced-form specifications that are not tightly guided by theory. These studies, therefore, fail to provide deeper insights into theoretically consistent mechanisms driving regional wage disparities.

In this paper, we use the Helpman (1998) model that is derived directly from the new economic geography (NEG) theory to explain the presence and persistence of apartheid regional wage disparities in South Africa over the period 1996-2011. The NEG theory emphasises how second nature geography factors, such as access to markets determine variation in regional wage levels (Krugman, 1991; Fujita & Mori, 2005). In this framework, the interaction of transport costs, increasing returns to scale and consumers' love of variety generate agglomeration forces that encourage economic agents (firms and workers/consumers) to concentrate in regions with greater access to markets, leading to regional wage divergence.

NEG models have increasingly been used to explaining regional wage disparities within many countries (Mion, 2004; Brakman et al., 2004; Kosfeld & Eckey, 2010; Hanson, 2005; Fallah et al., 2011; Paredes, 2015), but they have in general performed poorly and are sensitive to specifications and time period. One potential explanation is that the models do not incorporate first nature geography factors, such as natural endowments, that have played a prominent role in these countries' development. This is likely to hold for South Africa, whose spatial development is closely associated with the exploitation of natural resources such as gold and diamonds (Wilson, 2009).

Accordingly, we extend the NEG model to include potential first nature geography factors such as mineral resources, local climatic conditions, human capital, local unemployment and industrial structure. We apply the model to wage data for 354 regions that we obtain from the Population Censuses for 1996, 2001 and 2011. This augmented NEG model is then used to estimate the conditional wage gap in the former homelands in each year.

We find support for an augmented NEG model in explaining regional wage disparities in South Africa in all periods. Wage disparities in South Africa are thus well explained by a combination of market access forces plus location-specific first nature geography conditions. However, we also find persistently adverse wage effects associated with the apartheid policies, with wages substantially lower in the former homeland areas. The results reveal that on average

[2] Second nature geography, hence, relative geography is captured by many factors among them access to markets, which reflects the degree of economic interactions between firms, workers, and customers in space, which in turn promotes agglomeration of economic activities in central regions driven by transport costs, increasing returns to scale and consumer lover of variety.

wages of workers in homeland areas were 22% lower than predicted in 1996, with this gap rising to 39% in 2011[3]. Our findings show that the ending of apartheid was not sufficient to reduce the homeland wage gap despite the reintegration of homeland areas into South Africa and the continuous implementation of regional policies since 1994. Overall the study shows that first and second geography factors together with long-gone apartheid-era policies are key factors in explaining the persistence of apartheid regional wage disparities.

The study has three broader contributions. Firstly, the study presents a theoretically consistent approach to identifying the factors influencing regional wage disparities as well as the homeland wage gap in South Africa. Secondly, the study provides new insights into the sources of wage variation and wage disparities in South Africa, namely the influence of economic geography. Thirdly, it contributes to the emerging literature on the persistent effects of historical events (apartheid-era policies) on current regional economic development. Fourthly, it contributes to the NEG empirical literature by estimating a NEG model in an emerging country, whose characteristics might affect the empirical performance of the model. Finally, the study contributes toward the practical policy debate in South Africa, where mitigating regional imbalances and ensuring regional equalisation of living standards is regarded as a fundamental objective of the post-apartheid government. This objective can be aided by regional policy initiatives that are well informed on the sources of persistent regional wage disparities.

2 The South African context

South Africa is known for its extreme inequality. This inequality which is reflected not only at the individual level but also at the regional level has its roots in the apartheid system. However, its origin goes back as far as the 17th and 18th centuries when pre-existing geographic differences in access to waterways, climates, and natural resources played a vital role in the development of regional economies in the country. For instance, development of the port cities of Cape Town and Durban was driven by their close proximity to waterways that gave them an important role in the country as trading posts on the shipping route between Western Europe and Asia (Gelb, 2004; Bosker & Krugell, 2008). The trade enhanced development was reinforced by favourable climatic conditions that encouraged settlement of European colonisers who established institutions to suit their stay in these port cities (Krugell & Naude, 2003)[4].

The discovery of minerals shifted attention to the development of inland regions of Gauteng (Johannesburg and Pretoria). Mining activities and resulting infrastructure established to support the mining industry generated strong economic forces that promoted rapid industrialisation, urbanization and massive migration of workers to support the growing industry in and around Gauteng (Turok, 2012). This allowed Gauteng to develop into the large urban agglomeration that we see today (Beavon, 2001) that accounts for the bulk of the country's economic activities (Stats SA, 2014).

[3] The removal of apartheid government regional industrial development programme incentives in 1991 might explain the increase in the homeland wage gap. While these incentives were aimed at improving industrial development in regions near homelands and some nodes within the homelands, their removal in 1991 led to de-industrialisation, which in turn led to job losses.

[4] This outcome is supported by Acemoglu, Johnson, & Robinson. (2002), who note that colonies with favourable climatic conditions for European settlements are richer than other countries or regions.

The resulting unequal regional economic development process was reinforced by the apartheid system that was instituted in 1948 and consolidated the racial discrimination that was started under the Dutch and British colonial rule. Using the Land Act (1913), the Dutch and British colonial rule limited acquisition of land by the black majority to the native reserve areas that constituted 7% of the country's area and this area was increased to 13% by the Land Act (1936). Building on these Acts, the apartheid system further introduced the Group Areas Act (1950), the Bantu Authorities Act (1951) and the Bantu Resettlement Act (1954). These Acts were used to settle people according to their racial classification (white, black, coloured and indian), as codified in the Population Registration Act (1950). The subsequent Promotion of Bantu Self-Government Act (1959) and the Black Homeland Citizenship Act (1970) were also introduced to deny blacks their rights as citizens of South Africa by transforming native reserves into fully-fledged independent and self-governed homeland (or Bantustans) states for the country's black population (King and McCusker, 2007). Taken together, these legislations formed the basis of the homeland policy that led to the creation of ten homeland areas namely, Transkei, Bophuthatswana, Ciskei, Venda, Gazankulu, KaNgwane, KwaNdebele, KwaZulu, Lebowa, and QwaQwa.

As displayed in Figure 2, a key feature of these homelands is their location in periphery areas. They thus suffered problems associated with geographical remoteness including being far from major road networks, international airports, harbours, and major markets. The remoteness of the homelands acted as a barrier to investment, leading to high unemployment and when jobs were available, they were inferior, insecure and lowly paying compared to other areas. In these homelands, blacks were economical dependent on jobs created in areas reserved for whites, where a pass was needed for them to work. Those who remained in homelands, were heavily dependent on farming, despite the limited and poor agricultural land in homelands. The remoteness of the homelands was also pivotal to the grand apartheid vision of separate development along racial lines. The homeland areas were allocated limited resources, leading to the provision of services (such as education, health, water and electricity) markedly inferior compared to those provided in areas where whites lived and worked.

With the advent of democracy in 1994, the elected government embarked on a massive re-demarcation exercise of the country's administrative boundaries with the goal of dissolving the racial based spatial layout created by the apartheid system. This led to the creation of local authorities constitutionally responsible for the development of their areas (Bosker & Krugell, 2008). Furthermore, as already discussed the government introduced several policy initiatives aimed at levelling out regional economic differences to promote catch-up of peripheral regions. Despite these concerted efforts, existing evidence suggests that no major shifts have taken place in the South African spatial landscape (Nel and Rogerson, 2009). Consequently, many of the spatial patterns from the past persist to the present day. This paper provides an explanation of this persistence, focusing on regional wage disparities.

3 Theoretical model

To explain the persistence of apartheid regional wage disparities, we draw on the Helpman (1998) model that is derived directly from the NEG literature, as pioneered by Krugman (1991).

While Krugman (1991) developed a model that is normally used for cross country analysis, Helpman (1998) modified the model to allow for within-country analysis.

In the Helpman (1998) model a representative consumer in region r has a Cobb-Douglas utility function and consumes two bundles of goods; a homogeneous non-tradable housing service and differentiated tradable manufactured good. Housing stocks are produced in a perfectly competitive market and the supply is fixed in each region (Hanson, 2005; De Bruyne, 2010). Depending on supply and demand, the prices for housing services tend to be high in densely populated areas and low in sparsely populated areas (Kosfeld and Eckey, 2010). The manufactured good is produced in a monopolistic competitive market by a firm operating under increasing returns to scale. The firm uses mobile labour as the only factor of production. Manufactured good is traded across regions at a cost modelled in the form of an "iceberg" transport cost (τ). The implication of this assumption is that only a fraction of the shipped good arrives at the final destination. Consumers spend a fraction of their income on manufactured good (μ). The manufactured good is treated as a composite of differentiated varieties with the consumption of each variety determined by its relative price and the elasticity of substitution of the manufactured varieties ($\sigma > 1$).

In deciding where to locate in space, firms and consumers choose locations that maximise profits and utility, respectively. On one hand, manufacturing firms prefer to concentrate production in regions with greater access to markets to minimise transport costs and benefit from large-scale production (Redding, 2013). On the other hand, because of the need to avoid paying transportation costs in importing manufactured goods from other regions, consumers also favour locating in regions with greater access to large markets (Redding, 2010). These market effects, which we also refer to agglomeration forces, drive producers and consumers to concentrate in regions with greater access to large markets. This concentration usually leads to regional wage disparities, with lower nominal wages in remote areas with poor access to markets and higher nominal wages in areas closer to larger markets. However, the market effects are offset by crowding effects, which we also refer to as dispersion forces due to rising housing (land) prices as well as increasing competition in the goods and labour markets (Kosfeld and Eckey, 2010). The resulting wage distribution patterns depend on the tension between agglomeration and dispersion forces, with persistent regional wage disparities evident when agglomeration forces are domain.

Under these circumstances, the long-run spatial equilibrium of the economy can be summarised by five simultaneous equations related to real wages, housing expenditure, income, prices, and nominal wages. Using these equilibrium equations and assuming real wage equalisation across locations, Hanson (1998) derives an empirically testable wage equation[5] from the Helpman (1998) model given by:

$$w_r = \left[\sum_i^R Y_i^{\frac{\sigma(\mu-1)+1}{\mu}} H_i^{\frac{(1-\mu)(\sigma-1)}{\mu}} w_i^{\frac{\sigma-1}{\mu}} e^{-\tau(\sigma-1)d_{ri}} \right]^{\frac{1}{\sigma}} \tag{1}$$

[5] Redding & Venables (2004) derive a testable wage equation using the estimates of a gravity trade model based on bilateral trade data. The equation gives a theory-based market potential function consisting of two indices: market access and supplier access indices. While this strategy has largely been used for cross country analysis, it is less appealing for within country analysis because of the unavailability of regional-level trade data in most countries, and also because of its underlying assumption of labour immobility (labour is less mobile cross country but more mobile within country).

Equation (1) is the wage equation and generally referred to as the *"Helpman-Hanson model"*. The equation presents the average wage (w_r) that firms in region r are willing to pay their workers. This wage is a function of market potential, a measure of a region's access to markets, given by a distance (d_{ri}) weighted function of income (Y_i), housing supply (H_i) and wages (w_i) in all other regions. According to equation (1), average wages are higher in regions that are nearby other regions with higher levels of income, wages and housing stocks.

In estimating equation (1), the implications of the Helpman-Hanson model are captured by three structural parameters, namely, elasticity of substitution among manufactured varieties (σ), transport costs (τ) and share of income devoted to consumption of manufactured goods (μ) or housing services ($1-\mu$). The model is valid when these parameters satisfy the following constraints: $\sigma > 1$, $0 \leq \mu \leq 1$ and $\tau \geq 0$. In addition, two relations between these parameters should also be satisfied. The first is the market power condition that holds when $\sigma/(\sigma - 1)$ >1. The higher the ratio, the higher the market power of a firm.

The second is the no black hole condition that holds when $\sigma(\mu-1) < 1$. This condition is critical for the overall validation of the Helpman-Hanson model as it captures the tension between agglomeration and dispersion forces that are key in the determination of regional wage levels. When the condition holds, it implies that in the determination of regional wages, agglomeration and dispersion forces are interacting in a way that is consistent with the NEG. However, when the condition does not hold, it implies that the NEG forces are interacting in a way that is inconsistent with the observed distribution of wages across regions. Rather, the distribution of wages is determined by exogenous first nature geography factors (De Arcangelis and Mion, 2002).

4 Empirical Wage Equation

Our empirical wage equation is derived directly from the theoretical wage equation (1), discussed in the previous section. To estimate this equation, a transport cost function must be defined. While Hanson (2005) captures transport costs with an exponential distance decay function ($e^{-\tau d_{ir}}$), in this paper, we use a distance power function ($d_{ir}^{-\tau}$). As argued by Mion (2004), the distance power function is empirically appealing because of its strong theoretical foundations within the gravity trade model that provide the building blocks of NEG models[6]. Inserting the distance power function, taking logs, and imposing restrictions to equation (1), we derive the following estimation model:

$$\log(w_r) = \alpha_0 + \alpha_1 log\left[\sum_{i=1} Y_i^{\frac{1}{\alpha_1}-\alpha_2} H_i^{\frac{1}{\alpha_1}-1-\alpha_2} w_i^{\alpha_2} d_{ri}^{\alpha_3}\right] + \varepsilon_r \qquad (2)$$

The dependent variable, $\log(w_r)$ is log of average wage in region r. α_0 is a function of constants (σ, μ, τ, f) and the equilibrium real wage, ω. Y_i, H_i and w_i capture income, housing stock, and average wage in region i, respectively. d_{ir} is the distance between region i and r that we use to proxy transport costs and ε_r is the regression error term.

[6] For robustness checks, we also estimate the model using the exponential distance decay function. While the magnitude of the estimates differs between the functions, we derive qualitatively similar conclusions from the two functions. Hence, we choose not to report estimates based on the exponential distance function.

The importance of market potential in explaining the presence and persistence of regional wage disparities is confirmed by a positive and significant α_1 coefficient ($\alpha_1 > 0$). Estimation of the other reduced-form coefficients, α_2, and α_3 enable us to derive all the structural parameters (σ, μ, τ) of the Helpman-Hanson model. A positive wage coefficient ($\alpha_2 > 0$) and a negative distance coefficient ($\alpha_3 < 0$) are expected[7].

A major limitation of model (2) is that it explains regional wage disparities based on second nature geography forces only. However, as already discussed, first nature geography factors like mineral resources played a significant role in the development of regional economies and might be key determinants of regional wage levels in the country. Thus, to fully account for location we extend equation (2) to control for first nature geography factors:

$$\log(w_r) = \alpha_0 + \alpha_1 log\left[\sum_{i=1} Y_i^{\frac{1}{\alpha_1}-\alpha_2} H_i^{\frac{1}{\alpha_1}-1-\alpha_2} w_i^{\alpha_2} d_{ri}^{\alpha_3}\right] + \sum_{n=1}^{N} \beta_n X_{rn} + \varepsilon_r \qquad (3)$$

where in addition to the variables defined in equation (2), X_{rn} is a vector of first nature geography controls that includes measures of mineral resources, temperature, rainfall, human capital, unemployment rate and industrial structure. β_n is the vector of corresponding coefficients[8]. Of these measures, we expect human capital, mineral resources and share of manufacturing workers to have a positive effect on wages, while a negative effect is expected for local unemployment rate and share of agricultural workers. For temperature and rainfall, we expect either a positive or negative effect.

Having controlled for the effects of first and second nature geography factors, we use the extended model to determine the homeland wage gap by examining the effect of a unique historical event, the establishment of apartheid homeland areas. While the homeland areas were reintegrated into South Africa more than two decades ago, we argue that the legacy of the homeland policy continues to influence economic development across the country. Our claim finds support from a growing body of research that concludes that distinctive historical events such as colonisation and slave trade have a long-lasting effect and continue to influence economic development to the present day (Banerjee and Iyer, 2005; Nunn, 2009; Acemoglu & Dell, 2010; Dell, 2010; Michalopoulos and Papaioannou, 2013, 2014; Bandyopadhyay and Green, 2016; Angeles and Elizalde, 2017). The claim further finds support from a small emerging literature that examines the effects of the apartheid-era homeland policy on inter-ethnic trust (Kerby, 2014), social capital (Abel 2015), political outcomes (De Kadt and Larreguy, 2014) and the evolution of spatial income disparities (Bastos and Bottan, 2014) in South Africa.

Thus, we examine the effects of the apartheid homeland policy on the presence and persistence of regional wage disparities by augmenting equation (3) with a homeland status indicator (HS_r) as follows:

[7] We derive the Helpman-Hanson model structural parameters as follows: $\sigma = 1/\alpha_1$, $\mu = (1 - \alpha_1)/\alpha_1\alpha_2$ and $\tau = \alpha_1\alpha_3/(\alpha_1 - 1)$. From these parameters, two additional equilibrium conditions given by $\sigma/(\sigma - 1)$ – price-marginal cost ratio and $\sigma(1 - \mu)$ – no black hole condition, are also derived. Table 12 in appendix provide a summary of the conditions that the structural parameters need to satisfy for the Helpman-Hanson model to be consistent for the case of South Africa.

[8] A description of the variables is provided in the data section, while the motivation for their inclusion is given in the empirical results section.

$$\log(w_r) = \alpha_0 + \alpha_1 log\left[\sum_{i=1} Y_i^{\frac{1}{\alpha_1}-\alpha_2} H_i^{\frac{1}{\alpha_1}-1-\alpha_2} w_i^{\alpha_2} d_{ri}^{\alpha_3}\right] + \sum_{n=1}^{N} \beta_n X_{rn} + \delta HS_r + \varepsilon_r \qquad (4)$$

HS_r is a ratio ranging between 0 and 1, with 1 indicating that a region falls completely in a homeland area, while 0 indicates that a region fall in a non-homeland area (a detailed explanation of the construction of the variable is provided in the data section). Our coefficient of interest is δ which captures the conditional wage gap of the former homelands for 1996, 2001 and 2011. By estimating equation (4), we are able not only to determine the effects of the long-gone apartheid homeland policy on local wage levels but also the size of the apartheid homeland wage gap. In addition, by estimating the model for three time periods, we can determine the evolution of apartheid homeland wage gap. The expectation is that this should disappear or decline with the ending of apartheid.

5 Data and Descriptive Evidence

For our empirical analysis, we make use of two different data sources. We rely on the population census data collected by Statistics South Africa (Stats SA) and the climate data produced by Harris et al. (2014) for the Climatic Research Unit (CRU) at the University of East Anglia. Using this data, we construct a unique geographically consistent dataset for 354 regions for the years 1996, 2001 and 2011.

Data

Analysis using population censuses data constitutes an important contribution to the literature of the South African spatial economy. The censuses contain a rich set of information on demographics and labour market (including employment, industry, and income), among others. The greatest advantage of the censuses over other household surveys in South Africa is its large sample size and availability of labour market information at various geographical levels[9]. However, a major challenge with these geographical levels is their inconsistencies over time, which makes longitudinal analysis difficult. These inconsistencies are a result of government's ongoing re-demarcation of the country's administrative boundaries aimed at dissolving the apartheid-era racial based spatial planning.

To address this challenge, we use ArcGIS to overlay 2011 sub-place boundaries onto 1996/2001 magisterial district boundaries[10]. Population values from 2011 census at sub-place level are then re-aggregated to 1996/2001 magisterial district level leading to a geographically consistent dataset containing 354 magisterial districts that we use as our unit of analysis. Magisterial districts are an ideal unit of analysis as they closely define the location of cities and towns where most economic activities take place in South Africa. Magisterial districts are also ideal as they do not vary in size as much as other possible geographical units and are also sufficiently large to allow for the spatial analysis of various socioeconomic outcomes. Finally,

[9] There are, however, arguments that the censuses are not as accurate as the household surveys in measuring labour market outcomes, especially labour market income.

[10] Our choice of using 2011 sub-place units is motivated by existing literature that acknowledges that assigning population values from a smaller geographical unit minimises the error associated with the highly restrictive assumption underlying the areal-weighting interpolation technique of homogeneously distributed population across space (Maantay et al, 2008).

magisterial districts have commonly been used in the literature as the unit for analysis of local labour markets in South Africa (see, Naude and Krugell, 2003; 2005; Krugell, 2005; Bosker and Krugell, 2008; Magruder, 2012 among others).

For the estimation of equation 2, our main variables of interest are wage per worker (w_r) and market potential that is a function of income (Y_r), housing stocks (H_r), wage per worker (w_r) and distance (d_{ri}). We use income per worker to proxy for wage per worker for each region. However, in using census income information, three main challenges are worth noting. Firstly, the census income information is from many different sources. While acknowledging this challenge, income per worker is a good proxy for wage per worker given that labour income (wages) contributes the largest share to total income of employed individuals in South Africa. Secondly, the census income is bracketed with an open-ended top bracket. We address this challenge by assigning the midpoint of each bracket to everyone in that bracket and setting the midpoint of the open-ended bracket to two times the lower bound of the highest bracket. Finally, the census income has a high proportion of reported zero and missing information. Whereas this challenge is of great concern when using the whole sample, narrowing down to employed individuals who are the focus of this study significantly reduces the proportion of individuals with zero or missing income. A more detailed discussion of these challenges and how we addressed them is provided in the appendix. For convenience, from hereon we refer to regional income per worker as regional wage.

For the different components of the market potential index, we use the sum of personal income to proxy each region's total income (Y_r), a measure of market size. We further use the total number of rooms in a dwelling as a proxy for total housing stocks in a region (H_r). Finally, we calculate distance (d_{ri}), as the great-circle distance (in kilometres) between the geographical centres (centroids latitude and longitude) of region i and r.

To estimate equation 3 and 4, we include specific first nature geography factors. We incorporate the share of workers (in total working-age population) with tertiary education to capture the effects of regional differences in skilled workers (human capital)[11]. To capture the effects of differences in local labour market conditions, we include regional unemployment rate. This measure controls for several local labour market conditions including variations in regional wage rigidities and migration. To account for differences in regional sectoral composition, we add the share of agricultural and manufacturing sector workers in each region. To account for the influence of mineral resource endowments, we include the share of workers in the mining sector in each region. This measure enables us to control for the influence of mining activities that have played a prominent role in South Africa's spatial development.

To further control for first nature geography factors, we draw on climate data for temperature and rainfall. This data combines information from more than 4000 weather stations distributed around the world with satellite-based information. We use this data that is provided at a fine spatial resolution (0.5x0.5-degree grids) to obtain estimates of yearly (1996, 2001 and 2011) average temperature and rainfall by region[12].

[11] While this definition of skilled workers is subject to debate, we check the robustness of our definition by using the matric qualification (high school end-exam) as a cut off. Regardless of the definition used the importance of human capital (skilled workers) remain evident in our analysis.

[12] To derive these variables, we combine the climate data with the centroids (latitudes and longitudes) information of each region (magisterial district) obtained from the shapefile provided by Stats SA. Using the resulting data,

Finally, to account for the effects of the apartheid-era homeland policy, we incorporate a homeland status indicator calculated as the share of each region's area that falls in former homeland areas. To derive the indicator, we use ArcGIS to overlay magisterial district boundaries to former homeland boundaries. From the resulting mapping, we use areal-weighting interpolation technique to derive a ratio based on the area of each magisterial district that fall in a given homeland area. The ratio ranges between 0 and 1, with 1 indicating that a region falls completely in a homeland area, while 0 indicates that a region falls in a non-homeland area.

Descriptive Evidence

In Table 2, we present the summary statistics on our key variables for 1996 and 2011. Looking at the statistics, we see that on average, regional wage, total income, market potential, human capital, population and housing stocks have increased over time, while regional unemployment rate has decreased[13]. However, looking at the min and max statistics of these variables, we see that the regional averages mask significant disparities across regions and these differences persist over time. The homeland status variable shows that of the 354 regions, about 30% of these regions fall (fully or partially) in former homeland areas.

To give an indication of the association between our key variables, Table 3 presents some simple correlations between the variables. We focus our attention on column 2 that shows the pairwise correlation between our dependent variable (regional wage) and other variables. The results indicate that wages are negatively correlated with homeland status (-0.33) and unemployment rate, while they are positively correlated with market potential, human capital, housing stocks, population, and income. In summary, the descriptive statistics support the theoretical relationship between regional wages and the control variables.

6 Empirical results

We present four sets of results commencing with the results of the baseline model (2), then results for our extended model (3) and (4) respectively. Lastly, several results from various robustness tests are presented. Given the nonlinearity of equation (2) – (4) we follow existing literature (Roos, 2001; Brakman et al. 2004; Hanson, 2005) and estimate the models using nonlinear least-squares (NLS)[14]. All standard errors have been corrected for heteroscedasticity using a generalization of the White method (see White, 1980)[15].

Baseline results of the Helpman-Hanson model

Table 4 presents the estimates of the baseline model (equation 2) that includes second nature geography, proxied by market potential as the only explanatory variable. Column (1) shows estimates for 1996, column (2) for 2001, and column (3) for 2011. The coefficients for

we use QGIS zonal statistics tool to extract yearly average temperature and rainfall data for each magisterial district for the years 1996, 2001 and 2011.

[13] The climate variables - rainfall and temperature show little change over the time period.

[14] The main advantage of the NLS method is that it allows estimation of the nonlinear relationship between regional wages and market potential without having to linearise the model (Cieślik and Rokicki, 2016). The method also takes into account the constraints due to the links between the model parameters.

[15] After estimating the models, we use the White test to test the null hypothesis of homoskedasticity standard errors (White, 1980). The homoskedasticity assumption is rejected for all models.

α_1 to α_3 are all statistically significant, with signs consistent with theoretical expectations. The estimated coefficient of market potential (α_1) is positive and significant in all the years. A 10% increase in market potential is associated with a 1.13%, 0.98% and 1.7% increase in regional wages in 1996, 2001, and 2011, respectively. Overall, the results indicate that market potential differentials explain between 42% and 49% of the variation in regional wages.

Looking within the market potential indicator, the estimated coefficient on income per worker (α_2) is positive and statistically significant, and on distance (α_3) is negative and statistically significant in all the years[16]. The negative distance coefficient shows that as a region's distance to consumer markets increase, its wage levels decrease. Thus, remote regions in South Africa face a market access penalty that lowers their wage levels. This finding is consistent with theoretical expectations, as well as empirical findings from other countries such as USA (Hanson, 2005), Italy (Mion (2004), Brazil (Fally et al. 2010) and China (Moreno-Monroy, 2011).

To assess the validity of the model we also look at the implied structural parameters. The implied value of σ is greater than 1, as required by the NEG model and ranges from 5.9 to 10.2. The value of σ suggest that firms across regions in South Africa are operating under increasing returns to scale, enjoying mark-ups (given by $\sigma/(\sigma - 1)$) of 10.9 (2001) and 20.4 (2011) percent. These mark-ups are quite close to those of other studies from South Africa using industry and firm data. For example, Zalk (2014) finds mark-ups of between 10 and 20 percent, while Aghion et al. (2008) find a mark-up of 23.3 percent for the manufacturing sector in South Africa. However, the estimates for σ vary from what is reported in other countries. For example, they are higher than the 4.9 to 7.6 range reported for the US (Hanson, 2005) and lower than the 41.1 to 46.1 reported for Chile (Paredes, 2015).

The implied value of μ, the share of income devoted to manufactured goods is statistically significant and satisfies the restriction $0 < \mu < 1$, suggested by theory. However, with estimates of μ ranging from 0.85 to 0.88, these values imply that about 15% $(1 - \mu)$ of total household income is spent on housing services. As in other studies (Hanson, 2005; Mion, 2005), these values seem to be an overestimation of the share of income devoted to manufactured goods. According to Stats SA, about 32 percent of total household income is devoted to housing, water, electricity, gas and other fuels, with housing services taking up the largest share (Stats SA, 2012).

The estimate of τ is statistically significant and positive ($\tau > 0$) in all the years. While the positive (τ) value is consistent with findings in the literature (Paredes, 2015; Mion, 2004; Pires, 2006), it is hard to compare it with other studies, due to the sensitivity of τ to the unit of analysis, the transport cost function used and the method of measuring distance.

Finally, we consider the no black-hole condition, which holds when $\sigma\,(1 - \mu) < 1$. Our results show that the no black-hole condition holds for 2011, while it is rejected for 1996 and 2001. The implication of this is that the data are interacting in a way that is inconsistent with the Helpman-Hanson model. The estimated model for 1996 and 2001 therefore do not provide a foundation for fully explaining regional wage disparities and estimating the homeland wage gap.

[16] We expect α_2 given by $\alpha_2 = (\sigma - 1)/\mu$ to be positive as $\sigma > 1$ and $0 < \mu < 1$.

Effects of specific first nature geography factors

To strengthen our empirical findings, we include controls for specific first nature geography factors by estimating equation (3). The results from the estimation are reported in Table 5.

The main insight from the results is that the inclusion of specific first nature geography factors improves the fit of the model significantly. This is shown by increasing values of the adjusted R-squared in all columns. The estimated models now explain between 66% and 76% of the variation in regional wage levels. The results show that the estimates associated with market potential (both reduced-form and structural parameters) remain significant and consistent with theory. However, important changes can be seen in these estimates. For instance, the effect of market potential becomes stronger, with the estimate (α_1), increasing to 0.23 in 1996 and 0.46 in 2011. Thus, failure to control for first nature geography factors result in a downward bias in the effect of market potential. At the same time, the coefficients associated with wages decrease to 2.70 (1996) and 4.22 (2011), while the size of distance coefficient also decreased to -0.79 and -1.37 over the same period.

In line with the changes in the reduced form estimates, significant changes can also be seen in the corresponding structural parameters $(\sigma;\ \mu$ and $\tau)$. Most notably, these changes now provide evidence in support of the Helpman-Hanson model as applied to South Africa. The no black hole condition $(\sigma\,(1 - \mathrm{\mu}) < 1)$ is now satisfied in all the years. The case of South Africa is consistent with the Helpman-Hanson model once we account for the effects of first nature geography factors.

Looking at the specific first nature geography factors, except for rainfall in 2011 and share of manufacturing workers in 2001, all coefficients are statistically significant and have the expected signs. The coefficient of the share of skilled workers is positive, suggesting that a 10% increase in skilled workers is associated with a 30% (1996) and 22% (2011) increase in regional wages. This highlights that regions with larger shares of working-age population with tertiary education tend to have higher average wage levels. The result confirms the findings by Combes et al. (2008) for France and Huang and Chand (2015) for China. Regional unemployment rate is negatively associated with regional wages and a 10% increase in local unemployment rate correspond with an 8% (1996) and 18% (2011) decrease in regional wages. This result provides support for the wage curve theory as is also found for South Africa by Kingdon & Knight (2006), Magruder (2012) and Von Fintel (2015).

The relationship between the share of mining workers and regional wages is positive and statistically significant. On average, a 10% increase in mineral resource endowments is associated with an increase in regional wages of between 3.5% and 8.4%. The share of manufacturing workers has a positive effect on regional wages, suggesting that a 10% increase in manufacturing sector employment is associated with a 7% increase in regional wages. In contrast regions with high shares of agricultural employment have lower wage levels. Furthermore, regional temperature is significant and positively associated with regional wages, while rainfall is statistically significant and negatively related.

Based on the evidence above, we can conclude that differences in second nature geography, proxied by access to markets and specific first nature geography factors play a key role in explaining the presence and persistence of regional wage disparities in South Africa.

This finding is consistent with findings from other emerging economies that are highly endowed with natural resources, such as Chile (see Paredes, 2015).

The effects of the legacy of apartheid-era rule

Having established the effects of access to markets and specific first nature geography factors, we turn our attention to determining the homeland wage gap by examining the effects of the long-gone apartheid-era homeland policy on regional wages. While the ideal strategy to pick the true effect of the legacy of the apartheid homeland policy is to estimate equation (4) pre-1994 and post-1994, this is not possible due to the unavailability of regional data during the apartheid-era rule[17]. Accordingly, we use 1996 as our base year and 2011 as our terminal year. While apartheid policies officially ended in 1994, by 1996 not much had changed to the South African spatial economy as policies aimed at addressing regional imbalances were still in the initial stages of implementation. As a result, 1996 is an ideal base year in examining the effects of the apartheid homeland policy. The results from the estimation of equation (4) are reported in Table 6[18].

Two main findings can be seen. Firstly, the importance of first and second nature geography factors remain evident in all columns after the inclusion of the homeland status indicator. Secondly, the coefficient of homeland status is negative and significant in all the years, suggesting that regions in homeland areas have persistently low wages compared to other regions, even after controlling for their location (differences in first and second nature geography factors). Thirdly, over time, we see that the conditional homeland wage gap increases from 22% in 1996 to 39% in 2011, although much of the increase took place over the 1996-2001 period. This result highlight that the wage gap has not only persisted but has even increased over time. Thus, despite the abolishment of apartheid, the reintegration of homeland areas into South Africa in 1994 and continuous implementation of various regional policies, the legacy of apartheid-era homeland policy continues to negatively influence regional wage levels in South Africa.

Robustness tests

The results above show that the combination of first and second nature geography factors, together with long-gone apartheid-era homeland policy explains the persistence of regional wage disparities in South Africa. One concern with our estimates is that our results might be influenced by the way we define the homeland status indicator, namely the share of each region’s area that falls in former homeland areas. To test the robustness of our results, we re-estimate equation (4) using alternative homeland measures. Specifically, we re-define a region as homeland if: (1) at least 20% of its area falls in a homeland area; (2) its centroid[19].

[17] For instance, 1991 census does not include information for 4 homeland states of Transkei, Bophuthatswana, Venda and Ciskei (TBVC) who were independent from South Africa.

[18] An initial check of the association between regional wages and homeland status in Table 15 in the appendix. The results reveal lower average wages in homelands areas compared to other areas, showing that the unconditional wages in the homelands were 10.5% lower in 1996 and this homeland wage gap rose to 30.2% in 2011. The magnitude of these estimates is remarkably similar to the descriptive estimates revealed in Table 1.

[19] The centroid of a region is defined as the central point (defined by the latitude and longitude) of a given polygon.

falls in a homeland area and (3) its distance from a given homeland boundary equals 0; ≤10; ≤20; ≤30 km; ≤33 km.

The results are reported in Table 7 - 9. The coefficients of the alternative homeland status variables in table 7 and 8 remain negative and statistically significant in all columns, although the magnitude of the coefficients have decreased to 12% and 13% in 1996 as well as 27% and 25% in 2011. However, the main findings of a high and rising homeland wage gap remain evident. The results of the distance-based homeland status measure presented in table 9 show a significantly negative homeland status coefficient whose effect decreases with increasing distance from a given homeland boundary[20]. Beyond 33 km the effect is no longer significant. The results show that the wage levels for regions within 0 km, 10 km, 20 km, 30 km and 33 km of a given homeland boundary are 24.6%, 21.1%, 12.7%, 6.8% and 5.3% lower compared to other regions, respectively. These findings highlight that the apartheid-era homeland policy does not only affect regions in homeland areas but its negative effect spillover to surrounding regions that are within 33 km from a homeland boundary.

The results above confirm that our finding of a persistently negative and significant homeland effect on wage levels is not influenced by how we measure homeland status. However, the cross-section models that we have estimated this far do not allow for proper control for time effects in the determination of the homeland wage gap. Yet, time might have an influence on the effect that apartheid homeland policy has on regional wage levels. Thus, to control for time effects, we pool the data and estimate the following pooled regression models: the baseline model (2) and model (4) controlling for time fixed effects and model (4) controlling for time fixed effects as well as including interaction terms between homeland status and time fixed effects (2001 and 2011 dummies).

The estimation results are reported in table 10. The results show negative and highly significant estimates for the time dummies in all columns. Notwithstanding this evidence, the homeland status variable remains negative and significant after controlling for time fixed effects. The results show a homeland wage gap of about 35% (column 2) and this decrease to 8% once we include the interaction terms (column 3). The significantly negative estimates of the interaction terms confirm that time has an influence on the effect that homeland policy has on wage levels. The results indicate that homeland average wages were 28% lower in 2001 and this gap rose to 43% in 2011. Thus, the core findings of a high and rising homeland wage gap remain evident. Furthermore, the importance of market potential and first nature geography controls also remain evident in all columns.

Apart from the homeland issue, another concern with our results is the bias arising from two potential sources of reverse causality. First, regional wage is present in both (left and right) sides of the equations. Secondly, regional wage is also a component of regional income (Y_i). The standard approach to deal with the problem of reverse causality is to use the instrumental variable (IV) technique[21]. However, in this paper, we utilise two alternative strategies to check the robustness of our results to potential bias due to reverse causality.

[20] In table 8, we only present results for 2011, the more recent period since results for 1996 and 2001 did not add any additional insights.

[21] However, it is difficult to come up with reliable instrumental variables, given that most economic variables are also endogenous (Redding, 2010). In addition, established evidence acknowledges that the nonlinearity form of

First, we follow López-Rodríguez & Faíña (2006) and proxy each region's market size with regional population instead of regional income. This reduces the possible correlation between market potential and the error term, as regional population is strongly correlated with regional income, but less strongly correlated with regional wages (see Table 14). The results are reported in Table 11 and the bulk of the estimates are not very different from those reported in table 6 and where they differ, they differ by very small margins.

Second, we use lagged data to construct the market potential index. Specifically, we construct a market potential index based on 5, 10, and 15 year lagged income, housing stocks and wages data[22]. In addition, we also use lagged variables for all the other controls. This strategy should reduce the correlation between the explanatory factors and the error term significantly. The results are reported in Table 12. While the importance of all the control factors remains evident in all columns, the magnitude of the estimates differs from those reported in table 6. Despite these differences, we derive qualitatively similar conclusions. In light of this evidence, we can conclude that reverse causality does not substantially affect our preferred results in Table 6.

7 Conclusion

Despite the ending of apartheid, regional wage disparities remain prevalent in South Africa with the former homelands characterised by persistently low wages and incomes. In this paper, we use a new economic geography (NEG) model to estimate the extent to which the persistence in apartheid regional wage disparities are an outcome of economic forces such as access to markets. We estimate a structural wage equation derived directly from the NEG theory for 354 regions over the period 1996 to 2011.

We find strong support for the NEG model in explaining regional wage disparities, but only after we augment the model to include specific first nature geography factors such as human capital, mineral resources, local climatic conditions, industrial structure and unemployment. We also find persistently adverse wage effects associated with the apartheid policies, with wages substantially lower in the former homeland areas, even after controlling for NEG and first nature geography factors. Wages in homeland areas are thus lower than what we would predict given their location (first and second nature geography factors) Average wages of workers in homeland areas were 22% lower than predicted in 1996, with this gap rising to 39% in 2011. These findings show that the reintegration of homeland areas into South Africa and continuous implementation of regional policies since the end of apartheid were not sufficient to reduce the homeland wage gap. Our results are robust to a number of sensitivity tests carried out to check for potential bias due to the way we measure homeland status as well as due to reverse causality.

Our study highlights the need for regional policy initiatives aimed at improving the underlying conditions of lagging regions, especially in homeland areas. It lends support to regional policy measures aimed at promoting human capital accumulation, greater access to markets, expansion of the manufacturing and mining sectors. In addition, it supports regional policies

the wage equation makes estimation incorporating instrumental variables extremely complicated, leading to non-convergence of the model (Moreno-Monroy, 2008; Paredes, 2015).

[22] This enables us to estimate three regression models using regional wage for 2001 and market potential for 1996, regional wage for 2011 and market potential for 2001, and regional wage for 2011 and market potential for 1996.

aimed at addressing the problems of high unemployment and low productivity in the agricultural sector.

References

Abel, M. (2015). *Long-run effects of forced removal under apartheid on social capital*.

Acemoglu, D., Johnson, S., & Robinson, J. A. (2002). Reversal of Fortune: Geography and Institutions in the Making of the Modern World Income Distribution. *The Quarterly Journal of Economics, 117*(4), 1231–1294.

Aghion, P., Braun, M., & Fedderke, J. (2008). Competition and productivity growth in South Africa. *Economics of Transition*, *16*(4), 741–768.

Allison, P. D. (2001). *Missing Data* (Vol. 136). SAGE Publications.

Ardington, C., Lam, D., Leibbrandt, M., & Welch, M. (2006). The sensitivity to key data imputations of recent estimates of income poverty and inequality in South Africa. *Economic Modelling*, *23*(5), 822–835.

Bosker, M, & Garretsen, H. (2012). Economic geography and economic development in Sub-Saharan Africa. *The World Bank Economic Review*, *26*(3), 443–485.

Bosker, M, & Krugell, W. (2008). Regional income evolution in South Africa after apartheid. *Journal of Regional Science*, *48*(3), 493–523.

Bosker, Maarten, & Krugell, W. (2008). *Regional Income Evolution in South Africa*. *48*(3), 493–523.

Brakman, S., Garretsen, H., & Schramm, M. (2004). The spatial distribution of wages: estimating the Helpman-Hanson model for Germany. *Journal of Regional Science, 44*(3), 437–466.

Breinlich, H. (2006). The spatial income structure in the European Union—what role for Economic Geography? *Journal of Economic Geography*, *6*(5), 593–617.

Burger, P. (2015). Wages, Productivity and Labour's Declining Income Share in Post-Apartheid South Africa. *South African Journal of Economics*, *83*(2), 159–173.

Burger, R., & Yu, D. (2007). Wage trends in post-apartheid South Africa: Constructing an earnings series from household survey data. *DPRU Working Paper 07/117. Development Policy Research Unit.*

Collins, L. M., Schafer, J. L., & Kam, C.-M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods*, *6*(4), 330.

Combes, P., Duranton, G., & Gobillon, L. (2008). Spatial wage disparities: Sorting matters! *Journal of Urban Economics*, *63*(2), 723–742.

De Arcangelis, G., & Mion, G. (2002). Spatial Externalities and Empirical Analysis: The case of Italy. *Working Paper No 66. Sapienza University of Rome, CIDEI.*

De Bruyne, K. (2010). Explaining the Location of Economic Activity. Is there a Spatial Employment Structure in Belgium? *International Journal of Economic Issues, 3*(2), 199–222.

Fallah, B. N., Partridge, M. D., & Olfert, M. R. (2011). New economic geography and US metropolitan wage inequality. *Journal of Economic Geography, 11*(5), 865–895.

Frame, E., De Lannoy, A., Koka, P., & Leibbrandt, M. (2016). Multidimensional Youth Poverty: Estimating the Youth MPI in South Africa at ward level. *Working Paper Series, Number 189. Southern Africa Labour and Development Research Unit, University of Cape Town*.

Gelb, S. (2004). An overview of the South African economy. *State of the Nation: South Africa*, *2005*, 367–400.

Hanson, G. (2005). Market potential, increasing returns and geographic concentration.

Journal of International Economics, *67*(1), 1–24.

Harris, I., Jones, P. D., Osborn, T. J., & Lister, D. H. (2014). Updated high-resolution grids of monthly climatic observations--the CRU TS3. 10 Dataset. *International Journal of Climatology*, *34*(3), 623–642.

Huang, Q., & Chand, S. (2015). Spatial spillovers of regional wages: Evidence from Chinese provinces. *China Economic Review*, *32*, 97–109.

King, B. H., & McCusker, B. (2007). Environment and development in the former South African bantustans. *The Geographical Journal*, *173*(1), 6–12.

Kingdon, G., & Knight, J. (2006). How flexible are wages in response to local unemployment in South Africa? *Industrial and Labour Relations Review*, *59*(3), 471–495.

Kosfeld, R., & Eckey, H. (2010). Market access, regional price level and wage disparities: the German case. *Jahrbuch Für Regionalwissenschaft*, *30*(2), 105–128.

Leibbrandt, M., Poswell, L., Naidoo, P., Welch, M., & Woolard, I. (2005). Measuring Recent Changes in South African Inequality and Poverty using 1996 and 2001 Census Data. *In: Bhorat, H. and R. Kanbur (Eds.) Poverty and Policy in Post-Apartheid South Africa, Pretoria, HSRC Press*, 1–51.

López-Rodríguez, J., & Faíña, J. A. (2006). Does distance matter for determining regional income in the European Union? An approach through the market potential concept. *Applied Economics Letters*, *13*(6), 385–390.

Magruder, J. (2012). High Unemployment Yet Few Small Firms: The Role of Centralized Bargaining in South Africa. *American Economic Journal: Applied Economics*, *4*(3), 138–166.

Mion, G. (2004). Spatial externalities and empirical analysis: the case of Italy. *Journal of Urban Economics*, *56*(1), 97–118.

Moreno-Monroy, A. (2008). The dynamics of spatial agglomeration in China: an empirical assessment. *Working Paper 08-06, Economics Program.*

Nel, E., & Rogerson, C. (2009). Re-thinking spatial inequalities in South Africa: Lessons from international experience. *Urban Forum*, *20*(2), 141–155.

Noble, M., & Wright, G. (2013). Using indicators of multiple deprivation to demonstrate the spatial legacy of apartheid in South Africa. *Social Indicators Research*, *112*(1), 187–201.

Noble, M., Zembe, W., & Wright, G. (2014). Poverty may have declined, but deprivation and poverty are still worst in the former homelands. *Southern African Social Policy Research Institute*.

Ntuli, M., & Kwenda, P. (2014). Labour unions and wage inequality among African men in South Africa. *Development Southern Africa*, *31*(2), 322–346.

Posel, D., & Casale, D. (2005). “Who Replies in Brackets and what are the Implications for Earnings Estimates?: An Analysis of Earnings Data from South Africa.” *Working Paper No. 07,Economic Research Southern Africa*.

Redding, S. J. (2013). Economic Geography: A review of the theoretical and empirical literature. *In Palgrave Handbook of International Trade. Palgrave Macmillan UK*, 497–531.

Redding, SJ. (2010). The empirics of new economic geography. *Journal of Regional Science*, *50*(1), 297–311.

Redding, Stephen, & Venables, A. J. (2004). Economic geography and international inequality. *Journal of International Economics*, *62*(1), 53–82.

Roth, P. L. (1994). Missing data: A conceptual review for applied psychologists. *Personnel Psychology*, *47*(3), 537–560.

Rubin, D. B. (1987). Multiple imputation for nonresponse in surveys. *New York: Wiley*.

Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. New York: Chapman & Hall.

von Fintel, D. (2007). Dealing with Earnings Bracket Responses in Household Surveys–How Sharp are Midpoint Imputations? *South African Journal of Economics*, *75*(2), 293-312.

von Fintel, D. (2017). Institutional wage-setting, labour demand and labour supply: Causal estimates from a South African pseudo-panel. *Development Southern Africa*, *34*(1), 1–16.

von Fintel, D. (2018). Long-run spatial inequality in South Africa: early settlement patterns and separate development. *Studies in Economics and Econometrics*, *42*(2), 81–102.

Von Fintel, D. (2015). *Wage flexibility in a high unemployment regime: spatial heterogeneity and the size of local labour markets*. REDI3x3 Working Paper 8. Research Project on Employment, Income Distribution and Inclusive Growth, Cape Town.

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, *48*(4), 817–838.

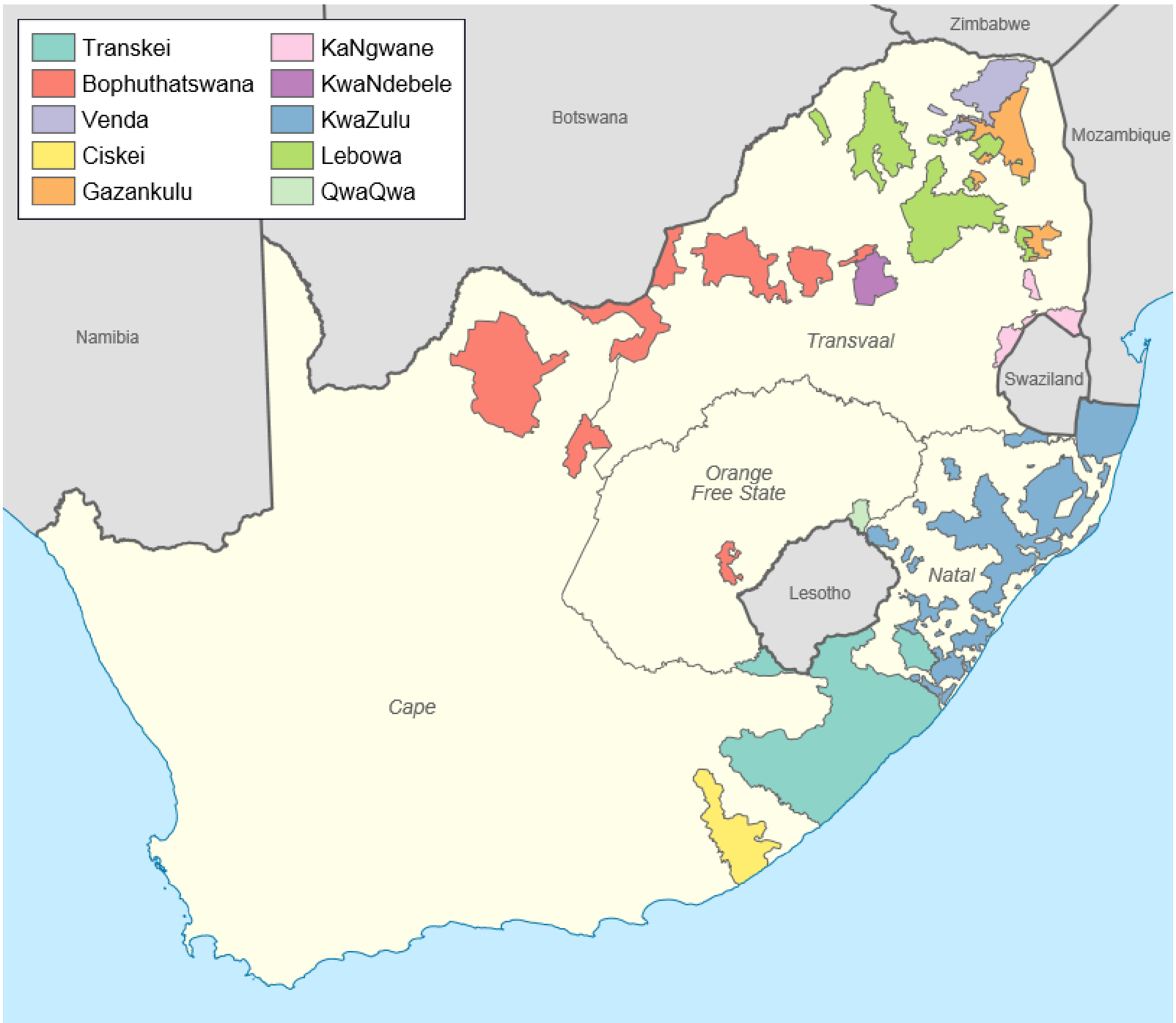
Wittenberg, M. (2014). Analysis of employment, real wage, and productivity trends in South Africa since 1994. *Series No. 994847703402676. International Labour Organization*.

Wittenberg, M. (2016). Wages and Wage Inequality in South Africa 1994–2011: Part 1–Wage Measurement and Trends. *South African Journal of Economics*, *00*(00), 1–21.

Wittenberg, M. (2017). Wages and Wage Inequality in South Africa 1994-2011: Part 2 - Inequality Measurement and Trends. *South African Journal of Economics*, *85*(2), 298–318.

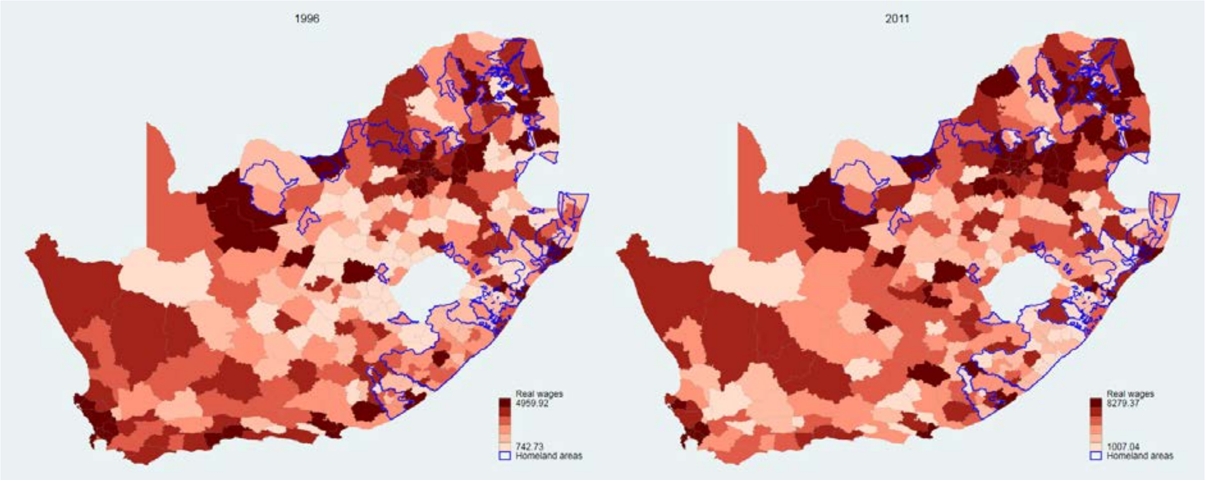
Zalk, N. (2014). Markups in South African Manufacturing-Are they high and what can they tell us? *TIPS. Manufacturing Led Growth for Employment and Equality. Conference Papers. Pretoria*.

Figure 1: Location of former homelands areas.



Source: http://commons.wikimedia.org/wiki/File:Bantustans_in_South_Africa.svg.

Figure 2: Spatial distribution of monthly average wages across regions in South Africa.



Notes: Author's calculations based on 1996 and 2011 census data aggregated to 354 magisterial districts. We proxy regional wages with regional income per worker, where income per worker is derived by weighting total income from employed individuals with total employed individuals in each region with a positive income and aged 15-64 years.

Table 1: Average wage for Homeland and Non-homeland areas in 1996 and 2011

	Homelands		Non-Homelands	
Variable	1996	2011	1996	2011
Mean	1409	1835	1614	2456
Std. Dev.	306	494	689	1048
Min	899	1194	743	1007
Max	3184	4224	4960	8279
Max/Min	3.54	3.54	6.68	8.22

Notes: We proxy regional wages with regional income per worker and to eliminate the influence of inflation, we convert nominal income for 2011 to its 1996 real income equivalent using the national consumer price index (CPI) provided by Stats SA. In deriving the homelands and non-homeland areas our unit of analysis is the 354 magisterial districts of South Africa. Of these districts, 105 are classified as Homelands and these are districts who have at least 20% of their area falling in former homeland areas. The other 249 are classified as non-homelands and these are districts who have less than 20% of their area in former homeland areas.

Table 2: Summary Statistics of Key variables in 1996 and 2011.

		1996				2011			
Variable	Obs	Mean	Std. Dev.	Min	Max	Mean	Std. Dev.	Min	Max
Wages	354	7.29	0.33	6.61	8.51	7.66	0.34	6.91	9.02
Market potential	354	19.90	1.12	18.02	23.56	21.63	1.17	19.53	25.61
Homeland status	354	0.30	0.46	0	1	0.30	0.46	0	1
Unemployment rate	354	0.37	0.19	0.03	0.84	0.32	0.11	0.06	0.58
Human capital	354	0.03	0.02	0.00	0.15	0.07	0.04	0.02	0.31
Housing stocks	354	10.78	1.23	8.20	13.79	11.25	1.28	8.54	14.59
Population	354	10.99	1.23	8.25	13.71	11.16	1.28	8.36	14.17
Income	354	16.84	1.41	14.10	21.05	18.53	1.47	15.70	23.09
Rainfall	354	3.99	0.49	1.96	4.82	3.92	0.37	2.51	4.60
Temperature	354	2.85	0.13	2.25	3.14	2.88	0.13	2.29	3.16

Note: Of these variables Wages, market potential, income, population, housing stocks, temperature, and rainfall are in logs, while homeland status, human capital and unemployment rate variables are proportions.

Table 3: Simple Correlations of Key variables, 2011

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
(1) Wages	1									
(2) homeland status	-0.33	1.00								
(3) Market potential	0.63	-0.02	1.00							
(4) Unemployment rate	-0.40	0.69	-0.20	1.00						
(5) Human capital	0.85	-0.25	0.58	-0.42	1.00					
(6) Housing stocks	0.42	0.30	0.87	0.17	0.44	1.00				
(7) Population	0.35	0.36	0.84	0.22	0.37	0.99	1.00			
(8) Income	0.63	0.07	0.94	-0.10	0.64	0.95	0.92	1.00		
(9) Rainfall	-0.08	0.40	0.05	0.47	-0.06	0.30	0.35	0.19	1.00	
(10) Temperature	0.01	0.44	0.32	0.20	0.02	0.40	0.43	0.33	0.10	1.00

Note: Of these variables wages, market potential, income, population, housing stocks, temperature, and rainfall are in logs, while homeland status, human capital and unemployment rate variables are proportions.

Table 4: Estimation of the Helpman-Hanson Model.

Reduced form coefficients	1996	2001	2011
Log market potential.	0.113***	0.098***	0.170***
	(0.036)	(0.037)	(0.046)
Log wages.	8.939***	10.511**	5.763***
	(3.015)	(4.183)	(1.710)
Log distance.	-3.103***	-4.250***	-2.209***
	(0.774)	(1.422)	(0.377)
Implied Values			
σ.	8.823***	10.17***	5.897***
	(2.812)	(3.867)	(1.606)
μ.	0.875***	0.872***	0.850***
	(0.020)	(0.022)	(0.027)
τ.	0.397***	0.463***	0.451***
	(0.050)	(0.048)	(0.076)
$\sigma/(\sigma-1)$.	1.128	1.109	1.204
$\sigma(1-\mu)$.	1.102	1.298	0.886
Adjusted R-squared	0.491	0.477	0.416
F-statistic	114.4	108.2	84.88
Obs	354	354	354

Asterisks indicate the level of significance, where: *** p<0.01, ** p<0.05, * p<0.1 and the values in parentheses are heteroscedasticity corrected standard errors. The estimated model includes a constant.

Table 5: Helpman-Hanson Model incorporating additional controls

Reduced form coefficients	1996	2001	2011
Log market potential.	0.235***	0.230***	0.462***
	(0.072)	(0.072)	(0.086)
Log wages.	4.084***	4.386***	1.875***
	(1.426)	(1.535)	(0.436)
Log distance.	-0.937***	-1.293***	-0.764***
	(0.111)	(0.181)	(0.085)
Implied Values			
σ.	4.261***	4.112***	2.166***
	(1.312)	(1.372)	(0.403)
μ.	0.798***	0.765***	0.622***
	(0.044)	(0.046)	(0.072)
τ.	0.287***	0.385***	0.655***
	(0.101)	(0.115)	(0.219)
$\sigma/(\sigma - 1)$.	1.307	1.321	1.857
$\sigma(1 - \mu)$.	0.859	0.966	0.819
Control variables			
Skilled workers	2.982***	1.153***	2.202***
	(0.290)	(0.239)	(0.186)
Share of mining workers	0.166***	0.349***	0.835***
	(0.091)	(0.165)	(0.151)
Share of agriculture workers	-0.606***	-1.032***	-0.254**
	(0.090)	(0.151)	(0.116)
Share of manufacturing workers	0.697***	0.247	0.692***
	(0.187)	0.307)	(0.216)
Temperature	0.408*	0.408**	0.149***
	(0.081)	(0.105)	(0.109)
Rainfall	-0.071***	-0.070***	-0.090
	(0.024)	(0.039)	(0.037)
Unemployment rate	-0.834***	-1.083*	-1.820**
	(0.072)	(0.146)	(0.136)
Adjusted R-squared	0.765	0.688	0.712
F-statistic	115.9	79.00	88.41
Obs	354	354	354

Asterisks indicate the level of significance, where: *** p<0.01, ** p<0.05, * p<0.1 and the values in parentheses are heteroscedasticity corrected standard errors. The estimated model includes a constant. Of the controls, wages, market potential, income, housing stocks, temperature, and rainfall are in logs, while, human capital, share of mining, manufacturing and agriculture workers, as well as unemployment rate, are proportions (%).

Table 6: Extended Helpman-Hanson Model with homeland status control

Reduced form coefficients	1996	2001	2011
Log market potential.	0.210***	0.229***	0.292***
	(0.078)	(0.075)	(0.088)
Log wages.	4.594**	4.371***	3.285***
	(1.892)	(1.598)	(1.128)
Log distance.	-0.864***	-1.089***	-0.739***
	(0.119)	(0.147)	(0.110)
Implied Values			
σ.	4.755***	4.363***	3.421***
	(1.755)	(1.433)	(1.029)
μ.	0.817***	0.769***	0.737***
	(0.047)	(0.048)	(0.062)
τ.	0.230***	0.324***	0.305***
	(0.098)	(0.112)	(0.123)
$\sigma/(\sigma-1)$.	1.266	1.311	1.413
$\sigma(1-\ \mu)$.	0.869	0.973	0.899
Control variables			
Skilled workers	3.387***	1.831***	2.382***
	(0.278)	(0.232)	(0.169)
Share of mining workers	0.161***	0.371***	0.650***
	(0.083)	(0.144)	(0.148)
Share agriculture workers	-0.666***	-1.018***	-0.486***
	(0.090)	(0.136)	(0.105)
Share manufacturing workers	0.681***	0.339	0.359**
	(0.171)	(0.277)	(0.183)
Temperature	0.481*	0.540**	0.263
	(0.092)	(0.122)	(0.097)
Rainfall	-0.063***	-0.073***	-0.039***
	(0.022)	(0.035)	(0.029)
Unemployment rate	-0.580***	-0.767**	-1.165***
	(0.081)	(0.150)	(0.141)
Homeland status	-0.218***	-0.368***	-0.389***
	(0.034)	(0.051)	(0.036)
Adjusted R-squared	0.785	0.731	0.786
F-statistic	118.1	88.13	119.2
Obs	354	354	354

Asterisks indicate the level of significance, where: *** p<0.01, ** p<0.05, * p<0.1 and the values in parentheses are heteroscedasticity corrected standard errors. The estimated model includes a constant. Our key variable is Homeland status, which is a ratio is given by the share of each region's area that falls in former homeland areas.

Table 7: Extended Helpman-Hanson Model with alternative homeland status measure

Reduced form coefficients	1996	2001	2011
Log market potential.	0.218***	0.226***	0.332***
	(0.082)	(0.076)	(0.093)
Log wages.	4.423**	4.437***	2.799***
	(1.874)	(1.646)	(0.921)
Log distance.	-0.879***	-1.179***	-0.713***
	(0.116)	(0.144)	(0.104)
Implied Values			
σ.	4.596***	4.419***	3.008***
	(1.731)	(1.476)	(0.844)
μ.	0.813***	0.771***	0.717***
	(0.049)	(0.048)	(0.067)
τ.	0.245**	0.345***	0.355**
	(0.104)	(0.120)	(0.143)
$\sigma/(\sigma-1)$.	1.278	1.292	1.498
$\sigma(1-\mu)$.	0.859	1.014	0.850
Control variables			
Skilled workers	3.204***	1.532***	2.319***
	(0.250)	(0.205)	(0.167)
Share of mining workers	0.154*	0.325*	0.651***
	(0.093)	(0.172)	(0.136)
Share agriculture workers	-0.647***	-1.027***	-0.425***
	(0.090)	(0.136)	(0.103)
Share manufacturing workers	0.688***	0.271	0.416**
	(0.186)	(0.267)	(0.203)
Temperature	0.475***	0.542***	0.280***
	(0.087)	(0.106)	(0.106)
Rainfall	-0.066***	-0.070*	-0.048
	(0.024)	(0.042)	(0.034)
Unemployment rate	-0.675***	-0.870***	-1.308***
	(0.087)	(0.129)	(0.140)
Homeland status	-0.121***	-0.206***	-0.265***
	(0.035)	(0.045)	(0.034)
Adjusted R-squared	354	354	354
Obs	0.773	0.706	0.757

Asterisks indicate the level of significance, where: *** p<0.01, ** p<0.05, * p<0.1 and the values in parentheses are heteroscedasticity corrected standard errors. The estimated model includes a constant. Homeland status is defined as homeland status = 1 if at least 20% of a region's area falls within a homeland area and 0 otherwise.

Table 8: Extended Helpman-Hanson Model with alternative homeland status measure

Reduced form coefficients	1996	2001	2011
Log market potential.	0.226***	0.252***	0.346***
	(0.080)	(0.079)	(0.087)
Log wages.	4.229**	3.924***	2.660***
	(1.708)	(1.388)	(0.789)
Log distance.	-0.879***	-1.096***	-0.759***
	(0.113)	(0.121)	(0.095)
Implied Values			
σ.	4.432***	3.972***	2.887***
	(1.580)	(1.249)	(0.725)
μ.	0.812***	0.757***	0.709***
	(0.048)	(0.051)	(0.064)
τ.	0.256**	0.369***	0.403***
	(0.104)	(0.131)	(0.146)
$\sigma/(\sigma-1)$.	1.291	1.336	1.530
$\sigma(1-\mu)$.	0.134	0.141	0.204
Control variables			
Skilled workers	3.265***	1.597***	2.242***
	(0.250)	(0.199)	(0.166)
Share of mining workers	0.150	0.332**	0.691***
	(0.092)	(0.169)	(0.134)
Share agriculture workers	-0.639***	-1.050***	-0.420***
	(0.088)	(0.134)	(0.102)
Share manufacturing workers	0.612***	0.226	0.386*
	(0.186)	(0.262)	(0.204)
Temperature	0.470***	0.509***	0.250**
	(0.086)	(0.105)	(0.105)
Rainfall	-0.073***	-0.089**	-0.078**
	(0.024)	(0.041)	(0.033)
Unemployment rate	-0.691***	-0.925***	-1.360***
	(0.080)	(0.120)	(0.135)
Homeland status	-0.129***	-0.231***	-0.246***
	(0.030)	(0.038)	(0.031)
Adjusted R-squared	354	354	354
Obs	0.777	0.717	0.578

Asterisks indicate the level of significance, where: *** p<0.01, ** p<0.05, * p<0.1 and the values in parentheses are heteroscedasticity corrected standard errors. The estimated model includes a constant. Homeland status is defined as homeland status = 1 if a region's centroid falls within a homeland area and 0 otherwise.

Table 9: Extended Helpman-Hanson Model, alternative homeland status measure, 2011

Parameters	0 km	≤10 km	≤20 km	≤30 km	≤33 km
Log market potential.	0.346***	0.392***	0.450***	0.497***	0.486***
	(0.097)	(0.109)	(0.105)	(0.105)	(0.104)
Log wages.	2.661***	2.319***	1.949***	1.711***	1.753***
	(0.881)	(0.772)	(0.565)	(0.461)	(0.476)
Log distance.	-0.760***	-0.730***	-0.730***	-0.705***	-0.714***
	(0.110)	(0.112)	(0.102)	(0.100)	(0.100)
Implied Values					
σ.	2.888***	2.554***	2.223***	2.014***	2.056***
	(0.808)	(0.712)	(0.520)	(0.425)	(0.440)
μ.	0.709***	0.670***	0.628***	0.593***	0.602***
	(0.071)	(0.086)	(0.088)	(0.092)	(0.091)
τ.	0.403**	0.470**	0.597**	0.695**	0.676**
	(0.163)	(0.202)	(0.243)	(0.285)	(0.278)
$\sigma/(\sigma-1)$.	1.530	1.644	1.818	1.986	1.947
$\sigma(1-\mu)$.	0.839	0.843)	0.828	0.820	0.818
Control variables					
Skilled workers	2.242***	2.363***	2.326***	2.296***	2.276***
	(0.198)	(0.206)	(0.217)	(0.226)	(0.229)
Share mining workers	0.691***	0.741***	0.772***	0.798***	0.806***
	(0.169)	(0.168)	(0.172)	(0.178)	(0.179)
Share agriculture workers	-0.420***	-0.277**	-0.247**	-0.255**	-0.253**
	(0.104)	(0.109)	(0.115)	(0.122)	(0.124)
Share manufacturing workers	0.385*	0.729***	0.686***	0.684***	0.682***
	(0.205)	(0.186)	(0.198)	(0.208)	(0.211)
Temperature	0.250**	0.393***	0.294**	0.221	0.199
	(0.114)	(0.126)	(0.134)	(0.137)	(0.138)
Rainfall	-0.078**	-0.026	-0.040	-0.059	-0.064
	(0.038)	(0.040)	(0.044)	(0.046)	(0.047)
Unemployment rate	-1.358***	-1.444***	-1.622***	-1.776***	-1.785***
	(0.166)	(0.176)	(0.153)	(0.145)	(0.145)
Homeland status	-0.246***	-0.211***	-0.127***	-0.068**	-0.053*
	(0.040)	(0.037)	(0.036)	(0.032)	(0.032)
Adjusted R-squared	0.758	0.747	0.726	0.716	0.714
F-statistic	101.32	95.53	86.09	81.93	81.22
Obs	354	354	354	354	354

Asterisks indicate the level of significance, where: *** $p<0.01$, ** $p<0.05$, * $p<0.1$ and the values in parentheses are heteroscedasticity corrected standard errors. The estimated model includes a constant. Homeland status is defined as homeland status = 1 if a region's distance from any homeland boundary=0; ≤10; ≤20; ≤30 km; ≤33 km and 0 otherwise. We only report results for 2011 as we reach similar conclusions for 1996 and 2001.

Table 10: Pooled Helpman-Hanson Model with initial homeland status control

Reduced form coefficients	(1)	(2)	(3)
Log market potential.	0.141***	0.209***	0.215***
	(0.019)	(0.035)	(0.039)
Log wages.	7.133***	4.785***	4.651***
	(1.044)	(0.885)	(0.939)
Log distance.	-2.732***	-1.136***	-1.029***
	(0.275)	(0.077)	(0.074)
Implied Values			
σ.	7.111***	4.785***	4.655***
	(0.972)	(0.805)	(0.852)
μ.	0.857***	0.791***	0.786***
	(0.011)	(0.023)	(0.025)
τ.	0.447***	0.300***	0.282***
	(0.029)	(0.052)	(0.056)
$\sigma/(\sigma-1)$.	1.164	1.264	1.274
$\sigma(1-\mu)$.	1.019	1.000	0.998
Control variables			
Skilled workers		1.898***	2.092***
		(0.122)	(0.124)
Share of mining workers		0.259***	0.266***
		(0.090)	(0.083)
Share agriculture workers		-0.760***	-0.779***
		(0.066)	(0.064)
Share manufacturing workers		0.318**	0.374***
		(0.136)	(0.131)
Temperature		0.470***	0.446***
		(0.067)	(0.063)
Rainfall		-0.0557***	-0.0492***
		(0.017)	(0.016)
Unemployment rate		-0.587***	-0.784***
		(0.072)	(0.071)
Homeland status		-0.345***	-0.0822***
		(0.0266)	(0.031)
Dummy 2001	-0.104***	-0.241***	-0.180***
	(0.021)	(0.0250)	(0.027)
Dummy 2011	-0.0665***	-0.326***	-0.250***
	(0.019)	(0.0195)	(0.020)
Homeland status*Dummy 2001			-0.275***
			(0.0364)
Homeland status*Dummy 2011			-0.433***
			(0.031)
Observations	1,062	1,062	1,062
Adj. R-squared	0.465	0.717	0.751

Asterisks indicate the level of significance, where: *** p<0.01, ** p<0.05, * p<0.1 and the values in parentheses are heteroscedasticity corrected standard errors. The estimated model includes a constant. Column (1) report estimates of the baseline model with time fixed effects, column (2) adds other controls, while column (3) adds interaction terms between homeland status and time fixed effects. Our key variable is Homeland status, which is a ratio is given by the share of each region's area that falls in former homeland areas.

Table 11: Sensitivity tests – Using regional population as a proxy regional market size

	Regional population		
Reduced form coefficients	1996	2001	2011
Log market potential.	0.208***	0.275***	0.281***
	0.058	0.060	0.073
Log wages.	4.596***	3.500***	3.413***
	1.459	0.888	1.020
Log distance.	-0.884***	-1.023***	-0.758***
	0.106	0.099	0.105
Implied Values			
σ.	4.802***	3.633***	3.557***
	1.343	0.795	0.921
μ.	0.827***	0.752***	0.749***
	0.034	0.039	0.049
τ.	0.232***	0.389***	0.296***
	0.082	0.113	0.110
$\sigma/(\sigma-1)$.	1.263	1.380	1.391
$\sigma(1-\mu)$.	0.830	0.900	0.892
Control variables			
Skilled workers	3.395***	1.893***	2.378***
	0.276	0.229	0.168
Mineral endowments	0.162*	0.353**	0.650***
	0.083	0.144	0.148
Temperature	0.477***	0.528***	0.259
	0.087	0.121	0.095
Rainfall	-0.069***	-0.100***	-0.041
	0.024	0.037	0.030
Unemployment rate	-0.584***	-0.810***	-1.170***
	0.083	0.156	0.142
Share of agriculture workers	-0.659***	-0.994***	-0.486***
	0.090	0.136	0.105
Share of manufacturing workers	0.678***	0.343	0.359*
	0.172	0.276	0.184
Homeland status	-0.216***	-0.363***	-0.389***
	0.034	0.051	0.037
Adjusted R-squared	0.785	0.732	0.787
F-statistic	118.4	88.7	119.2
Obs	354	354	354

Notes: Asterisks indicate the level of significance, where: *** p<0.01, ** p<0.05, * p<0.1 and the values in parentheses are heteroscedasticity corrected standard errors. Estimated models include a constant. In the estimations, we use regional population instead of regional income as a proxy for regional market size.

Table 12: Sensitivity test – Using lagged market potential and other control variables.

	Lagged variables		
Reduced form coefficients	5yrlag	10yrlag	15yrlag
Log market potential.	0.203**	0.229***	0.210**
	0.098	0.076	0.081
Log wages.	4.742*	4.370***	4.595**
	2.559	1.618	1.981
Log distance.	-0.973***	-1.089***	-0.864***
	0.177	0.129	0.116
σ.	4.937***	4.363***	4.755***
	2.379	1.452	1.829
μ.	0.830***	0.769***	0.817***
	0.056	0.048	0.048
τ.	0.247**	0.324***	0.230**
	0.124	0.117	0.099
$\sigma/(\sigma-1)$.	1.254	1.297	1.266
$\sigma(1-\mu)$.	0.838	1.006	0.869
Skilled workers	3.390***	1.831***	3.387***
	0.379	0.201	0.246
Mineral endowments	0.114	0.371**	0.161*
	0.110	0.165	0.090
Temperature	0.541***	0.540***	0.481***
	0.124	0.102	0.084
Rainfall	-0.072**	-0.073*	-0.063***
	0.032	0.040	0.023
Unemployment rate	-0.384***	-0.767***	-0.580***
	0.137	0.124	0.085
Share of agriculture workers	-0.923***	-1.018***	-0.666***
	0.127	0.130	0.087
Share of manufacturing workers	0.446	0.339	0.681***
	0.275	0.254	0.181
Homeland status	-0.361***	-0.368***	-0.218***
	0.056	0.050	0.039
Adjusted R-squared	0.662	0.731	0.785
F-statistic	63.9	88.1	118.1
Obs	354	354	354

Notes: Asterisks indicate the level of significance, where: *** p<0.01, ** p<0.05, * p<0.1 and the values in parentheses are heteroscedasticity corrected standard errors. Estimated models include a constant. The estimates are based on 5, 10 and 15 year lagged data for market potential and other control variables.

Appendix

Additional Tables

Table 13: Helpman-Hanson model – structural parameter constraints

Structural parameter	Parameter description
$\alpha_1 > 0.$	Market potential estimate
$\sigma > 1.$	Elasticity of substitution between manufactured varieties
$0 < \mu < 1.$	Share of income devoted to manufactured varieties
$\tau > 0.$	Unit transport cost
$\sigma/(\sigma - 1) > 1.$	Market power condition reflecting imperfect competition
$\sigma(1 - \mu) < 1.$	No-black-hole condition

Notes: These structural parameters are derived from the reduced-form coefficients obtained from estimating equations (5) and (6). Thus, given α_1, α_2 and α_3 the structural parameters are obtained as follows: $\sigma = 1/\alpha_1$, $\mu = (1 - \alpha_1)/\alpha_1\alpha_2$ and $\tau = \alpha_1\alpha_3/(\alpha_1 - 1)$. From these parameters, two additional equilibrium conditions given by $\sigma/(\sigma - 1)$ – price-marginal cost ratio and $\sigma(1 - \mu)$ – no black hole condition, are also derived.

Table 14: Correlation coefficients.

	Wages	Income	Population
Wages	1.0000		
Income	0.7387	1.0000	
Population	0.4025	0.8236	1.0000

Note: Variables are in levels and are for the entire sample period (1996–2011). Thus, the correlation coefficient is an average value for 1996, 2001 and 2011 data.

Table 15: Association between regional wage and homeland status indicator.

	1996	2001	2011
Homeland Status	-0.105***	-0.112***	-0.302***
	(0.030)	(0.038)	(0.033)
Constant	-0.376***	-0.509***	-0.378***
	(0.023)	(0.028)	(0.022)
Observations	354	354	354
R-squared	0.016	0.013	0.121
F-test	12.04	8.764	81.25

Asterisks indicate the level of significance, where: *** p<0.01, ** p<0.05, * p<0.1 and the values in parentheses are heteroscedasticity corrected standard errors.

Addressing the challenges with the census data

Inconsistent geographical units across the censuses over time

The greatest advantage of the censuses is the availability of information at different geographical levels (including municipalities, main places (cities/towns) and sub-places (villages/suburbs). However, a major challenge with these geographical units is their inconsistencies across the censuses over time. To address this challenge, we use ArcGIS to overlay 2011 sub-place boundaries onto 1996/2001 magisterial district boundaries. Based on the overlaying results, we use areal-weighting interpolation technique to assign 2011 sub-place population values to their corresponding 1996/2001 magisterial district population values. This leads to a consistent longitudinal and cross-sectional dataset containing a total of 354 magisterial districts that we use as our unit of analysis.

Looking at the results, we see that having started with 354 magisterial district units and 22108 sub-place units the overlay process produced 279366 union zone units, with an areal-weighting ratio of between 0 and 1[23]. These union zones account for 1219067 square km of South Africa's total land area of 1219602 square km as of 2001 (99. 95 percent). Furthermore, the union zone areas account of 13156028 of South Africa's total employment of 13179825 in 2011 (99.83 percent)[24]. The areal-weighting interpolation process, thus, fails to account for 535 square km of the land and 23797 employment. This translates to a prediction error of less than 1% for the total area (0.044%) and total employment (0.181%). Our analysis of the distribution of the error values based on the mean and mean absolute percentage error (MAPE) statistics show that these errors are of a reasonable size and are close to zero[25].

Challenges with the census income data

An additional challenge with the census is the unavailability of wage (labour income) data which is critical in a study of regional wage disparities. To overcome this challenge, we follow existing literature (Redding & Venables, 2004; Bosker & Garretsen, 2012; Breinlich, 2006) and use regional income per worker to proxy for regional wage per worker. Regional income per worker is calculated as the simple average income of all workers, aged 15-64 years, with a positive income in a region. The income variable used to derive regional income per worker is defined as the sum of basic salary, bonuses, allowances, income from grants, transfers, remittances and any other income source received by individuals. We acknowledge that regional income per worker is an imprecise proxy for regional wage per worker, as it contains income from other sources (non-wage income). However, we argue that it is a good proxy in the case of South Africa, given that labour income (wages) contributes the largest share to total income of employed individuals. To support our argument, we compare income per worker

[23] The ideal ratio is 1, which shows that a given sub-place falls completely in a given magisterial district. Of the 27366 union zones, 18277 union zones that accounts for 12.37% of the country's total area and a massive 71.47% of the country's total employment have a ratio of 1. If we consider a ratio of at least 0.7, we see that 21556 union zones have a ratio of between 0.7 and 1. These union zones account for 51.07% of the country's total area and a massive 92.02 percent of the country's total employment. On the other end, 3828 union zones which account for 3.94% of the country's total area and 0.85% of the country's total employment have a ratio of between 0 and 0.09.

[24] While we focus on employment only in this discussion, the same conclusions reached for employment also holds for other population variables.

[25] The error distribution is reasonably symmetrical with a mean of -1.85, which is close to zero. Further, the MAPE, which expresses the accuracy of the prediction as a percentage of the error, shows that on average the predicted values are off the true values by 0.24 percent.

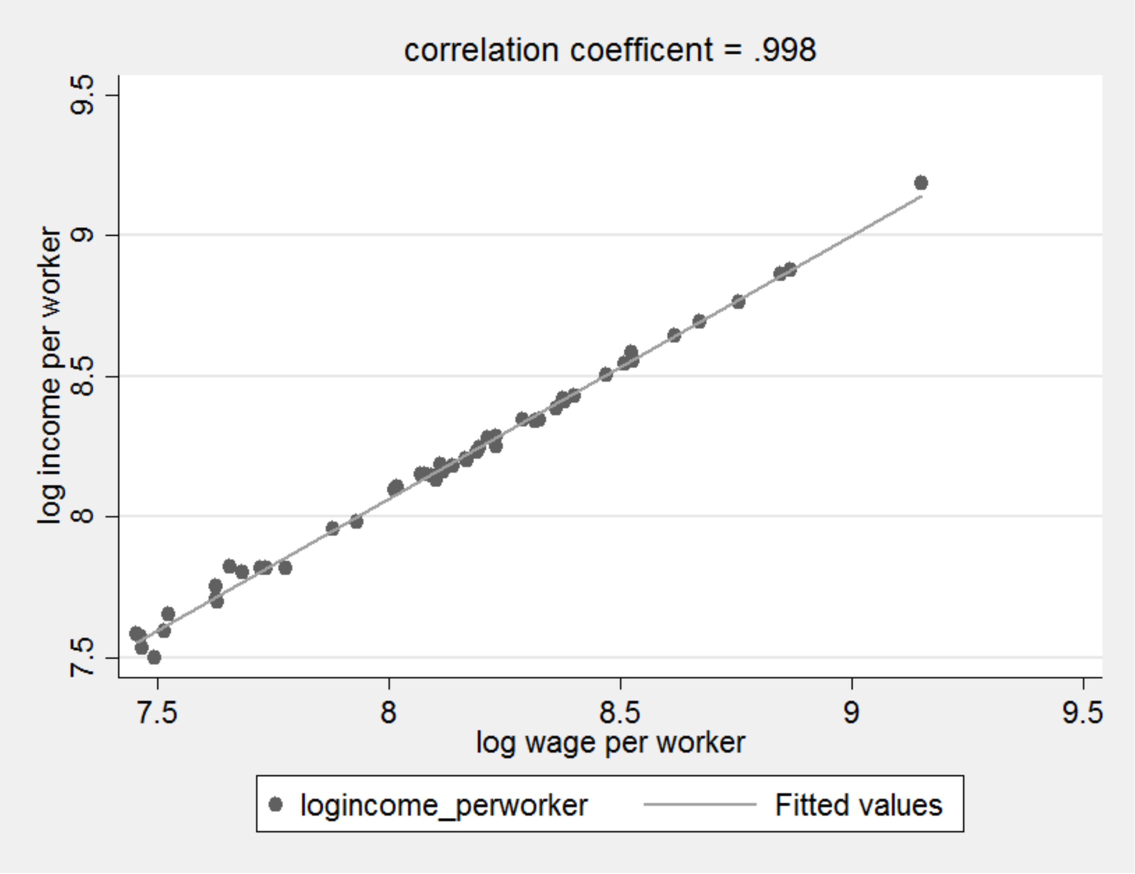
and wage per worker across 53 district councils in South Africa based on NIDS 2010/2011 household survey data.

The NIDS survey collects data on various sources of individual income such as labour (wage) income, government grant income, other government income, investment income, income of a capital nature and remittance income. Of these sources, labour (wage) income is the sum of income from main and secondary job wages, casual wages, self-employment income, 13th cheque, other bonus, profit share, income from helping a friend and extra piece-rate income. From these different sources, we derive three income variables: individual income, income from employed individuals and labour (wage) income. Based on these variables, we further derive district income per worker, given as total income from employed individuals divided by total employed individuals in each district, as well as district wage per worker, given as total labour (wage) income divided by total employed individuals in each district.

Analysis of these variables shows that, on average, income from employed individuals contributes about 72.5 percent to total income across all districts, a figure close to the 73.3 percent we find from the analysis of 2011 census data. Furthermore, on average, labour income accounts for about 69 percent of total income across districts, a figure consistent with the 70 percent normally found in most individual and household studies in post-apartheid South Africa (see Leibbrandt et al. 2010). By narrowing down to labour (wages) income and income of employed individuals only, we find that on average, labour income accounts for 94.7 percent of total income from employed individuals, across districts. This shows that the bulk of the income from employed individuals comes from labour income, with roughly 5.3 percent coming from other sources (for example, income from grants, transfers, remittances.). Thus, we expect regional income per worker to overstate regional wage per worker by a small proportion. Figure 3 that gives the relationship between income per worker and wage per worker across districts show a high correlation of 0.998 between these variables. This suggests that, on average, income per worker is a good predictor of wage per worker. Hence, income per worker can be said to be a good proxy of wage per worker.

We confirm this by assessing whether district-specific factors such as industrial composition, human capital, unemployment, and geographical location have any additional effect on the spatial variation of income per worker after the effects of wage per worker are accounted for. For income per worker to be a good proxy for wage per worker, we expect district-specific factors to have no additional contribution in explaining variations in district income per worker. To see whether this is the case, we regress district wage per worker on district income per worker, controlling for other district-specific factors.

Figure 3: Association between income and wage per worker across districts.



Notes: Calculation using income and wage data from NIDS Wave 2 for a sample of 53 district councils.

The estimation results are presented in Table 16, where column (1) reports estimates for the association between district income and wage per worker, while column (2) reports estimates where we add district-specific factors. Columns (1) and (2) show a highly positive and statistically significant association between district income and wage per worker. With the exception of the share of workers in the agricultural sector, all the other district-specific factors are not significant[26]. Based on this analysis, we conclude that income per worker is a good proxy for wage per worker in South Africa.

[26] Given that the agricultural sector is highly concentrated in rural areas, the significance of the share of workers in the agricultural sector suggests that income per worker might not be a good proxy for wage per worker in rural areas. It is the case that, while workers in rural locations might receive low average wages, they might receive income from other sources, which would show higher income per worker.

Table 16: Association between regional income and wage per worker in South Africa.

	(1)	(2)
Log wage per worker	0.940***	0.936***
	(0.009)	(0.010)
% Human capital		-0.505
		(0.415)
Rural dummy		0.007
		(0.007)
% Manufacturing sector workers		-0.047
		(0.069)
% Agricultural sector workers		-0.084**
		(0.032)
% Mining sector workers		-0.069
		(0.072)
% Participation rate		-0.016
		(0.035)
% Unemployment rate		0.003
		(0.034)
Constant	0.545***	0.599***
	(0.072)	(0.087)
Observations	53	53
R-squared	0.995	0.997
F-test	11137	1836

Notes: Standard errors in parentheses *** p<0.01, ** p<0.05, * p<0.1. Using income per worker as the dependent variable, column (1) reports the estimates of the association between regional income and wage per worker, while column (2) adds regional specific factors. Human capital is the share of each region's population with at least a tertiary degree.

A major challenge with the census income data that we use to derive income per worker is that it is collected in brackets (see Table 17). To construct a continuous income measure, we assign the midpoint of each bracket to everyone in that bracket[27]. For the highest band, which is open-ended, we set the midpoint to two times the lower bound of the highest bracket, a rule employed by Stats SA. While the midpoint approach has been found to exaggerate income inequality (Wittenberg, 2017)[28] and reduce income variability, it has been found to lead to similar conclusions as other complicated techniques such as the reweighting approach, hot deck approach, mean imputation, and multiple imputation (Posel & Casale, 2005; Ardington et al. 2006; Burger & Yu, 2007; von Fintel, 2007). Its appropriateness has seen it being used widely by other researchers who analyse South Africa's survey data (Hofmeyr, 1999; Casale, Muller

[27] For example, a band of 1 to 400 rand will take the value of 200.5 Rands. Alternative approaches include weighting, multiple imputation and non-parametric techniques. For more detail on these methods see Wittenberg (2017).

[28] Although the midpoint approach can exaggerate income inequality, this exaggeration will be consistent over the various censuses, such that income inequality conclusions across the censuses will also be consistent.

& Posel, 2004; Kingdon & Knight, 2004; Meth & Dias, 2004; Leibbrandt et al. 2005; Vermaak, 2005).

Table 17: Monthly personal income brackets in the various censuses.

1996 census	2001/2011 census
No income	No income
R1 - R200	R1 - R400
R201 - R500	R401 - R800
R501 - R1000	R801 - R1600
R1001 - R1500	R1601 - R3200
R1501 - R2500	R3201 - R6400
R2501 - R3500	R6401 - R12 800
R3501 - R4500	R12 801 - R25 600
R4501 - R6000	R25 601 - R51 200
R6001 - R8000	R51 201 - R102 400
R8001 - R11000	R102 401 - R204 800
R11001 - R16000	R204 801+
R16001- R30000	
R30001+	

Notes: Brackets for income in the censuses.

While we can use the mid-point approach, a challenge with these brackets is their inconsistency across the three censuses. For instance, 1996 income brackets are narrower than 2001 and 2011 brackets, which are similar (Table 17). In addition, the top end of the brackets also differs significantly. The value of the top-end bracket in 1996 is set at R30 000 or more and is R204801 or more for 2001 and 2011. To reduce the possible bias due to these inconsistencies and allow comparability of income over time, we compressed 2001- and 2011-income brackets into their 1996 real income equivalents, using the CPI values provided by Stats SA: 38.5, 52.4 and 92.6 for 1996, 2001 and 2011 census, respectively (Stats SA)[29].

A final challenge of the census income data is the high rate of reported zero and missing income. For example, 70.7%, 68.2% and 49.1% of the respondents in 1996, 2001 and 2011 had zero or missing income (see Table 18). Interestingly, narrowing down to employed individuals who are the focus of this study significantly reduces the proportion of individuals with zero or missing income to 5%, 2.2% and 13.3% in 1996, 2001 and 2011 (see Table 16). Given the employed individuals with missing and zero income, the question is how to deal with these individuals. The most common technique which we use in this study is to drop these individuals. However, dropping employed individuals with missing (missing plus zero) income information can introduce potential bias in parameter estimation (Rubin, 1987; Schafer, 1997)[30].

[29] http://www.statssa.gov.za/publications/P0141/CPIHistory.pdf?

[30] We assume that all employed individuals who reported zero income have missing income because any employed individual is highly likely to receive a positive income. Thus, we drop employed individuals with missing income information.

Table 18: Proportion of individuals with missing and zero income

	All Individuals			Employed Individuals		
Year	Missing	Zero	Missing & Zero	Missing	Zero	Missing & Zero
1996	10.1%	60.6%	70.7%	3.8%	1.2%	5%
2001	0%	68.2%	68.2%	0%	2.2%	2.2%
2011	7.9%	41.2%	49.1%	4.7%	8.5%	13.3%

Note: All Individuals includes all the people interviewed in the censuses.

The extent of this bias is negligible when data is missing completely at randomly (MCAR) and to some extent when data is missing at random (MAR) but not negligible when data is missing not at random (MNAR)[31]. Given this, it is important to check whether income information is MCAR, MAR or MNAR. While it is difficult to check whether data is MAR and MNAR, we can easily check whether data is MCAR. We achieve this by running a logistic regression predicting missingness (0 = not missing, 1 = missing) from specific observed variables. Significant coefficients, either singly or jointly, would indicate a violation of MCAR. Our results presented in Table 19 show highly significant coefficients for the bulk of the factors in all columns, indicating a violation of MCAR in 1996, 2001 and 2011. These results suggest that rather than missing completely at random, income information is missing systematically driven by age, education, race, gender, and location[32].

Given these results, excluding employed individuals with missing income information from our analysis might bias our results[33]. Since our analysis is at the regional level, if those workers with missing income data are concentrated at the bottom of the distribution, then the level of income per worker of a given region will be overestimated. Alternatively, if those workers with missing income information disproportionally fall at the top of the distribution, then the level of income per worker of a given region will be underestimated.

[31] Data is MCAR if the probability of missingness does not depend on any variable, either observed or unobserved, while data is MAR if the probability of missingness depends only on observed variables and not unobserved or missing information. MCAR is a special case of MAR. Finally, data is MNAR if the probability of missingness depend on unobserved factors which are not measured by the researcher.

[32] As a robustness check, we also estimated a logistic regression model with regional specific factors like income per worker, market potential, share of workers with higher education and unemployment rate. Our results continued to reveal highly significant coefficients for these factors for all the years, indicating a violation of MCAR.

[33] Interestingly, research suggests that violation of the MCAR does not seriously bias parameter estimates (Collins, Schafer, & Kam, 2001), especially after controlling for those factors highly correlated with the variable of interest (see Allison, 2001). Accordingly, our empirical analysis controlled for a number of regional specific factors highly correlated with income per worker, to reduce the potential bias due to omitted workers with missing income information.

Table 19: Logistic regression model predicting missingness

	1996	2001	2011
VARIABLES	Missing dummy	Missing dummy	Missing dummy
Age	-0.005***	-0.025***	-0.032***
	(0.001)	(0.001)	(0.000)
Education	-0.002	-0.053***	-0.085***
	(0.002)	(0.002)	(0.001)
Gender: Female	0.107***	0.251***	0.380***
	(0.011)	(0.015)	(0.006)
Race: Coloured	0.565***	-0.214***	0.127***
	(0.021)	(0.031)	(0.011)
Indian	0.435***	-0.293***	0.193***
	(0.030)	(0.046)	(0.015)
White	1.146***	0.174***	0.378***
	(0.015)	(0.023)	(0.009)
Location: Urban	-0.013	-0.211***	-0.157***
	(0.016)	(0.020)	(0.008)
Province: Eastern Cape	0.114***	0.327***	0.272***
	(0.024)	(0.034)	(0.013)
Northern Cape	-0.372***	-0.017	-0.261***
	(0.045)	(0.056)	(0.023)
Free State	-0.257***	0.014	-0.143***
	(0.031)	(0.040)	(0.016)
Kwazulu-Natal	0.217***	0.260***	0.203***
	(0.023)	(0.032)	(0.012)
North West	-0.021	-0.197***	-0.089***
	(0.029)	(0.040)	(0.015)
Gauteng	0.176***	0.026	0.076***
	(0.019)	(0.029)	(0.010)
Mpumalanga	0.331***	-0.190***	-0.118***
	(0.028)	(0.041)	(0.015)
Limpopo	0.327***	0.010	-0.075***
	(0.031)	(0.040)	(0.015)
Constant	-3.425***	-2.397***	-0.030*
	(0.033)	(0.045)	(0.018)
Observations	722,718	767,180	1,073,587

Note: Missing dummy is an indicator variable taking a value of 1 if a worker has missing income information and 0 otherwise. Standard errors in parentheses *** p<0.01, ** p<0.05, * p<0.1

While acknowledging this potential bias, we argue that dropping workers with missing income information will not change our overall conclusions given the small sample size of workers with missing income information in the census (5%, 2% and 13% in 1996, 2001 and 2011). Although there is no established cut-off from the literature regarding an acceptable percentage of missing information for valid statistical inferences, our claim finds support from Roth (1994) who argued that the choice of a missing data estimation technique can have substantial implications for the parameter estimates as the portion of missing data reaches 15% to 20%. To further support our claim, we will also carry out robustness checks to see whether our main results that exclude workers with zero income differ significantly from the results when we include workers with zero income.