# Ethnic Diversity and Labour Market Outcomes: Evidence from Post-apartheid South Africa*

Sara Tonini † Peng Zhang ‡

December 4, 2020

***Abstract.*** Using 1996 and 2001 South African census data, this paper investigates how the ethnic diversity amongst the black population affects labour market outcomes. To address endogeneity concerns, we employ an original instrumental variable approach, which relies on the historical spatial distribution of the ethnic-specific homelands. In particular, we exploit the fact that a district which is equidistant to all the different homelands is more likely to be ethnically heterogenous than a district which is close to the homeland of a specific group and distant from the others. We find that ethnic diversity significantly increases the employment rate of black South Africans and that this effect is driven by an increase in high-skill employment. To explain this result, we propose a model based on human capital and network theories. In more homogenous districts, where individuals can rely on larger networks to get a low skill job, they tend to invest less in education. The opposite occurs in more diverse districts where the absence of large networks incentivise individuals to get higher education in order to find an occupation. Consistent with the model predictions, our empirical evidence confirms the role of education as a key mechanism in shaping the labor market effects of ethnic diversity.

**Keywords**: Ethnic diversity; Employment; Human capital; South Africa; Homeland.

*JEL classification*: O12, J40, Z13, N37.

*We are deeply indebted to Pramila Krishnan, Toke Aidt and Enrico Rettore for their continuous support and valuable suggestions throughout the project. We are grateful to Kaivan Munshi, Hamish Low, Sriya Iyer, Elie Murard, Arnaud Chevalier, Ingo Isphording, Mirco Tonin and Oskar Nordstrom Skans whose comments significantly improved the paper. We also thank participants in Simon Fraser University, Luxembourg Institute of Socio-Economic Research, IZA Brown Bag seminar, IZA Labour and Development Workshop, AEHN 2017 Annual Meeting and the Empirical Microeconomics Seminar at the University of Cambridge for helpful suggestions and Jake Bradley, Yujiang Chen, Stephanie De Mel, Sihua Ding, Lu Han, Jiaqi Li, Simon Lloyd, John Spray, Alessandro Tondini, Chen Wang and Jasmine Xiao for insightful discussions. We thank researchers from the University of Cape Town for their support, especially Tim Brophy and Adrian Frith who kindly provide the map of Bantustans. Financial support from the PODER program (Policy Design and Evaluation Research in Developing Countries) as part of the Marie-Curie Fellowship by the European Commission is highly acknowledged.

†University of Alicante, Spain and Institute for Labor Economics (IZA) . Email: sara.tonini@ua.es.

‡Simon Fraser University, Canada. Email: pza25@sfu.ca.

# 1 Introduction

In the list of the most diverse countries in the world, some of the wealthier economies - Belgium, Switzerland and Canada - appear alongside many poor African states and Brazil[1]. It is thus not surprising that the link between ethnic diversity and economic performance has been at the centre of a lively debate in the social sciences for the past two decades. Following the initial work by Easterly and Levine (1997), scholars have examined the costs and benefits of ethnic diversity (for a review, Alesina and La Ferrara (2005)). The majority of the literature finds a negative association between ethnic diversity and many socio-economic indicators. More ethnically fractionalised communities can experience slower economic development as measured by GDP per capita (Easterly and Levine, 1997). They may also have higher social costs which are reflected in lower levels of trust and participation in social activities (Alesina and La Ferrara, 2000, 2002), inefficient public goods provision (Alesina et al., 1997; Gomes et al., 2016) and higher inequality (Alesina et al., 2016). The ethnic cleavage may also be detrimental to the establishment of a culture of inclusiveness and tolerance which is favourable to economic growth. While the literature relying on cross-country differences in ethnic diversity seems to provide conclusive evidence on the negative relationship between ethnic heterogeneity and economic growth, the more recent contributions focusing on smaller geographical areas find a positive effect of diversity on wages and productivity (Montalvo and Reynal-Querol, 2020).

In the literature, the effect of ethnic diversity on individual outcomes, especially labour market performance, which has been shown to be of great importance in driving economic development (Anand et al., 2016), is still understudied. This paper adds to this discussion by investigating how ethnic diversity amongst black South Africans (i.e. within-African ethnic diversity) affects their labour market outcomes in the post-apartheid period. As a legacy of colonialism and the apartheid system, the labor market in South Africa is still largely stratified by race.[2] While there is a wealth of studies focusing on the racial gap in labour market outcomes and on racial inequalities in accessing employment (Allanson et al., 2000, 2002; Kingdon and Knight, 2004a, 2007), the within-African heterogeneity is mainly overlooked. We focus on how the employment rate of the black South Africans responds to the composition of black ethnic groups in the district of their residence. Post-apartheid South Africa provides a unique and interesting setting for the study of the diversity - labour market nexus. On the one hand, the apartheid government that ruled the country from 1948 to 1994 deteriorated inter-ethnic relationships by reinforcing strong ethnic identity and solidarity to prevent black ethnic groups from forming a coalition to fight against the white government (Gradin, 2014). On the other hand, the regime compressed the educational and employment opportunities of the black South Africans, imposing strict labour regulations to reserve skilled and semi-skilled jobs for Whites. Blacks were also banned from setting up their own business. Only with the end of apartheid were Africans able to freely choose where to live, their education and their occupation.

Using census data for 1996 and 2001, we examine the relationship between ethnic diversity, as

[1] Alesina et al. (2003) provide measures of ethnic, linguistic and religious fractionalization for about 190 countries. They compared data from an array of diverse sources: national censuses, Encyclopedia Brittanica, the CIA, Minority Rights Group International and a 1998 study called "Ethnic Groups Worldwide".

[2] Leibbrandt et al. (2010) provide a throughout description of the labor market's trend, institutions, and policies in post-apartheid South Africa.

measured by the universally used ethno-linguistic fractionalisation index, and labour market outcomes in a sample of more than 697,000 individuals living across 210 South African districts. We find that individuals living in a more ethnically diverse district are more likely to be employed. One challenge in interpreting this association as a causal relationship is that the level of ethnic diversity in a district is likely not random. For example, a district with more job opportunities may well attract people from diverse ethnic backgrounds. To deal with the endogeneity of the ethnic diversity measure, we implement an instrumental variable approach, which relies on the historical location of black settlements (known as "homelands"). Building upon a standard gravity model from the migration literature (Alesina et al., 2016), we assume that people's mobility decreases with distance from their homeland. In particular, our instrument exploits the fact that a district tends to host a more diverse population when it is equally distant to multiple homelands. On the contrary, a district is more ethnically homogeneous when it is relatively close to one homeland and far away from the others. Importantly, the equidistance to multiple homelands, which we use as an IV, remains a strong predictor of ethnic diversity even after controlling for the proximity of the district to the closest homeland. Our results show that a one standard deviation increase in ethnic diversity index in 1996 (2001) increases employment rate by 2.98 (4.56) percentage points, which corresponds to about 8% (13%) of the average employment rate in 1996 (2001). Importantly, this positive effect on employment is driven by an increase in high-skill employment.

To investigate the possible mechanism at play, we propose a model in the spirit of the human capital and network theories. From a theoretical perspective, ethnic-specific networks could help individuals get a job, by reducing search frictions and asymmetric information (i.e. network effect). However, while personal referrals and networking could be sufficient to find a low-skill job, they are not useful to get into high-skill occupations, for which higher education and/or training are required. If the likelihood of getting a (low-skill) job through ethnic connections is overestimated, having a large network to rely upon may disincentivize individuals from investing in human capital and, consequently, trap them in low-skill jobs or unemployment. On the other hand, people in more diverse districts, where the relative size of one's own group (social network) is smaller, are incentivised to acquire higher education. Our findings are indeed consistent with the key predictions from the model. In fact, the chance of getting a college degree and of working in a high-skilled occupation increases with the level of ethnic diversity. Moreover, an increase in one's own group size reduces the likelihood of obtaining higher education and being employed.

This paper contributes to our understanding of the relationship between ethnic diversity and economic development, by estimating the impact of ethnic heterogeneity on the labour market outcomes of black South Africans. In so doing, we are filling the gap in the literature focusing on small geographical areas for developing countries. In fact, previous contributions investigating the effect of local cultural diversity on labour market outcomes (specifically, wages and productivity) studied the case of the US (Ottaviano and Peri, 2005, 2006; Alesina et al., 2016), the United Kingdom (Nathan, 2011), Germany (Suedekum et al., 2014) and the Netherlands (Bakens et al., 2013). Additionally, we contribute theoretically to the mechanism through which ethnic diversity affects economic performance. We provide a framework in which social network and higher education are substitute in accessing employment. This could be the case in many contexts where ethnicity is a salient feature and there

are significant access barriers (for example, high college fees and little public financial support) to post schooling education. Our work also speaks to the literature on the South African labour market, by emphasising the importance of the within-African heterogeneity for economic performances and providing an additional factor to consider when studying the causes of the infamously high unemployment rate in the country.

The paper unfolds as follows. In Section 2, we provide a historical and institutional background. Section 3 is dedicated to describing the data sources. Section 4 details the empirical methodology, focusing on the instrumental variable and its validity. In Section 5, we present and comment on the results about the effect of ethnic diversity on labour market outcomes in our setting. Section 6 proposes a theoretical model with empirical evidence to explain the main empirical results and rule out some alternative explanations. Finally, we draw some conclusions and policy implications.

## 2 Institutional Setting

### 2.1 Ethnic groups in South Africa and the formation of ethnic diversity

None of the black ethnic groups are indigenous in South Africa. All of them migrated from eastern and central Africa to southern Africa starting from centuries ago, as part of the so-called "Bantu migration".

Before explaining the narratives, two concepts should be made clear. The first is "homeland" which refers to the original settlements of those ethnic groups when they first moved to South Africa. The second is "white areas" or "white South Africa"[3] which refers to places in South Africa outside those homelands. Many years after arrival in South Africa, those black people moved out of their original homelands and ended up in these "white areas" due to different reasons, mainly the pressure of conflicts with the British and Dutch colonisers as well as other ethnic groups. Therefore, "white areas" are not areas where only white people reside, but places outside original black homelands (the proportion of the black over the whole population can still be large in those "white areas").

Based on Mwakikagile (2010) and Gradin (2014), we provide historical narratives on the mass migration of ethnic groups from central Africa towards South Africa, the original settlements of these ethnic groups and the migration of these people out of their homelands to "white areas" in South Africa. The timeline about the history of the settlements and migration of the black ethnic groups outside their own settlements up to the time of South Africa's independence can be found in the upper panel of Figure 1.

The indigenous groups in South Africa are San and Khoikhoi (both are "coloured" groups) residing in the southwestern and southeastern coast about 2000 years ago. Around 700s A.D., black Africans had settled in the northern part of what is South Africa today.[4] They were members of different Bantu

[3] It became an official terminology during the Apartheid regime.
[4] Some argue it is as early as the third century (Gradin, 2014).

ethnic groups who had moved southward from East-Central Africa (the Great Lake district around Congo) and spoke related languages.

Ethnicity-specific information on the Bantu migration from eastern and central Africa towards South Africa and the formation of ethnic diversity in "white areas" are summarised in Appendix A. The table contains information on the timing of their migration into South Africa, geographical location of original homelands, timing of migration outside homelands and the Bantustans assigned to them during Apartheid (which will be explained in constructing our instrumental variable). For example, Zulu are believed to be descended from a leader named Zulu born in the Congo Basin area. In the 16th century, they migrated to the south and eventually settled in the eastern part of South Africa, an area now known as Kwazulu-Natal. The Zulu empire in the 1800s witnessed their vast migration and expansion of territory.

One indication from the narratives is that the black had settled in the country long before Europeans arrived. For example, the diaries of shipwrecked Portuguese sailors attest to a large Bantu-speaking population in present-day Kwazulu-Natal by 1552. In 1652 Jan van Riebeeck and about 90 other people set up a permanent European settlement as a provisioning station for the Dutch East India Company at Table Bay on the Cape of Good Hope, beginning the era of European colonisation.

Due to the pressure from the potential conflicts with white colonisers and the other ethnic groups, the nine black ethnic groups began to move out of their homelands or change their territories. By the early 1700s, there were already some African groups migrating into the interior of the country to shield themselves from European domination. By 1750 some white farmers, known as Boers, expanded to the region where they encountered the Xhosa and Zulu. Starting from 1789, a series of wars and conflicts over land and cattle ownership broke out between the Boers and the black ethnic groups. In early 1800s the British replaced the Dutch at the Cape as the dominant force. The Boers, defeated by the British, migrated eastwards into today's Kwazulu-Natal and Free State where the conflicts between the Boers and Zulu people continued. Many other ethnic groups have encountered similar conflicts.

The destination of their migration is not well-documented. This information, however, can be reflected from today's distribution of ethnic groups across South Africa. This pattern of migration will also affect today's distribution of ethnic diversity. For example, a place would be more diverse potentially if more ethnic groups moved in. Details will be shown in the next section. One thing which needs to be emphasised here is that in most of the cases the key driving force of emigration from ethnic homelands is the conflict either with the white or with other ethnic groups rather than the economic benefits in the destination.

Importantly, further evidence shows that the mass migration both from central to southern Africa and from homelands to "white areas" within South Africa took place mainly before the spur of industrialisation and modern economy. The discovery of mineral resources is a milestone in the economic development and transformation towards modern South Africa. Diamonds were first discovered in 1867 along Vaal and Orange rivers, and in Kimberley in 1871. In 1886, gold was first discovered in Witwatersrand, around today's Johannesburg, which stimulated trade and construction in large di-

mensions. All this took place after the Bantu migration. This means the migration from homelands to "white" areas, although not completely random, may not be purely driven by the economic prospects in the destination.

In 1910 the Union of South Africa was established, which declared the superior socio-economic status of the white politically and created a white-dominated society. Since then racial discrimination has been a prominent feature of South African society even before the official institution of Apartheid, and the mobility of the black was largely restricted.

### 2.2 The role of apartheid in shaping inter-ethnic relations and labour market outcomes

Since mid-1900s, inter-ethnic relationships and labour market outcomes have been significantly shaped by the apartheid regime and related regulations. The regime reinforced the ethnic identity and destroyed much of the path dependence in the opportunities for education and labour market for the black. The timeline of the Apartheid regime can be found in the lower panel of Figure 1.

Starting in 1948, the ruling Afrikaner National Party (NP) implemented a program of *apartness* and formalized a racial classification system, which transformed into official *Apartheid* by the 1951 *Bantu Authorities Act* and 1953 *Bantu Self-Govern Act*. Each individual living in South Africa belonged to one of the four races (White, Indian, Colored, Black), which essentially defined an individual's social and political rights. In addition, the government over-emphasised the differences among the various ethnic groups, in the spirit of the *"divide et impera"* principle. The ethnic segregation, on top of the racial separation, was to guarantee the political and economic supremacy of the white minority. This exacerbated division of ethnic groups served as a tool for the white to control the black in an easier way (Gradin, 2014).

With the introduction of the *Promotion of Black Self-Government Act* in 1959, the government delimited a number of scattered rural areas as "native reserves" for blacks (called "Bantustans"), one for each ethnic group. The designated areas for the reserves amounted to 13 percent of the total South African territory, while the blacks accounted for more than 75 percent of the total population. Blacks' land ownership was restricted, as well as their ability to freely move and settle in the white South Africa. Internal migration was severely regulated until the repeal of the *Pass Laws Act* in 1986. With the forced removal of the blacks from the "white areas" of South Africa, the Bantustans became over-densely populated territories, where land was overgrazed and afflicted with serious soil erosion. The economic development of these reserves never materialized, leaving their inhabitants in acute poverty (Christopher, 2001). In 1970, the regime promulgated the *National States Citizenship Act*, which provided citizenship to blacks in their homelands. The ultimate aim was to create a number of ethnicity-based independent states.

In conclusion, the apartheid regime used separation along racial lines and ethnic lines as a fundamental device for the demarcation of physical and social boundaries for all interactions. It also severely limited the job opportunities and resources among the black (Posel, 2001). The *Bantu Education Act* of 1953 ensured that non-whites received a substandard quality of education, while access to occupa-

tion was regulated by the 1956 *Industrial Conciliation Act*. Whites were authorized to determine the racial allocation of jobs (Mariotti, 2012) and to reserve certain professions for themselves, especially in the manufacturing sector. In particular, the black were banned from semi-skilled and skilled occupations. Similarly, blacks were not allowed to run their own businesses in white areas. In fact, only with the advent of the democracy, in 1993, non-whites were able to make their free occupational choices. This, together with the reallocation of industries, changed the industrial and occupational structures in white areas, which partly weakened the path-dependence in regional demand for black labour.

### 2.3 Labour market in post-apartheid South Africa

High unemployment rates and large proportion of discouraged workers remain important issues in the South African labour market in the post-apartheid era (Bhorat and Oosthuizen, 2005; Leibbrandt et al., 2009). Based on 1996 census data, over 60 percent of the working-age black population are either unemployed or out of the labour force. Skill-biased technological changes lead to an increase in capital-labour ratio in the late 1980s and the whole 1990s, further reducing demand for unskilled labour. At the same time, real wage has been stable or decreasing between 1995 and 2005 (Banerjee et al., 2008). The increase in the supply of unskilled labour, together with the shrinkage in labour demand due to skill-biased technical change largely leads to this persistent unemployment issues in the contemporary South African labour market (Banerjee et al., 2008). Furthermore, there is a very low informal employment rate in South Africa, which is only 7.7% - 9.7% based on various measures of informality in September 2004 Labour Force Survey (Heintz and Posel, 2007), possibly because there are also entry barriers in those informal sectors (Kingdon and Knight, 2004b). This means that the formal wage-employed sector is still the main force in absorbing increased labour supply.

There is, however, large heterogeneity among different ethnic groups. In general, groups with medium and small sizes are more active in the labour market and more likely to be employed, both in self- and wage-employed jobs. This indicates that groups with smaller size are in general more active in the labour market and more competitive in job search, which can be explained by the theoretical model that we provide in the paper.

## 3 Data

For our empirical analysis, we make use of census data. There are three years of census data in the post-apartheid era: 1996, 2001 and 2011, all of which are a 10% sample of the population universe. The classification and boundary of magisterial districts have changed dramatically after 2001, making it less reliable to match the new system of magisterial districts in 2011 to the older ones. Moreover, in publicly available 2011 census data, there is no information on which magisterial district each individual resides in. For these reasons, we use data from 1996 and 2001 censuses only.

The unit of analysis is the Magisterial District (MD). There are 354 magisterial districts in South Africa, with an average territory size of 3447.5 $km^2$ and an average population size of 0.1 million in 1996.

It is particularly convenient to use the MD as a small-scale geographical unit for comparative analysis, given that all the other administrative divisions have been revised and re-demarcated repeatedly since the first democratic election in 1994. It also provides a reasonably large geographical unit to define labour market. Our final sample consists of 210 districts in 2001 census (205 in 1996 census), which are the "white" areas outside the historical homelands. Take 2001 census as an example. The excluded districts are either part of the homelands and thus had distinct political status and partially different laws and labour market regulations (124 districts)[5], or districts where the black population in 2001 accounted for less than 1% of the overall population (11 districts[6]), or they cannot be matched with 1985 census data that is explored in the instrumental variable approach (9 districts).[7]

**Status in employment.** In both 1996 and 2001 census data, we construct an individual-level binary variable for unemployment. **The dummy takes value 1 if one is unemployed or economically inactive and 0 if one is employed (either self-employed or an employee)**. Among workers who are employed, we also consider the allocation of them between self-employment and wage-employment jobs. More in details, an individual is considered to be self-employed if s/he declares to be either self-employed, an employer or work in the family business. To do this, we create another dummy variable only for employed people. It equals 1 if one is self-employed and takes value 0 if s/he declares to be an employee. We only consider working-age black population (15-64 years old).

**Ethnicity.** Following Amodio and Chiovelli (2018), the ethnolinguistic group each individual belongs to is identified using the information on the first language they speak in the 1996 and 2001 census. There are nine black ethnic groups in the country: Xhosa, Zulu, Swazi, Ndebele, North Sotho, South Sotho, Tswana, Tsonga, and Venda. Following Desmet et al. (2012), we rely on Lewis' *Ethnologue* tree of ethnolinguistic groups (Lewis et al., 2009) to build our measures of ethnic diversity.[8] For each magisterial district and census year, we calculate the relative shares of each ethnic group within the black population and combine them into ethnic diversity index: the *fractionalisation index*.[9] Universally used in the empirical literature on ethnic diversity (Desmet et al., 2017; Easterly and Levine, 1997; Alesina et al., 2003; Alesina and La Ferrara, 2005), the ethno-linguistic fractionalisation index (ELF) is a decreasing transformation of the Hirschmann-Herfindahl concentration index and is defined as

$$ELF = 1 - \sum_{k=1}^{m} s_k^2$$

where $s_k$ is the population share of ethnolinguistic group $k$ and $m$ is the overall number of groups. Intuitively, the index measures the probability that two individuals who are randomly drawn from the population belong to different ethnic groups. A larger value of the fractionalisation index indicates higher level of diversity in the magisterial district.

[5]The boundary of the homelands does not coincide with the boundary of contemporary MD. Taking a conservative method, we define district with less than 10 % overlap with homelands as "white" districts.

[6]This figure is 16 in 1996 census data, which is why the total number of districts of our interest is 205 in 1996.

[7]OLS regression results remain unchanged if we include the nine districts which cannot be matched with 1985 census data.

[8]The nine black ethnolinguistic groups of South Africa belong to the Niger-Congo language family and correspond to level 11 in the tree of ethnolinguistic groups.

[9]We consider another index: polarization index in the robustness check. It has been proved that fractionalisation index performs better in explaining economic outcomes than polarisation index (Alesina et al., 2003).

Figure 2 shows how ethnic diversity, measured by the ELF index, is distributed in the districts of our interest in 1980, 1985, 1996 and 2001. Districts in darker colours are those with higher ethnic diversity. There is large variation in ethnic diversity levels across South Africa. In general, districts in the northeastern part of the country are more ethnically diverse than those in the southwestern part. In addition, some districts in the middle part of the country are the most ethnically diverse ones. These patterns will be explained when we construct instrumental variables. Districts coloured in white are those inside original homelands, with less than 1% of the black population or that cannot be matched to 1985 census data. A cross-year comparison shows that the degree of ethnic diversity in these districts is very stable. The patterns are extremely similar between year 1996 and 2001. The spatial distribution of ethnic diversity during Apartheid (1980 and 1985) is slightly different but places with higher (lower) degree of diversity remain ethnically diverse (homogenous) over time. This reveals that the formation of ethnic diversity is a historical event and not largely driven by contemporary migration. A comparison between 1980 and 1996 (or 2001) confirms that the Apartheid regime did not drastically shift the spatial distribution of ethnic diversity.

**Demographic, socio-economic and geographical controls.** From the censuses, we also derive a number of controls, which we introduce in our regressions either at the individual level or as aggregated information at the district level. Individual characteristics include gender, age, educational attainment, marital status and whether one's father is alive. Among the district-level controls, we consider population density, proportion of the blacks, proportion of people working in manufacturing and service sectors, whether the district is mainly rural or urban, and whether there is a river and road crossing the district. Additionally, we introduce other geographical factors, which can potentially shape the economic activities of a region. Starting from the Mineral Resources Data System[10], we compute the density of mine for each district. Our geographical unit here, magisterial district, is large enough to capture activities related to the mining sector. Furthermore, the density of mine has two advantages over a simple dummy for the presence of mining activities. Firstly, it takes into account the number of mineral resources in each district as the magnitude of the effect of mines can increase with the number of mines available at the district level. Secondly, it captures the fact that mineral resources have larger economic effects in more condensed districts either due to higher population density or lower travel cost to the mines. In order to account for the agricultural suitability of land, we use the measure of terrain ruggedness from Nunn and Puga (2012).[11] We also include the measure of soil quality as another proxy for agricultural suitability. Data comes from the Harmonized World Soil Database from the Food and Agricultural Organization of the United Nations. It is a discrete index ranging from 1 to 7, with a descending order of soil quality.[12] As a proxy for the economic development at the local level, we use the National Oceanic and Atmospheric Administration night-time light satellite images data for 1996 and 2001 (Michalopoulos and Papaioannou, 2013).[13] We also

[10]Mineral Resources Data System, MRDS, is a collection of reports describing metallic and nonmetallic mineral resources throughout the world. Spatial data is available at: https://mrdata.usgs.gov/mrds/.

[11]We also tried the measure of slope from the same data source. The results are very similar. We do not include ruggedness and slope at the same time as they are highly correlated (the correlation is larger than 0.9), which potentially leads to multicollinearity issues in regressions.

[12]In the soil quality index, 1 = No or slight limitations; 2 = Moderate limitations; 3 = Severe limitations; 4 = Very severe limitations; 5 = Mainly non-soil; 6 = Permafrost area; 7 = Water bodies.

[13]Night-light data is at 30-second grid level. Here we take the average night-time light density within each magisterial district by summing up the night-light measure over these grids and dividing it by area of the district.

include the number of conflicts in each district as it has been proved to be correlated with ethnic diversity (Amodio and Chiovelli, 2018) and potentially affects economic prosperities. "Conflicts" here incorporate violence outside the context of a civil war, including violence against civilians, militia interactions, communal conflict, and rioting. A detailed discussion of conflicts in post-Apartheid South Africa can be found in Amodio and Chiovelli (2018).

The rationale of taking into account these control variables is to control for the main drivers of economic development especially employment which are correlated with ethnic diversity. A detailed discussion is in the section about empirical model specification. Details on the sources of data and methods in constructing district-level control variables are presented in the Appendix B.

Before looking into the data, it is worthwhile to point out some differences in information collected in 1996 and 2001 census. Firstly, 1996 census distinguishes between those who are unemployed and out of labour force (i.e. economically inactive) while 2001 census combines these two categories. We thus conduct analysis separately as well as jointly for these two groups in 1996 data, and compare the results based on the joint group with the corresponding results using 2001 census.

Secondly, we also explore labour market outcomes other than employment status to enrich our analysis on South African labour market, including wage, income and working hours. Information on working hours is only available in 2001 census data. We thus focus on 2001 census in calculating hourly income. A drawback of the income information in the census data in both years is that it calculates income from all possible income sources, including labour market income, social grant and other sources like bonus, rent or interest. As a result, another dataset (i.e. Labour Force Survey) is required for a more precise measurement of wage, which will be discussed in the empirical results.

The negative correlation between unemployment and ethnic diversity at the district level is further confirmed in Figure 3 where we plot the proportion of unemployed (including economically inactive) people over the whole working-age black population against ethnic diversity in each district. The downward-sloping line between these two variables is observed in both 1996 and 2001.

# 4 Empirical Methodology and Specification

## 4.1 Baseline model specification and potential bias

We study the relationship between ethnic diversity among the black population living in "white areas" in South Africa and their labour market outcomes. In particular, we examine whether the within-black ethnic diversity affects blacks' employment opportunities. We start by examining the cross-sectional evidence and investigating the relationship separately for year 1996 and 2001. For both of the years we specify our linear probability model as follows:

$$Empl_{ikdp} = \alpha + \beta ELF_{dp} + \gamma \mathbf{X}_{ikdp} + \delta \mathbf{Z}_{dp} + v_{ikdp} \tag{1}$$

where $Empl_{ikdp}$ is a dummy variable for the labour market outcome for individual $i$ of ethnicity $k$ in district $d$ in province $p$, taking value 1 if one is unemployed or economically inactive, and 0 if employed. We also report the results for wage-employment, self-employment (including self-employed, employer and working in the family business) and the substitution between wage-employment and self-employment within the subsample of the employed people. $ELF_{dp}$ takes the value of the within-black ethnic diversity index (i.e. fractionalisation index computed in Section 3[14]) in district $d$ in province $p$. $X_{ikdp}$ is a vector of individual-level characteristics (age, gender, educational attainment, marital status, whether one's father is alive which is a proxy for family financial and non-financial support). $Z_{dp}$ is a set of both time-varying demographic and economic controls and time-invariant geographical characteristics at the district level, which will be explained in more detail below. Standard errors are clustered at the district level to allow for correlation of the error term across individuals in the same district.

Unobservables which potentially affect employment rate are included in the term $v_{ikdp}$. $v_{ikdp}$ can therefore be decomposed into the following items:

$$v_{ikdp} = \theta_p + \lambda_k + \epsilon_{ikdp} \tag{2}$$

$\epsilon_{ikdp}$ is the random error term. $\theta_p$ is province fixed effect which mainly controls for historical path dependence in job opportunities in each province, as well as province-level fiscal variables including social grant provision and policies on taxation and redistribution. There is also evidence that there is inequality between ethnic groups (Alesina et al., 2016) and that the gaps between different ethnic groups lie in their demographic structure, location, education and labour market outcomes (Gradin, 2014). Therefore we introduce $\lambda_k$, ethnic group fixed effects, which allows us to control for mechanical compositional effect and ensures we are comparing individuals from the same ethnic group across districts exposed to different levels of diversity.

We include a rich set of district controls $Z_{dp}$ to limit the information in unobserved items. To account for market size effects, we introduce the population density and urban/rural status of the district. As proxies for local economic development, we use the average night-time light density across 30-second grid areas within each district, and the share of blacks in the district population. For the industrial structure of the district which potentially leads to differences in labour intensity of firms, we control for the proportion of people employed in manufacturing and service sectors. Furthermore, to control for the direct spillover from homelands, we include the distance to homelands. To control for the potential cost of ethnic diversity like conflicts, we add the number of violence in each district in the corresponding years, as conflict has been proved to be associated with ethnic diversity (Amodio and Chiovelli, 2018)) and potentially job opportunities for the blacks (for example, there might be more closure of factories in more turbulent districts). Finally, to control for agricultural suitability and other geographic factors relevant to the local economic activities we use the terrain ruggedness, the existence of a river and a road crossing the district and the density of mineral resources.

[14]We use the results about polarization index as a robustness check.

**The relationship between ethnic diversity and labour market outcomes can also be investigated at the district level. Then model (1) would change accordingly. $Empl_{dp}$ would represent the proportion of individuals who are unemployed, wage employed and self-employed in district $d$ in province $p$ and the ethnicity fixed effect would be removed. The set of individual characteristics $X_{ikdp}$ should therefore be aggregated at the district level (e.g. average education in each district). The district-level regression becomes**:

$$Empl_{dp} = \alpha + \beta ELF_{dp} + \delta \widetilde{\mathbf{Z}_{dp}} + \theta_p + \epsilon_{dp} \tag{3}$$

Here $\widetilde{Z_{dp}}$ include both the individual-level variables in $X_{ikdp}$ aggregated at the district level, and the original district-level variables in $Z_{dp}$. Similarly, after controlling for province fixed effect $\theta_p$, the remaining items in $\epsilon_{dp}$ are still sources of omitted variable bias which will be dealt with using the same instrumental variable approach.

As individual-level regressions contain more information (especially ethnic-specific characteristics captured by ethnicity fixed effects), we mainly report results based on individual-level regressions in our analysis whilst presenting the results of district-level regressions for robustness check.

### 4.2 Instrumental variable approach

Our instrument for ethnic diversity exploits the historical origins of the location of blacks' homelands. As explained in the institutional setting, the nine black ethnic groups moved long ago from the northern territories of the African continent and settled in different regions of today's South Africa, with one ethnic group occupying one settlements (i.e. defined as "homelands"). Let us assume that the magnitude of migration from the homelands to outside districts decreases with the distance between them and distance is the only determinant in migration. When they moved out of these homelands to the outside districts (i.e. "white" districts which we are focusing on in this paper), the territories that are equally distant to multiple homelands are more likely to be inhabited by individuals with different ethnic origins, and therefore the ethnic diversity will be the highest. On the contrary, places close to only one homeland and far away from the rest become ethnically homogeneous as they have one group dominant in population size migrating from the closest homeland. Visually, this prediction is confirmed by the distribution of ethnic diversity in South Africa in 1996 (Figure 2). As shown before, places with relatively higher level of diversity are not necessarily places at the border or close to economic centres of the country, but are those in the middle and northeastern part of the territory surrounded by multiple homelands. Furthermore, districts on the far western part of the country present reasonably high level of ethnic diversity although being far away from all homelands. This is because these districts are still equally equidistant from all the homelands.

We therefore need an instrument that captures the equidistance of each district to all the original homelands. Our instrumental variable strategy proceeds in two stages. First, similar to Alesina et al. (2016), we estimate a parsimonious gravity model of migration based on 1985 census data (i.e. pre-

1994 distribution of ethnic groups). We aim at predicting the level of within-black ethnic diversity in each white district $d$, solely as a function of a factor that is plausibly exogenous to labour market outcomes of the blacks: the distance of the district to the homelands. Second, we start from the predicted stocks to construct a diversity index. Specifically, we estimate:

$$N_{dk85} = \alpha + \beta_1 Dis_{dk} + \gamma_k + \epsilon_{dk85} \tag{4}$$

where $N_{dk85}$ is the actual stock of individuals belonging to ethnic group $k$ in (white) district $d$ in 1985; $Dis_{dk}$ is the bilateral Euclidian distance between the centroid of district $d$ and the closest border of homeland for ethnic group $k$[15]; and $\gamma_k$ is the homeland fixed effect. The determinants in our model are the ones traditionally used in the related literature (Mayda, 2010; Beine et al., 2013; Ortega and Peri, 2014; Dumont et al., 2010). In particular, the physical distance between two districts (the homelands and the white areas) accounts for the migration costs, while the homeland fixed effects take into account common shocks in living conditions in the original settlement and the stock of population of each ethnic group in homelands, which can also influence migration decision. Following Santos Silva and Tenreyro (2006), we estimate the model by using the pseudo poisson maximum likelihood (PPML) estimator, which better suits the count data in the dependent variable.[16]

By imposing a universal $\beta_1$ to all ethnic groups, we assume that the per-unit migration cost is the same for everyone, regardless of their ability and ethnicity. In addition, by ignoring any characteristics of the destination (e.g. population size, economic development and job opportunities) in the gravity model, we impose the condition that the benefit of migration is the same for everyone. Therefore by construction our predicted number of migrants from each homeland is only determined by the distance between homeland and destination.

In principle, the migration stocks could be predicted by 1996 and 2001 data. Nevertheless, we prefer to use the 1985 census data to rule out the selection of migration resulting from the movements of the black population after 1994 (this happened even as early as the repeal of the Pass Law in 1986). In fact, as previously documented (Section 2), the blacks had very limited freedom in choosing their own residential location and were strictly regulated in inter-district migration before 1986. After 1986 these constraints were loosened and the blacks had some freedom to decide where to resettle. Therefore, the distribution of ethnic groups in 1985 is less affected by the simultaneous change of labour market conditions and blacks' selection into "white areas" in the post-Apartheid era. Another reason why we use the 1985 distribution of the black population is that the equidistance to different homelands is a feature which stays stable over time. By sticking to 1985 data we can construct an instrumental variable whose value stays the same between 1996 and 2001 to make the IV regression

[15]The reason why we use the centroid of the districts instead of capital city is that capital cities are not well-defined at the magisterial district level. We use the border instead of the centroid of the homeland because the shape of the homeland is highly irregular and scattered. Furthermore, the distribution of population within homeland is highly uneven, making the centroid of homeland a less reliable measure in capturing the distance between the destination and the location of potential migrants from homeland.

[16]We do not control for the population size in the destination in the gravity model as it might be endogenously determined by the level of economic development in the destination which potentially affects the flow of migrants into the destination. Here our aim is not to get the most precise estimate of bilateral migration but to construct the counterfactual number of migrants in each district under a hypothetical setting where bilateral migration is only determined by distance between the original homeland and destination.

results in these two years more comparable.[17]

Using the predicted stocks $\widehat{N_{dk}} = \widehat{\alpha} + \widehat{\beta}_1 Dis_{dk} + \widehat{\gamma}_k$, we calculate the predicted share of ethnic group $k$ in the black population of district $d$ and construct the instrument for the fractionalisation index $ELF$:

$$\widehat{ELF} = 1 - \sum_{k=1}^{m} \widehat{s_k}^2 \quad \text{with } \widehat{s_k} = \frac{\widehat{N_{dk}}}{\sum_{k=1}^{m} \widehat{N_{dk}}} \tag{5}$$

The same instrumental variable approach with the same model specification at the first stage can be applied to district level regressions.

The remaining challenge is to find a proper measure of the original homelands for each ethnic groups. As there is no document about the exact location and boundary of these homelands, we use the territories of Bantustans during Apartheid as proxies for these original homelands. As is discussed in the institutional setting, with the ascent of the apartheid regime, the white-dominated government of South Africa designated specific territories as pseudo-national homelands (i.e. "native reserves", called "Bantustans" in the official documents) for the country's black African population. The Bantustans were organized on the basis of ethnic and linguistic groupings and were a major administrative device for the exclusion of blacks from the "white areas" of South Africa. The location of the Bantustans is based on the government's knowledge and documents about the historical location of homelands of each ethnic group. Ten Bantustans were created for these nine ethnic groups (There are two Bantustans for Xhosa people - Transkei and Ciskei. Other groups each occupies one Bantustan).[18]

To verify that the location and territory of Bantustans can be treated as proxies for the original homelands for the black people, we compare the distribution of these Bantustans and the "Murdock map". This map, drawn by an anthropologist George Murdock in 1953[19], provides the information on what the dominant ethnic group is in each geographical unit on the map of the whole African continent at the end of the $19^{th}$ century. As reflected in the Murdok's map (panel (a) in Figure 4) (each colour represents a certain group dominating the corresponding place in terms of population size), up to the end of the $19^{th}$ century, each of the nine groups has occupied some specific areas of the country. The Murdock map reveals the distribution of dominant ethnic group in each geographical unit rather than the exact location of original homelands. And the boundary of the geographical units on this map does not coincide with the border of magisterial districts in South Africa. Therefore, the Murdock map can only roughly implies the spatial distribution of each ethnic group in South Africa, which is a result of both the distribution of original homelands and the migration of ethnic groups from these original settlements to other places.

[17] We do not find much variation in fragmentation index between 1996 and 2001, which means ethnic diversity stays relatively stable over time.

[18] Therefore we treat Transkei and Ciskei as one homeland in the gravity model. When we calculate the distance between each district and the original homeland of Xhosa people, we measure the distance between each district and Transkei and Ciskei respectively and choose the smaller one.

[19] The map has been digitized by Nathan Nunn, starting from "Tribal Map of Africa" which is a fold out map from the book "Africa: Its peoples and Their Culture History" by George Murdock, 1959.

Comparing Murdock's map in panel (a) and the distribution of Bantustans under the Apartheid system in panel (b) in Figure 4, we can find large overlaps between the Bantustan and the region where the same group have dominated historically in Murdock's map. For example, places around the Bantustan designed for Tswana people (the dark green part in panel (b)) are also the places dominated by Tswana people (labeled with the same dark green colour) at the end of the $19^{th}$ century in Murdock's map in panel (a). Therefore, it is reasonable to use the distribution of Bantustans as proxies for the location of original ethnic homelands.

The map in Figure 5 presents the value of predicted diversity index together with the distribution of Bantustans across the country. The white places with slashes are either places which cannot be plausibly considered as "white" South Africa of our interest as they have more than 10% overlap with Bantustans, or places which cannot be matched with 1985 census data. The spatial pattern of predicted value of ethnic diversity in this figure is similar to the distribution of ethnic diversity in Figure 2 based on the real data. Again, places with the highest predicted ethnic diversity are those amid multiple homelands (mainly in the middle and northeastern part of the country). A more important feature is that the distance to the closest homeland (proxied by Bantustans) does not completely determine the level of predicted ethnic diversity. That is to say, places close to a specific Bantustan (and far from the other ones) may not be highly diverse. It is particularly the case for the districts around the Bantustans of Transkei, Ciskei, Kwazulu and Bophuthatswana.

#### 4.2.1 Test of validity of the instrumental variable

Identification requires the instrument to capture the ethnic diversity pattern observed in 1996 and 2001 and to be uncorrelated with any other determinants of the blacks' labour market outcomes. The first condition is satisfied provided that: 1) The historical distribution of ethnic groups within the country varies with and is closely related to the distance of the destination region ("white" district) from multiple Bantustans, and 2) Apartheid did not overturn the historical pattern. As for the second condition, the non-randomness of blacks' homelands could cast doubts on its fulfillment. The proximity to the Bantustans might well be correlated with unobserved factors other than diversity, affecting the blacks' labour market outcomes.

However, the instrument exploits the distance to *multiple* ethnic homelands as a predictor for diversity. As mentioned above, the map in Figure 5 shows that districts with higher predicted diversity are the ones that are "equally" distant to multiple homelands, and not necessarily the ones that are the closest to a specific homeland. For example, although being contiguous to one of the Bantustans - Transkei (identified with the red color in Figure 5), districts in the South-East are among the most ethnically homogeneous areas because they are located at the periphery of other homelands. To further ensure that the instrument only captures the relative distance to multiple homelands and not the proximity to a single Bantustan, in the regressions we control for the distance to the closest homeland. As all the homelands are located at the eastern part of the country, controlling for distance to the closest homeland can also deal with the problem that the instrumental variable might purely capture the west-east division of the country.

We argue that, conditional on proximity to a single homeland, the distance to multiple homelands is as good as random. The most direct narrative evidence is that according to the timeline in the institutional setting, the mass migration of the black largely occurred before the discovery of mines, rise of industrial sectors and modern development. This means the migration from homelands to "white" areas is not purely driven by the higher economic prosperity in the destination.

For a more rigorous test of the validity of our instrumental variable, we run regressions to show that the predicted ethnic diversity index is not correlated with potential confounders which determine ethnic diversity and employment simultaneously, conditional on all the control variables in our first stage regressions. Firstly, we test the correlation between the instrumental variable and potential job opportunities. According to agglomeration economics, economic centres, as clusters of economic activities, business and capital inflow, may act as the hub of job creation. Therefore, distance to economic centres may capture the potential job opportunities an individual is exposed to, based on the spillover of economic prosperity from the economic centres. There are five main economic centres in South Africa: Cape Town, Pretoria, Durban, Port Elisabeth and Johannesberg. In the validity test we calculate the distance from the centroid of each magisterial district to the closest economic centre and correlates it with predicted fragmentation index discussed above.

The second potential confounding factor is the economic activity of the white. On the one hand, as the Apartheid regime destroyed the self-employment opportunities, leadership and the training towards skilled occupations of the black in the "white" South Africa, the majority of the employers of wage-employed black people are the whites. Although our main regressions focus on the blacks, the population size and the employment status of the whites are also important in determining black people's employment rate, as they might be the providers of potential jobs to the black workers. On the other hand, the dominance and wealth of the white might potentially affect the migration decision of the early black migrants. Black people from different ethnic groups may move to a district where the white behave relatively better as there are more opportunities (or poorer as there is less stress/competition from the white) and thus the ethnic diversity of the black might be correlated with the behaviours of the white. We then calculate the employment rate of the white among their working-age population for each magisterial district in our sample and see if it relates to ethnic diversity of the black.

Thirdly, path dependence also matters in determining contemporary employment opportunities. As the distribution of black settlements is not completely random, the equidistance to multiple original settlements might reveal some socio-economic characteristics besides the distance itself (i.e. customs, early conflict or the distribution of ancient civilisations) which have long-term impact on contemporary development. This persistence of particular socio-economic features is usually a concern in literature which constructs instrumental variables with geographical variables. However, in our special setting, the Apartheid regime before our sample period compressed the opportunities of education, job opportunities and residential choice nationwide among the black and potentially destroyed part of such historical path dependence. If we can show that the path dependence which potentially correlates with equidistance to homelands was largely destroyed by the Apartheid regime due to the shift in residential patterns and the re-allocation of economic activities both for the black and the white, we

will be safer to claim that the historical persistence is not likely to affect contemporary employment opportunities directly. As there is no reliable data to reveal the employment pattern of the black during apartheid, we use the employment pattern of the white in 1980 as a proxy for the remaining path-dependence in employment close to the end of the apartheid and see if it correlates with our instrumental variable measured with 1996 and 2001 data. For the employment status of the white in 1980, we do not consider self-employment as the definition of self-employment is not quite clear under Apartheid regime and therefore has large measurement errors.[20] We also consider the population size of the white in 1980.

The fourth potential confounding factor is the magnitude of contemporary migration. Although historical migration was not mainly driven by economic prospects, it might still be the case that contemporary diversity results from contemporary migrants which are driven by economic opportunities. Therefore, we need to show that our predicted diversity does not relate to the magnitude of contemporary migration which refers to cross-district migration ever happening in one's life in 1996 census and cross-district migration between 1996 and 2001 in 2001 census.

Table 3 shows the results on the validity of the instrumental variable based on 1996 and 2001 census data. We regress a set of variables that potentially affect employment rate on predicted fractionalisation index conditional on all the control variables in the main regressions discussed above. Panel A, B, C and D present the tests on the relationship between predicted ethnic diversity and job opportunities, economic activities of the white, path dependence and contemporary migration, respectively. We obtain the coefficients of the tests by regressing the corresponding dependent variables (as reported in the table) on predicted ethnic diversity conditional on all the control variables in the main regression. These dependent variables include: distance to the closest economic centre, proportion of white people who are self-employed over the white population in 1996 and 2001, proportion of white people who are employees over the white population in 1996 and 2001, proportion of white people over the whole population in 1996, 2001 and 1980, proportion of white people who are employees over the white population in 1980 and the number of contemporary migrants in each district. We do not find systematic relationships between these potential confounders and our instrumental variable, which means that the predicted ethnic diversity can be considered as a valid instrumental variable.

#### 4.2.2 Other potential threats to the instrumental variable

This section discusses some remaining potential threats to the instrumental variable which are not likely to be measured with available data.

Firstly, one may argue that the original distribution of ethnic homelands is not completely random. The fact that one place is close to multiple homelands might mean that these homelands are themselves close to each other. Similarly, one possible pre-requisite for a place to be close to only one homeland is that those homelands might be scattered and relatively far away from each other. If the whole

[20]There are four censuses during Apartheid: 1960, 1970, 1980 and 1985 census. We only consider 1980 census as the data quality is higher than that in 1960 and 1970 census. Publicly available 1985 census data has no information on employment status.

region is equipped with better endowments (geography, climate or soil quality) than the others at the time of the Bantu migration from central Africa, this place could attract more than one ethnic groups to establish their homelands, whilst regions with only one ethnic homeland or regions where the distribution of homelands is more scattered might be less attractive in resources and endowments. Therefore, our instrumental variable - the predicted diversity index might just capture the distribution of homelands and the original endowments of the whole surrounding region.

This is not likely to be the case for the following reasons. The first reason is that our instrumental variable captures the equidistance to different homelands conditional on the distance to the closest homeland. By construction places far away from all homelands can still have reasonably high predicted diversity, as long as it is of equidistance to all these homelands. These places are less likely to be affected by the original endowments and resources of ethnic homelands. The second reason is that we have already controlled for geographical endowments (ruggedness, soil quality and river) in each district which are potentially correlated with their initial development by affecting their agricultural production. The third reason is that if our instrumental variable mainly captures the initial economic development and the endowments or resources of the region rather than ethnic diversity, the predicted diversity index should be correlated with the labour market outcomes among both black and white population. However, as is shown in Table 3, our instrumental variable is not systematically correlated with the employment rate of white workers. Therefore, it is unlikely that the initial endowments in the regions surrounding ethnic homelands challenge the exclusion restriction of the instrumental variable.

Secondly, there is a possibility that districts close to multiple homelands might be the trading centres for people from those homelands whilst trade flows in districts close to only one homeland are less. This might also lead to the difference between these two types of places in the initial economic prosperity and the establishment of cities resulting from trade. Here we show this is unlikely to severely violate the validity of our instrumental variable. Our instrumental variable by construction allows for the case that a place far away from all homelands can be reasonably diverse if it is equidistant to different homelands. And this place is less affected by the initial trade flows among homelands. Furthermore, places with more initial trade flows might become contemporaneous economic centres due to the path dependence in city development and the accumulation of capital and labour. In our validity test we do not find a systematic pattern of the distance to the closest economic centre and predicted diversity index.

Thirdly, one may worry that certain events which attract diverse migrants might happen coincidentally in places close to multiple homelands. For example, the homeland for Tswana group (i.e. the Bantustan of Bophuthatswana) and places in Mpumalanga and Limpopo Province (in the northeastern part of the country) are rich in mineral resources. If our instrumental variable mainly captures the distribution of mineral resources, and if the discovery of mines in a district motivates people of diverse backgrounds to migrate into the district and at the same time boosts economic development, what can be reflected in the predicted ethnic diversity is mainly the effect of mineral resources. In our analysis we have controlled for the density of the mines in each district. More importantly, narrative evidence reveals that the mass migration from central Africa (which can be dated back to the $11^{th}$ and $12^{th}$ century) and the emigration from homelands to "white" South Africa happened well before

the discovery of mineral resources (mainly starting from the $19^{th}$ century). Therefore, the discovery of mines and the related events are not likely to violate the validity of our instrumental variable.

#### 4.2.3 First stage results

Table A0 in the Appendix reports the estimated parameters of the gravity model. It suggests that the distance between a white district $d$ and an ethnic group's homeland is strongly negatively correlated with the size of the same ethnic group's population living in district $d$. Table 4 presents the first-stage regression of the instrument at the individual level. We are interested in both working-age population (age 15-64) and a subsample which have already finished full education (age 25-64). All regressions include province fixed effects and all control variables. Columns 1 and 2 (3 and 4) report the first-stage regression results based on 1996 (2001) census data. In both years the predicted fragmentation index $\widehat{ELF}$ is positively associated with the observed index $ELF$. The F-statistics are very high in all regressions (i.e. much larger than 10), indicating that the instrument is a very strong predictor of ethnic diversity. Comparing column 1 and 2 reveals that the F-statistics remain stable in both the full sample and the subsample. Comparison between columns 3 and 4 confirms the same pattern in year 2001.

District-level regressions in Appendix Table A1 reveal the same pattern. Predicted ethnic diversity is positively and strongly correlated with the ethnic diversity index in real data. F-statistics of the instrument are still large in all regressions in both year 1996 and 2001. All results consistently show that our predicted ethnic diversity index is strong enough as an instrumental variable.

### 4.3 Supplementary approach: district-level fixed effects

The fact that we have two-year cross-sectional census data and that the territory of magisterial districts stay stable between 1996 and 2001 motivate us to find a way to construct panel data at district level as a supplementary approach to the instrumental variable specification. From the district-level model specification (3), we realise that the main source of bias comes from the unobserved district-level confounders. Therefore an alternative way to instrumental variable approach to deal with this bias is to control for it directly by including district fixed effect based on a panel of districts. Therefore we construct a balanced panel by matching the magisterial districts between 1996 and 2001[21] and conduct the model (3) by adding magisterial district fixed effect directly. Any time-invariant variables in $Z_{dp}$ and $\theta_p$ are dropped automatically. Instead we add time fixed effect $u_t$ in the model.[22]

[21] Among 205 magisterial districts in 1996 and 210 districts in 2001, 205 of them can be matched, given that we exclude districts with less than 1% of black people over the whole population.

[22] A potential further specification is to combine the above two approaches and rely on fixed effect-IV approach. The rationale to do this is that some district-level unobservables might change over time which cannot be captured by time-invariant $\sigma_d$. In this case, we have the first difference specification:

$$\Delta Empl_{dt} = \alpha + \beta \Delta ELF_{dt} + \delta \Delta \widetilde{\mathbf{Z}_{dt}} + \Delta f_{dt} + \epsilon_{dt}$$

Ideally we can find an instrumental variable for $f_{dt}$. A similar case to this specification can be found in Dustmann et al. (2017). However, this first-difference specification at district level with instrumental variable is not appropriate here because there is little variation in both the real-world ethnic diversity and the predicted ethnic diversity (i.e. the

$$Empl_{dt} = \alpha + \beta ELF_{dt} + \delta \widetilde{\mathbf{Z}_{dt}} + \sigma_d + u_t + \epsilon_{dt} \tag{6}$$

We report the results of this district-level fixed effects model right after the main analysis.

# 5 Empirical Results

## 5.1 Ethnic diversity and labour market outcomes

### 5.1.1 Ethnic diversity on employment

Table 5 summarizes the main results on the effect of ethnic diversity (measured by fractionalisation index) on unemployment rate. The dependent variable is a dummy which equals 1 if one is unemployed or out of labour force and 0 otherwise (including people who are self-employed and employees). In 1996 census data which distinguishes people who are unemployed and out of labour force, we create dummies for unemployment and labour force participation and look at how they respond to ethnic diversity separately. Columns 1-6 report the results in year 1996 while columns 7-8 are for year 2001 when unemployed workers and people out of labour force are combined into one category in the original census data. Panel A in Table 5 reports the results based on the cross-sectional OLS regressions at the individual level. Panel B in Table 5 provides the corresponding estimates based on the instrumental variable regressions. We provide results both for the full sample (which gives a lower bound of the effect of diversity on employment) and a subgroup of people aged from 25 to 64 who have finished their education (which gives an upper bound of the effect of diversity on employment). All regressions control for the individual and district level characteristics as well as ethnicity fixed effects discussed above.

In most of the OLS and IV regressions in Table 5 the coefficients of ethnic diversity on unemployment (or labour force participation or these two outcomes altogether) are significantly negative, indicating that within-black diversity increases the rate of employment and labour force participation. Comparing panel A and panel B, the negative and significant coefficients of ethnic diversity remain in IV regressions in many columns. In panel B, comparing columns 2, 4 and 6 reveals that ethnic diversity increases employment mainly by decreasing the number of people who are actively looking for jobs but still unemployed, rather than bringing people into the labour force. Table 5 also shows that the coefficients of ethnic diversity are larger and more significant for the subgroup of people who have finished education than those for the full sample, which confirms that the full-sample analysis gives a lower bound of the effect of ethnic diversity.

We focus on the full sample (i.e. lower bound) to calculate the magnitude of the effects of ethnic diversity on employment based on the results in columns 5 and 7. In panel A in column 5, a one

equidistance to different homelands does not change in the time dimension) over time, which is not sufficient for reliable statistical inference.

standard deviation increase in ethnic diversity index in 1996 is associated with a 2.15 percentage point decrease in unemployment (including inactivity), which is 3.51% of the average unemployment (including inactivity) rate. Similarly, in panel A in column 7, a one standard deviation increase in ethnic diversity index in 2001 is associated with a 3.88 percentage point decrease in unemployment (including inactivity), which is 5.97% of the average unemployment (including inactivity) rate. Correspondingly, in IV regressions, a one standard deviation increase in ethnic diversity index in 1996 (2001) decreases unemployment (including inactivity) by 2.61 (4.40) percentage point, which is 4.24% (6.91%) of the average unemployment (including inactivity) in 1996 (2001).

Comparing the magnitude of estimates in OLS and IV regressions in both years shows that the magnitude of the effects of ethnic diversity on unemployment rate increases largely between 1996 and 2001 (from 3.51% of the average unemployment rate to 5.97% in OLS and from 4.24% to 6.91% in IV) and IV estimates are slightly larger than OLS estimates. This can be explained by the fact that IV regressions capture LATE for workers at the margin of being affected by ethnic diversity, who might be the most responsive to diversity levels.

Appendix Table A3 further breaks down employment status into two categories: self-employment and wage-employee. All the independent variables remain the same as those in Table 5. In columns 1 and 3 in Appendix Table A3, the dependent variable is a dummy which equals 1 if one is self-employed and 0 otherwise (including unemployed, inactive and wage employed). The dependent variable in columns 2 and 4 is a similar one which equals 1 if one is a wage employee and 0 otherwise. Again, panel A (B) reports the results for OLS (IV) regressions.

The results show that in the post-Apartheid South African context, within-black ethnic diversity has a positive effect on the labour market outcomes of the blacks, mainly in wage employment. Specifically, a one standard deviation increase in the fractionalisation index is associated with a 2.31 (3.54) percentage point increase in the wage-employment rate of the working-age black individuals in 1996 (2001), according to the OLS results. This corresponds approximately to a 6.52% (10.39%) increase of the average wage-employment rate among the black working-age population in 1996 (2001). In IV regressions, a one standard deviation increase in the fractionalisation index increases wage-employment rate by 2.74 (4.60) percentage points in 1996 (2001), which is around 7.71% (13.50%) of the average wage-employment rate in 1996 (2001).

Similar to the patterns in Table 5, the effect of ethnic diversity on wage-employment increases from year 1996 to 2001. IV estimators have slightly larger magnitude than OLS estimators for possibly the same reason. We do not find anything significant about self-employment rate.

The corresponding district-level regressions based on the model specification 3 are reported in the Appendix Table A4. In these district level regressions, the dependent variables are the proportion of working-age black people who are unemployed or inactive; who are wage-employed; who are self-employed and the proportion of people who are self-employed relative to wage-employed (columns 1-4 and columns 5-8, for year 1996 and 2001 respectively). OLS (IV) estimators are shown in panel A (B).

The OLS and IV estimates reported in Table A4 confirm the positive impact of diversity on the employment of the blacks. And this positive impact mainly takes place in wage-employment. The effect on employment (and wage-employment) in OLS regressions is slightly smaller than the one estimated with the individual-level regressions, while the magnitude of the effect in IV regressions is slightly larger than that in the individual-level regressions.[23]

#### 5.1.2 Ethnic diversity on wage, income and working hours

In this section we investigate labour market outcomes other than employment to get a more thorough picture of how labour market responds to ethnic diversity in post-Apartheid South Africa. We replicate the above individual-level regressions (both OLS and IV) by replacing the dependent variables with other labour market outcomes, including working hours, hourly wage and monthly earnings. As information on working hours is only available in 2001 census data, we only conduct these analyses based on 2001 data. For data on working hours, if values of self-reported weekly working hour are larger than 80, we treat them as outliers and exclude them from regressions. In addition, we trim the income data by excluding values above 5 standard deviation of the mean income. Hourly wage is constructed by dividing monthly earnings by monthly working hours (i.e. four times of weekly working hours).

Data on monthly income in 2001 census includes both labour market earnings and income from other sources such as dividend, rent or social grant. We first report the results based on these rough measures of monthly earnings and replicate the regressions with more precise data on labour market earnings and working hours.

Panel A in Table 7 reports the OLS and IV regression results on these labour market measures based on 2001 census data. Dependent variables include: log monthly income, log hourly earnings and weekly working hours. As self-employed workers and employees have very different determinants of working hours and earnings, and that ethnic diversity mainly increases wage-employment rate, we only focus on employees in all regressions.[24] Columns 3 and 6 indicate that ethnic diversity does not affect weekly working hours among the employees. Therefore the increase in employment in response to ethnic diversity comes from the extensive margin by increasing employability of unemployed and inactive people, rather than the intensive margin (measured by weekly working hours). And this extension of the extensive margin of labour is not achieved at the sacrifice of decreased intensive margin.

Columns 1, 2, 4 and 5 show some evidence on the increase in both monthly and hourly income among the black employees in response to ethnic diversity. As stated above, information on income in census data incorporates all potential income sources. Therefore we need another dataset which asks information on labour market earnings in particular. We turn to October Household Survey 1996 to

[23]Columns 4 and 8 report the results on the effect of ethnic diversity on the rate of self-employment relative to wage-employment by only including black people who are employed. Results in other columns are based on the whole working-age black population.

[24]There are more observations in columns 3 and 6 than others because there are missing values in income and we trim the income values above 5 standard deviation from the mean.

replicate all the results in Panel A.[25] We do not choose year 2001 because starting from year 1998 there is no information on the magisterial districts each individual lives in. The results are in Panel B in Table 7. Columns 3 and 6 confirm that weekly working hours are not responsive to ethnic diversity. In columns 1, 2, 4 and 5 the effects of ethnic diversity on measures of labour market earnings are not significant, possibly because the increase in employment can come from both the supply and demand side of the labour market, or because the measures of nominal earnings are not adjusted for price levels (as there is no price or living cost data at the magisterial district level).

### 5.2 Supplementary approach: district-level fixed effects

As a supplementary approach to the instrumental variable approach, we provide estimation results on district-level fixed effects models based on the model specification (6) in Table 8. We construct a balanced panel between 1996 and 2001 (205 magisterial districts each). The measures of labour market outcomes (i.e. dependent variables) are: proportion of people who are unemployed or inactive among the whole working-age black population; proportion of employed workers among the whole working-age black population (excluding self-employed people); ratio of the number of self-employed workers versus employees and log monthly income among employees.

Similar to the main IV regression results, larger diversity is associated with more employment, mainly in wage-employment, while there is no significant correlation between ethnic diversity and monthly income. In particular, in district fixed effects regressions we find some evidence that larger diversity is associated with a higher ratio of wage-employment in relation to self-employment.

## 6 How Does Ethnic Diversity Affect Employment: A Theoretical Model and Mechanism

We propose a theoretical framework consistent with our empirical findings above to explain the positive effects of ethnic diversity on employment and occupational choices. More specifically, we focus on human capital investment which increases with ethnic diversity. The story is as follows. In ethnically homogenous districts, people are trapped in low education and low-skilled jobs as those occupations are largely based on ethnicity-specific networks which help them with job hunting. In more diverse places, as the size of their own groups becomes smaller, they are motivated to get higher education because they cannot rely much on ethnicity-specific networks to look for jobs. Employment rates will therefore be higher in more diverse districts as the chance of being employed increases with educational attainments. People in more diverse places will also shift to high-skilled occupations as these jobs are less ethnically affiliated.

[25] It is an annual survey staring from 1993 (which was renamed as Labour Force Survey conducted twice a year from 2000 and became a quarterly survey from 2008). In 1996 survey 72890 individuals are covered, among which 16082 have information on work status.

### 6.1 Model setup

We assume each district has $n$ different ethnic groups. Any district with $k$ ($k < n$) groups will be interpreted as having $k$ groups with positive group sizes and $n - k$ groups with 0 population. This is to make sure we fix the number of ethnic groups in the district and the variation in diversity level only comes from the dissimilarity in group sizes, consistent with the empirical strategy. Ethnic group $i$ has a population share $s_i$, $i = 1, 2, \ldots n$ and $\sum_i^n s_i = 1$. The total population size in the district is standardised to 1.

Each individual $j$ in group $i$ lives for two periods. They make decisions on education in the first period and get income from jobs (or unemployment insurance if being unemployed) in the second period. At the beginning of the first period each of them is endowed with family wealth $y_0$ and an ability $\omega_{ij}$ with a uniform distribution at the interval $[0, 1]$ in all districts (which follows Munshi (2011)).[26] There are two types of educational attainments: high education (i.e. a college degree) or low education. The first one channels people to high-skilled jobs while low educated people can only work in low-skilled occupations. For simplicity we assume the demand for high-skilled (low-skilled) occupations is $H$ ($L$) and is fixed in all districts.

High education incurs a monetary cost of $S$ to each family, including the tuition fee and the transport cost of going to colleges. For each individual, the net return to college education, conditional on finding a job, is an increasing function of ability. It is because people with higher abilities either benefit more from high education or incur less opportunity cost (e.g. spending less time) in learning. This net return is denoted as $r_h\omega_{ij}$ for individual $j$ in ethnic group $i$.

In each period, individuals get utility $u(c)$ from consumption $c$. Suppose there is no saving or borrowing in all households. Consumption in the first period is $c_1 = y_0 - S$. In the second period, consumption $c_2$ is equal to the income from labour market. Individuals with high education have the exogenous probability $p_h$ in finding a job in high-skilled occupations. His consumption in the second period is therefore $r_h\omega_{ij}$ if he gets a job and $I$ which is unemployment insurance if he is unemployed. Following Acemoglu and Shimer (1999), we use the isoelastic utility function to capture the risk aversion of workers and assume that individuals always get positive utility from consumption. $\beta$ is a discount for the future. Therefore, $u(c_1, c_2) = \frac{c_1^{1-\alpha}}{1-\alpha} + \beta\frac{c_2^{1-\alpha}}{1-\alpha}$ with $0 < \alpha, \beta < 1$. This indicates that $\alpha$ is not too large, which is verified in Brick et al. (2012) in their research on South Africa. They estimate a moderate degree of risk aversion with $\alpha = 0.393$, which is lower than many other developing countries.

Individual $j$ in ethnic group $i$ gets the following expected utility if he chooses high education:

$$EU_{ij}^h = u(y_0 - S) + \beta(p_h u(r_h\omega_{ij}) + (1 - p_h)u(I)) = \frac{(y_0 - S)^{1-\alpha}}{1-\alpha} + \beta(p_h\frac{(r_h\omega_{ij})^{1-\alpha}}{1-\alpha} + (1 - p_h)\frac{I^{1-\alpha}}{1-\alpha})$$

The monetary cost to the family to get low education is standardized to 0. The net return to low

[26] The result also holds with a normal distribution of endowed abilities. We use the uniform distribution for simplicity in calculation.

education, conditional on finding a job, is assumed to be independent of own ability as ability is not well-observed in low-skilled jobs (Munshi and Rosenzweig, 2006). This net return is denoted as $r_l$ and is fixed in all districts (which can be interpreted as minimum wage). There are different types of network effects in the labour market for low educated people. According to Munshi and Rosenzweig (2006); Munshi (2011), networks can either provide low-skilled workers with more information regarding job opportunities and as a consequence lower their cost in job searching, or increase their job opportunities directly by providing referrals to job candidates. In both cases, ultimately, individuals benefiting from network effects will have higher chances of finding a job, compared with their low-skilled counterparts. Assume the effect of network increases with the network size and network effects only exist among low-skilled individuals (which follows (Munshi and Rosenzweig, 2006)), the chance of finding a job for low-skilled individuals $p_{li}$ in group $i$ is a function of both the network size of own group $s_i$ and the proportion of own group who receive low education. It is written down as $p_{li}(\cdot)$ for the moment.

Individual $j$ in ethnic group $i$ gets the following expected utility if he chooses low education:

$$EU_{ij}^{l} = u(y_0) + \beta(p_{li}(\cdot)u(r_l) + (1 - p_{li}(\cdot))u(I)) = \frac{y_0^{1-\alpha}}{1-\alpha} + \beta[p_{li}(\cdot)\frac{r_l^{1-\alpha}}{1-\alpha} + (1 - p_{li}(\cdot))\frac{I^{1-\alpha}}{1-\alpha}]$$

Mincer equation predicts that returns to high education are larger than those to low education. It is also reasonable to assume that unemployment insurance is less than labour market income from any job. Therefore, $0 < I < r_l < r_h$. Human capital theory indicates that individuals with high education are more likely to be employed than those with low education (i.e. $p_h > p_{li}(\cdot)$ for all $i$) (Cairo and Cajner, 2018). Without loss of generality, we assume that $p_h = 1$ and $0 < p_{li}(\cdot) < 1$. This makes sense as there is an excess demand of high-skilled jobs in all districts. As the highest wage employment rate in our data is 0.457 (see table 1, the Venda group), we further assume that $0 < p_{li}(\cdot) \leq 1/2$.

It is obvious that individual $j$ will choose low education when $\omega_{ij} = 0$, as $EU_{ij}^{h} < EU_{ij}^{l}$ in this case. We assume that individuals with the highest level of endowed ability will always choose high education (i.e. $EU_{ij}^{h} > EU_{ij}^{l}$ when $\omega_{ij} = 1$):

$$\frac{(y_0 - S)^{1-\alpha}}{1-\alpha} + \beta\frac{r_h^{1-\alpha}}{1-\alpha} > \frac{y_0^{1-\alpha}}{1-\alpha} + \beta(\frac{1}{2}\frac{r_l^{1-\alpha}}{1-\alpha} + \frac{1}{2}\frac{I^{1-\alpha}}{1-\alpha})$$

Let $A = \frac{y_0^{1-\alpha} - (y_0 - S)^{1-\alpha}}{\beta}$, we have:

$$A < r_h^{1-\alpha} - \frac{1}{2}r_l^{1-\alpha} - \frac{1}{2}I^{1-\alpha} \tag{7}$$

This condition is easily satisfied when the return to high education $r_h$ is much higher than the return to low education $r_l$, which is reasonable in our institutional setting where highly-educated workers are very limited.

For $0 < \omega_{ij} < 1$, individual $j$ in ethnic group $i$ chooses high education if $EU_{ij}^h > EU_{ij}^l$.[27] That is:

$$\frac{(y_0 - S)^{1-\alpha}}{1-\alpha} + \beta\frac{(r_h\omega_{ij})^{1-\alpha}}{1-\alpha} > \frac{y_0^{1-\alpha}}{1-\alpha} + \beta[p_{li}(\cdot)\frac{r_l^{1-\alpha}}{1-\alpha} + (1 - p_{li}(\cdot))\frac{I^{1-\alpha}}{1-\alpha}]$$

We have the following lemma:

**Lemma 1.** *For each ethnic group $i$, there is a threshold of ability $\omega_i$. Everyone whose ability is above (below) this level will go for high (low) education.*

*Proof.* For each individual $j$ in ethnic group $i$, the expected utility of having low education $EU_{ij}^l$ is fixed. The expected utility of receiving high education $EU_{ij}^h = \frac{(y_0-S)^{1-\alpha}}{1-\alpha} + \beta\frac{(r_h\omega_{ij})^{1-\alpha}}{1-\alpha}$ is an increasing function of $\omega_{ij}$. Since $EU_{ij}^h = 0$ if $\omega_{ij} = 0$ and $EU_{ij}^h > EU_{ij}^l$ if $\omega_{ij} = 1$, there exists a threshold of ability above (below) which everyone chooses high (low) education. □

This threshold is denoted as $\overline{\omega_i}$ so that $EU_{ij}^h = EU_{ij}^l$. Everyone in group $i$ whose ability is at the interval $[0, \overline{\omega_i}]$ chooses to have low education while those whose ability at the interval $(\overline{\omega_i}, 1]$ will have high education.

As indicated above, the network effect is determined by the number of own groups working in low-skilled jobs (assume information on high-skilled job openings is not useful to low-skilled workers), which is a function of both the size of own group and the proportion of people from own group who work in low-skilled jobs. The network effect $p_{li}(\cdot)$ can be further expressed as $p_{li}(\cdot) = p_l(s_i\overline{\omega_i})$. As the population size is standardised to 1, $s_i$ indicates the size of own group and $\overline{\omega_i}$ represents the proportion of own group working in low-skilled jobs. The threshold of ability above which everyone goes for high education is therefore determined by the following equation:

$$\frac{(y_0 - S)^{1-\alpha}}{1-\alpha} + \beta\frac{(r_h\overline{\omega_i})^{1-\alpha}}{1-\alpha} = \frac{y_0^{1-\alpha}}{1-\alpha} + \beta[p_l(s_i\overline{\omega_i})\frac{r_l^{1-\alpha}}{1-\alpha} + (1 - p_l(s_i\overline{\omega_i}))\frac{I^{1-\alpha}}{1-\alpha}] \tag{8}$$

As the probability of finding a job increases with network effect, we have $\frac{dp_l(s_i\overline{\omega_i})}{d(s_i\overline{\omega_i})} > 0$.

### 6.2 Ethnic diversity and human capital investment

Suppose $f(\overline{\omega_i}) = \frac{(y_0-S)^{1-\alpha}}{1-\alpha} + \beta\frac{(r_h\overline{\omega_i})^{1-\alpha}}{1-\alpha}$, $g(\overline{\omega_i}) = \frac{y_0^{1-\alpha}}{1-\alpha} + \beta[p_l(s_i\overline{\omega_i})\frac{r_l^{1-\alpha}}{1-\alpha} + (1 - p_l(s_i\overline{\omega_i}))\frac{I^{1-\alpha}}{1-\alpha}]$. we have the following lemma:

**Lemma 2.** *Given $s_i$, $g(\overline{\omega_i})$ is an increasing function of $\overline{\omega_i}$. Given $\overline{\omega_i}$, $g(\overline{\omega_i})$ is an increasing function of $s_i$.*

[27] Without loss of generality, this also incorporates people with credit constraints (i.e. $y_0 < S$). They will spontaneously choose low education as $EU_{ij}^h < EU_{ij}^l$ is easier for them to satisfy.

*Proof.* As $\frac{dp_l(s_i\overline{\omega_i})}{d(s_i\overline{\omega_i})} > 0$ and $s_i$ is given, we have $\frac{dp_l(s_i\overline{\omega_i})}{d\overline{\omega_i}} > 0$. Therefore:

$$\frac{dg(\overline{\omega_i})}{d\overline{\omega_i}} = \frac{dp_l(s_i\overline{\omega_i})}{d\overline{\omega_i}} \cdot \frac{r_l^{1-\alpha} - I^{1-\alpha}}{1-\alpha} \cdot \beta$$

Since $r^l > I$, we have $\frac{dg(\overline{\omega_i})}{d\overline{\omega_i}} > 0$.

Similarly, $\frac{dg(\overline{\omega_i})}{ds_i} > 0$. □

Note that in our institutional setting, the unemployment rate among the low-skilled individuals is still high even in the most diverse districts, which means network effect cannot increase the chance of being employed drastically. We therefore impose the following condition:

*Condition 1.* $\frac{dp_l(s_i\overline{\omega_i})}{d(\overline{\omega_i})} \cdot \frac{r_l^{1-\alpha} - I^{1-\alpha}}{1-\alpha} < r_h^{1-\alpha}$.

Intuitively, $\frac{dp_l(s_i\overline{\omega_i})}{d(\overline{\omega_i})}$ is the increase in the probability of finding a low-skilled job with respect to the increase in the proportion of own group with low education. $\frac{r_l^{1-\alpha} - I^{1-\alpha}}{1-\alpha}$ is the corresponding increase in utility from getting a job, as opposed to being unemployed. $\frac{dp_l(s_i\overline{\omega_i})}{d(\overline{\omega_i})} \cdot \frac{r_l^{1-\alpha} - I^{1-\alpha}}{1-\alpha} < r_h^{1-\alpha}$ is therefore associated with an increase in the expected utility from a bigger network size, which is bounded due to the overall high unemployment rate in the labour market.

This leads to the following proposition:

**Proposition 1.** *For ethnic group $i$, the threshold of ability above which individuals go for high education $\overline{\omega_i}$ increases with the group's population share $s_i$.*

*Proof.* We focus on group $i$ and omit the notation $i$ for simplicity. Suppose $h(\overline{\omega}) = f(\overline{\omega}) - g(\overline{\omega}) = \beta[\frac{(r_h\overline{\omega})^{1-\alpha}}{1-\alpha} - p_l(s\overline{\omega})\frac{r_l^{1-\alpha}}{1-\alpha} - (1 - p_l(s\overline{\omega}))\frac{I^{1-\alpha}}{1-\alpha}] - \frac{\beta A}{1-\alpha}$, we have:

$$\frac{dh(\overline{\omega})}{d\overline{\omega}} = \beta(r_h^{1-\alpha}\overline{\omega}^{-\alpha} - \frac{dp_l(s\overline{\omega})}{d\overline{\omega}}(\frac{r_l^{1-\alpha}}{1-\alpha} - \frac{I^{1-\alpha}}{1-\alpha}))$$

Since $\overline{\omega}^{-\alpha} \geq 1$, given condition 1, $r_h^{1-\alpha}\overline{\omega}^{-\alpha} - \frac{dp_l(s\overline{\omega})}{d\overline{\omega}}(\frac{r_l^{1-\alpha}}{1-\alpha} - \frac{I^{1-\alpha}}{1-\alpha}) \geq r_h^{1-\alpha} - \frac{dp_l(s\overline{\omega})}{d\overline{\omega}}(\frac{r_l^{1-\alpha}}{1-\alpha} - \frac{I^{1-\alpha}}{1-\alpha}) > 0$. That is, $h(\overline{\omega})$ is an increasing function of $\overline{\omega}$.

Starting from the population share $s_0$, the threshold $\omega_0$ is determined by:

$$h_0(\overline{\omega_0}) = \beta[\frac{(r_h\overline{\omega_0})^{1-\alpha}}{1-\alpha} - (p_l(s_0\overline{\omega_0})\frac{r_l^{1-\alpha}}{1-\alpha} + (1 - p_l(s_0\overline{\omega_0}))\frac{I^{1-\alpha}}{1-\alpha})] - \frac{\beta A}{1-\alpha} = 0$$

Suppose $s_0$ now increases to $s_1$. As $g(\overline{\omega})$ increases with $s$, at the level of $\overline{\omega_0}$, $h_1(\overline{\omega_0}) = \beta[\frac{(r_h\overline{\omega_0})^{1-\alpha}}{1-\alpha} - (p_l(s_1\overline{\omega_0})\frac{r_l^{1-\alpha}}{1-\alpha} + (1 - p_l(s_1\overline{\omega_0}))\frac{I^{1-\alpha}}{1-\alpha})] - \frac{\beta A}{1-\alpha} < \beta[\frac{(r_h\overline{\omega_0})^{1-\alpha}}{1-\alpha} - (p_l(s_0\overline{\omega_0})\frac{r_l^{1-\alpha}}{1-\alpha} + (1 - p_l(s_0\overline{\omega_0}))\frac{I^{1-\alpha}}{1-\alpha})] - \frac{\beta A}{1-\alpha} <$

0.

As $h(\overline{\omega})$ is an increasing function of $\overline{\omega}$, we need to have $s_1 > s_0$ so that $h_1(\overline{\omega_1}) = 0$, which means $\overline{\omega}$ should increase with $s$ to satisfy equation 8.

□

Equation 8 determines the threshold $\overline{\omega}$ above which everyone has high education. Re-arranging the terms in this equation, we get:

$$\overline{\omega} = \frac{[A + p_l(s\overline{\omega})r_l^{1-\alpha} + (1 - p_l(s\overline{\omega}))I^{1-\alpha}]^{\frac{1}{1-\alpha}}}{r_h}$$

Since $0 < p_l(s\overline{\omega}) \leq 1/2$, we can get the range of $\overline{\omega}$:

$$\frac{[A + I^{1-\alpha}]^{\frac{1}{1-\alpha}}}{r_h} < \overline{\omega} \leq \frac{[A + 1/2 \cdot r_l^{1-\alpha} + 1/2 \cdot I^{1-\alpha}]^{\frac{1}{1-\alpha}}}{r_h} \tag{9}$$

### 6.3 Ethnic diversity and employment rates

The previous section proves that ethnic groups with larger population share will have less incentive to invest in high education. This section further shows that consequently, they may be trapped in low-skilled occupations due to the prevalence of ethnicity-specific networks.

The overall employment rate in group $i$ (both high-skilled and low-skilled jobs) is written as:

$$E = \underbrace{(1 - \overline{\omega})}_{\text{high-skilled job}} + \underbrace{\overline{\omega} \cdot p_l(s\overline{\omega})}_{\text{low-skilled job}} = 1 - \overline{\omega}[1 - p_l(s\overline{\omega})] \tag{10}$$

When the group size $s$ increases, we have:

$$\frac{dE}{ds} = \underbrace{-\frac{d\overline{\omega}}{ds}[1 - p_l(s\overline{\omega})]}_{\text{human capital effect}} + \overline{\omega} \cdot \underbrace{\frac{dp_l(s\overline{\omega})}{d(s\overline{\omega})} \cdot \frac{d(s\overline{\omega})}{ds}}_{\text{network effect}}$$

$\frac{d\overline{\omega}}{ds} > 0$ indicates that as group size increases, human capital effect decreases the number of people having high education, therefore leading to more unemployment in the low-skilled sector. $\frac{dp_l(s\overline{\omega})}{d(s\overline{\omega})} \cdot \frac{d(s\overline{\omega})}{ds} > 0$ means that the network effect increases the employment opportunities of low educated people. Therefore, intuitively, if the employment rate decreases with the population share (i.e. $\frac{dE}{ds} < 0$), the human capital effect should be more important than the network effect.

To further investigate when the human capital effect dominates network effect, from equation 8,

we have:

$$p_l(s\overline{\omega}) = \frac{r_h^{1-\alpha}\overline{\omega}^{1-\alpha} - I^{1-\alpha} - A}{r_l^{1-\alpha} - I^{1-\alpha}}$$

Replacing $p_l$ in equation 10, we can re-write the employment rate as:

$$E = 1 - \overline{\omega} + \overline{\omega} \cdot \frac{r_h^{1-\alpha}\overline{\omega}^{1-\alpha} - I^{1-\alpha} - A}{r_l^{1-\alpha} - I^{1-\alpha}}$$

$$\frac{dE}{ds} = \frac{d\overline{\omega}}{ds} \cdot (-1 + \frac{r_h^{1-\alpha}(2-\alpha)\overline{\omega}^{1-\alpha} - I^{1-\alpha} - A}{r_l^{1-\alpha} - I^{1-\alpha}})$$

Based on this, we get the following proposition:

**Proposition 2.** *When* $\overline{\omega} < (\frac{2}{3-\alpha})^{\frac{1}{1-\alpha}}$, *we have* $\frac{dE}{ds} < 0$. *This means human capital effect is more important than network effect and therefore the employment rate among an ethnic group decreases with its population share.*

*Proof.* $\frac{d\overline{\omega}}{ds} > 0$ always holds. Therefore, when employment rate decreases with own group size, we have $\frac{dE}{ds} < 0$, equivalently:

$$\overline{\omega} < \frac{(r_l^{1-\alpha} + A)^{\frac{1}{1-\alpha}}}{r_h(2-\alpha)^{\frac{1}{1-\alpha}}} \tag{11}$$

From equation 9, we get $\overline{\omega} > \frac{[A+I^{1-\alpha}]^{\frac{1}{1-\alpha}}}{r_h}$ for any $\omega$. As there should exist at least one $\overline{\omega}$ which satisfies equation 11, we have:

$$\frac{[A + I^{1-\alpha}]^{\frac{1}{1-\alpha}}}{r_h} < \frac{(r_l^{1-\alpha} + A)^{\frac{1}{1-\alpha}}}{r_h(2-\alpha)^{\frac{1}{1-\alpha}}}$$

Therefore:

$$A < \frac{r_l^{1-\alpha} - (2-\alpha)I^{1-\alpha}}{1-\alpha} \tag{12}$$

Both inequalities 7 and 12 should be satisfied. We consider the following two cases:

*Case 1:* $r_h^{1-\alpha} > \frac{(3-\alpha)(r_l^{1-\alpha} - I^{1-\alpha})}{2(1-\alpha)}$.

Equivalently,

$$r_h > [\frac{(3-\alpha)(r_l^{1-\alpha} - I^{1-\alpha})}{2(1-\alpha)}]^{\frac{1}{1-\alpha}} \tag{13}$$

This implies $r_h^{1-\alpha} - \frac{1}{2}r_l^{1-\alpha} - \frac{1}{2}I^{1-\alpha} > \frac{r_l^{1-\alpha}-(2-\alpha)I^{1-\alpha}}{1-\alpha}$. Plugging inequalities 12 and 13 into 11, we can get $\overline{\omega} < (\frac{2}{3-\alpha})^{\frac{1}{1-\alpha}}$.

The upper limit of $\overline{\omega}$ from inequality 9 is $\frac{[A+1/2\cdot r_l^{1-\alpha}+1/2\cdot I^{1-\alpha}]^{\frac{1}{1-\alpha}}}{r_h}$, which is smaller than 1 after plugging in inequalities 12 and 13.

*Case 2:* $r_h^{1-\alpha} \leq \frac{(3-\alpha)(r_l^{1-\alpha}-I^{1-\alpha})}{2(1-\alpha)}$.

Equivalently,

$$r_h \leq [\frac{(3-\alpha)(r_l^{1-\alpha} - I^{1-\alpha})}{2(1-\alpha)}]^{\frac{1}{1-\alpha}} \tag{14}$$

This implies $r_h^{1-\alpha} - \frac{1}{2}r_l^{1-\alpha} - \frac{1}{2}I^{1-\alpha} \leq \frac{r_l^{1-\alpha}-(2-\alpha)I^{1-\alpha}}{1-\alpha}$. Plugging inequalities 7 and 14 into 11, we can verify that $\overline{\omega} < (\frac{2}{3-\alpha})^{\frac{1}{1-\alpha}}$ makes sure $\frac{dE}{ds} < 0$.

The upper limit of $\overline{\omega}$, $\frac{[A+1/2\cdot r_l^{1-\alpha}+1/2\cdot I^{1-\alpha}]^{\frac{1}{1-\alpha}}}{r_h}$, is also smaller than 1 after plugging in inequalities 7 and 14.

□

It is easy to verify that the required $\overline{\omega}$ is larger when the group is less risk averse (i.e. $\alpha$ is small).[28] Higher $\alpha$ implies lower $\overline{\omega}$ because more risk averse people have a higher tendency to invest in high education to avoid being unemployed. As $\alpha$ is relatively low in South Africa, a higher $\overline{\omega}$ (which means a lower proportion of people with high education) can lead to a decrease in employment rate when the group size becomes larger.

So far we have discussed how the employment rate of an ethnic group is affected by its group size. This result links to ethnic diversity in the following way. A homogenous place should have a dominant ethnic group with a large group size. This group will be trapped in low-skilled jobs, and therefore will have a low employment rate. And the average employment rate among all the groups in the whole district will also be lower if the majority (i.e. the dominant group) have low employment rates.

A more diverse district has a more even distribution of ethnic groups and each group has relatively smaller group sizes. In this district people can rely less on ethnicity-based networks and will be motivated to invest in more education and look for high-skilled jobs which provide more employment opportunities.

### 6.4 Empirical evidence of the model

The theoretical model links human capital, networks and ethnic diversity. In addition to the main result that wage employment rate increases with the level of ethnic diversity, the model gives two key predictions. First, the probability of having a college degree should increase with the level of ethnic

[28] This is because $[\frac{2}{3-\alpha}]^{\frac{1}{1-\alpha}}$ is a decreasing function of $\alpha$.

diversity. Second, both educational attainment and wage employment rate should decrease with the relative group size. We find two series of empirical evidence to further support the model and its predictions.

First of all, we use the same instrument as the main analysis and conduct an instrumental variable regression of the probability of obtaining a college degree on the level of ethnic diversity. Consistent with what the model indicates, panel A in Table 7 reports that coming from a more ethnically diverse district increases the chance of having a college degree. Panel B shows that obtaining a college degree is indeed positively correlated with the chance of being employed. Panel C restricts the sample to wage employed people and further confirms that the increase in employment rate in more ethnically diverse districts mainly occurs in high-skilled occupations such as managerial jobs, consistent with the human capital effect in the theoretical model.

Secondly, we prove the key channel of the model that both educational attainment and wage employment rate should decrease with the relative size of own groups in the district. The results are in Table 8. Following the construction process of the instrumental variable in our main analysis, we use the predicted population share of own ethnic group in each district based on the equidistance of the corresponding district to all ethnic homelands as an instrument variable for the real population share for each ethnic group in each district. The results for IV regressions indicate that both the chance of achieving a college degree and the wage employment rate decrease with the population share of own groups.

### 6.5 Summary of the theoretical model and mechanism

In summary, diversity along ethnic lines motivates people to invest more in human capital. And this improvement in human capital increases an individual's chances of finding a job. Our model does not impose any intrinsic difference in taste, skills or attitudes between different ethnic groups. The tradeoff between the cost of and benefit from developing human capital mainly relates to the group size which determine the size of networks individuals can rely on. Higher level of human capital investments is the most likely to occur in a place where the distribution of group size is relatively even, which implies a larger ethnic diversity. It is because in ethnically homogeneous districts, the dominant group with a large group size benefits from large network effects. Network easily channels individuals into low-skilled jobs and disincentivises them to invest more in human capital.

Our theoretical model complements classical network theory by emphasising the downside of networks in the labour market. Classical network theory focuses more on the positive effect of networks in job searching, such as reducing information asymmetry and search frictions. In this paper, however, reliance on networks can trap individuals in low-level of human capital and low-skilled jobs, because the network effects are the most useful for low-skilled jobs.

# 7 Conclusion and Discussion

This paper provides empirical support for the positive effect of within-black ethnic diversity and Blacks' labour market outcomes in post-apartheid South Africa based on an instrumental variable approach.To provide explanations regarding the possible mechanisms at play, we propose a theoretical model we propose a model based on human capital and network theories. In more homogenous districts, where individuals can rely on larger networks to get a low skill job, they tend to invest less in education. The opposite occurs in more diverse districts where the absence of large networks incentivise individuals to get higher education in order to find an occupation.

Our findings could inform the South African policy makers on the importance of the ethnic heterogeneity within the black population for education and employment decisions. Unemployment rapidly increased in South Africa, especially among blacks, during the 1990s and 2000s, when the economy was unable to absorb the growing supply of semi!skilled labor (Gradin, 2019). According to Haver Analytics, in 2019 South Africa had the highest unemployment rate amongst the BRICS nations, more than double that of Brazil (11.6%) and considerably higher than that of India (7.1%), Russia (4.8%) and China (3.8%). In our setting, unemployment would be the consequence of heavy relying on one's own ethnic network to find a job instead of getting higher education. Policies that aim at encouraging ethnic interactions and at making higher education more accessible could have substantial effects on reducing the unemployment rate. Anand et al. (2016) estimate that a ten percentage point reduction in unemployment would lower the Gini coefficient for South Africa by 3 percent.

We believe that our results could be relevant also to other countries, especially the ones characterised by strong ethnic identities and high barriers to accessing education. Our results could partly be explained by the peculiarity of the South African case, which is characterised by labor market inflexibility and a small informal sector compared with other developing countries Kingdon and Knight (2007). In contexts where the informality sector acts as a refuge, taking in those workers who do not find opportunities in the formal sector, we might find that ethnic diversity doesn't increase employment per se, but rather its quality.

## References


Acemoglu, D. and R. Shimer (1999). Efficient Unemployment Insurance. *The Journal of Political Economy 107*, 893–928.

Alesina, A., R. Baqir, and W. Easterly (1997). Public goods and ethnic divisions. *The Quarterly Journal of Economics 114*(4), 1243–1284.

Alesina, A., A. Devleeschauwer, S. Kurlat, and R. Wacziarg (2003). Fractionalization. *Journal of Economic Growth 8*, 155–194.

Alesina, A., J. Harnoss, and H. Rapoport (2016). Birthplace diversity and economic prosperity.

Alesina, A. and E. La Ferrara (2000). Participation in heterogeneous communities. *The Quarterly Journal of Economics 115*(3), 847–904.

Alesina, A. and E. La Ferrara (2002). Who trusts others? *Journal of Public Economics 85*(2), 207–234.

Alesina, A. and E. La Ferrara (2005). Ethnic diversity and economic performance. *Journal of Economic Literature XLIII*(September), 762–800.

Alesina, A., S. Michalopoulos, and E. Papaioannou (2016). Ethnic inequality. *Journal of Political Economy 124*(2), 428–488.

Allanson, P., J. Atkins, and T. Hinks (2000). A multilateral decomposition of racial wage differentials in the 1994 south african labour market. *The Journal of Development Studies 37*(1), 93–120.

Allanson, P., J. Atkins, and T. Hinks (2002). No end to the racial wage hierarchy in south africa? *Review of Development Economics 6*(3), 442–459.

Amodio, F. and G. Chiovelli (2018). Ethnicity and violence during democratic transitions: Evidence from South Africa. *Journal of the European Economic Association 16*(4), 1234–1280.

Anand, R., S. Kothari, and N. Kumar (2016). South Africa: labor market dynamics and inequality.

Bakens, J., P. Mulder, and P. Nijkamp (2013). Economic impacts of cultural diversity in the netherlands: Productivity, utility, and sorting. *Journal of Regional Science 53*(1), 8–36.

Banerjee, A., S. Galiani, J. Levinsohn, Z. McLaren, and I. Woolard (2008). Why has unemployment risen in the New South Africa? *Economics of Transition 16*(4), 715–740.

Beine, M., F. Docquier, and M. Schiff (2013). International migration, transfer of norms and home country fertility. *Canadian Journal of Economics/Revue canadienne d'économique 46*(4), 1406–1430.

Bhorat, H. and M. Oosthuizen (2005). The post-Apartheid South African labour market.

Brick, K., M. Visser, and J. Burns (2012). Risk Aversion: Experimental Evidence from South African Fishing Communities. *American Journal of Agricultural Economics 94*(1), 133–152.

Cairo, I. and T. Cajner (2018). Human Capital and Unemployment Dynamics: Why More Educated Workers Enjoy Greater Employment Stability. *The Economic Journal 128*(609), 652–682.

Christopher, A. J. (2001). *The atlas of changing South Africa*. London and New York: Routledge.

Desmet, K., I. Ortuño-Ortín, and R. Wacziarg (2012). The political economy of linguistic cleavages. *Journal of Development Economics 97*(2), 322–338.

Desmet, K., I. Ortuno-Ortin, and R. Wacziarg (2017). Culture, ethnicity, and diversity. *American Economic Review 107*(9), 2479–2513.

Dumont, J.-C., G. Spielvogel, and S. Widmaier (2010). International migrants in developed, emerging and developing countries. *OECD Social, Employment and Migration Working Paper No.114*.

Dustmann, C., U. Schoenberg, and J. Stuhler (2017). Labor Supply Shocks, Native Wages, and the Adjustment of Local Employment. *The Quarterly Journal of Economics 132*(1), 435–483.

Easterly, W. and R. Levine (1997). Africa's growth tragedy: policies and ethnic divisions. *The Quarterly Journal of Economics 112*(4), 1203–1250.

Gomes, J., K. Desmet, and O.-O. Ignacio (2016). The Geography of Linguistic Diversity and the Provision of Public Goods.

Gradin, C. (2014). Poverty and ethnicity among black South Africans.

Gradin, C. (2019). Occupational segregation by race in South Africa after apartheid. *Review of Development Economics 23*, 553–576.

Heintz, J. and D. Posel (2007). Revisiting informal employment and segmentation in the South African labor market.

Kingdon, G. G. and J. Knight (2004a). Race and the incidence of unemployment in south africa. *Review of Development Economics 8*(2), 198–222.

Kingdon, G. G. and J. Knight (2004b). Unemployment in South Africa: the nature of the beast. *World Development 32*(3), 391–408.

Kingdon, G. G. and J. Knight (2007). Unemployment in south africa, 1995?2003: Causes, problems and policies. *Journal of African Economies 16*(5), 813–884.

Leibbrandt, M., I. Woolard, H. McEwen, and C. Koep (2009). Employment and inequality outcomes in South Africa.

Lewis, M. P., G. F. Simons, and C. D. Fennig (2009). *Ethnologue: languages of the world*, Volume 16. SIL international Dallas, TX.

Mariotti, M. (2012). Labour markets during apartheid in south africa. *The Economic History Review 65*(3), 1100–1122.

Mayda, A. M. (2010). International migration: a panel data analysis of the determinants of bilateral flows. *Journal of Population Economics 23*(4), 1249–1274.

Michalopoulos, S. and E. Papaioannou (2013). Pre-colonial ethnic institutions and contemporary african development. *Econometrica 81*(1), 113–152.

Montalvo, J. G. and M. Reynal-Querol (2020). Ethnic diversity and growth: Revisiting the evidence. *The Review of Economics and Statistics*, 1–43.

Munshi, K. (2011). Strength in numbers: networks as a solution to occupational traps. *Review of Economic Studies 78*(3), 1069–1101.

Munshi, K. and M. Rosenzweig (2006). Traditional institutions meet the modern world: caste, gender and schooling choices in a globalizing economy. *American Economic Review 96*(4), 1225–1252.

Mwakikagile, G. (2010). *South Africa as a multi-ethnic society*. Dar es Salaam, Tanzania: Continental Press.

Nathan, M. (2011). The long term impact of migration in british cities: diversity, wages, employment and prices. *SERC Discussion Paper 67*.

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

Ortega, F. and G. Peri (2014). Openness and income: The roles of trade and migration. *Journal of International Economics 92*(2), 231–251.

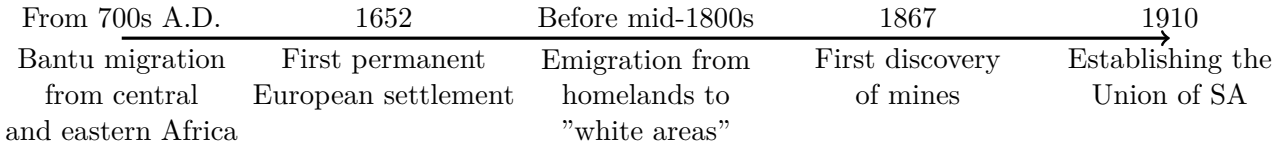
Ottaviano, G. I. and G. Peri (2006). The economic value of cultural diversity: evidence from us cities. *Journal of Economic geography 6*(1), 9–44.

Ottaviano, G. I. P. and G. Peri (2005). Cities and cultures. *Journal of Urban Economics 58*(2), 304 – 337.

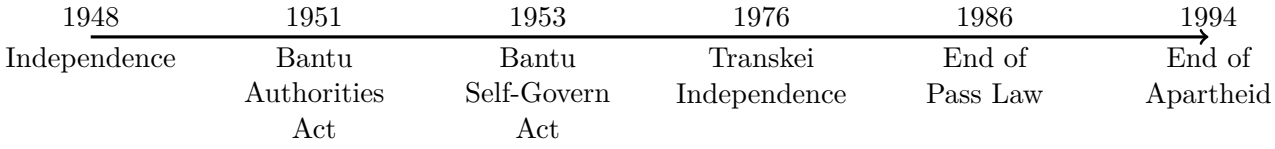
Posel, D. (2001). What's in a name? racial categorisations under apartheid and their afterlife. *Transformation 47*, 50–74.

Santos Silva, J. and S. Tenreyro (2006). The log of gravity. *The Review of Economics and statistics 88*(4), 641–658.

Suedekum, J., K. Wolf, and U. Blien (2014). Cultural diversity and local labour markets. *Regional Studies 48*(1), 173–191.



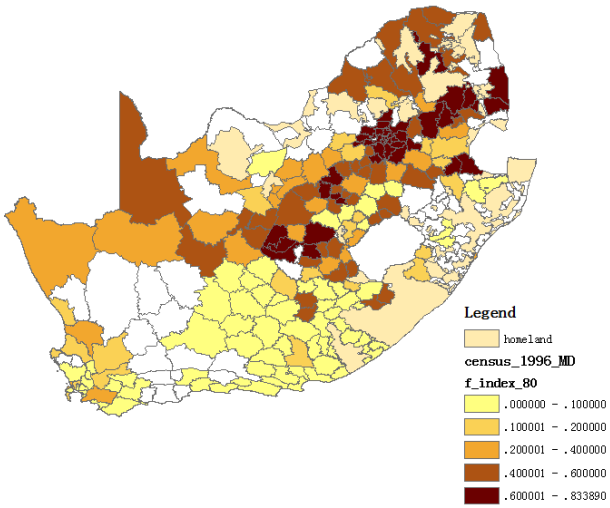

**(a)** Timeline of Bantu migration and early development in South Africa



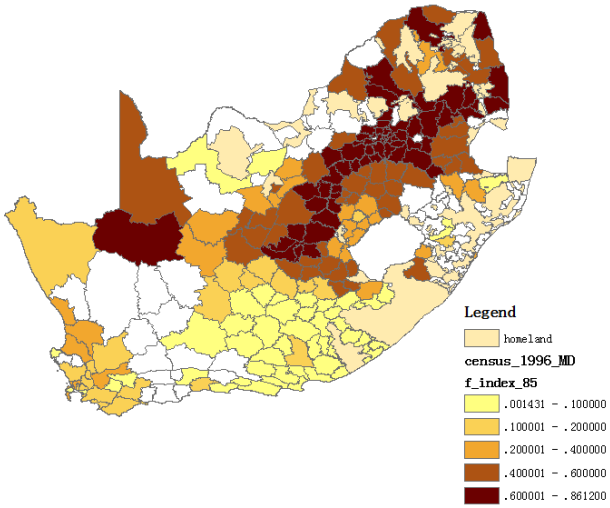

**(b)** Timeline of modern South Africa starting from Apartheid

Notes: The figures presents the timeline of important nodes in South African history: Bantu migration from central and eastern Africa, emigration of ethnic groups from original homelands, the White colonisation, the discovery of mines and Apartheid regime. Sources of narratives: Mwakikagile (2010) and Gradin (2014).

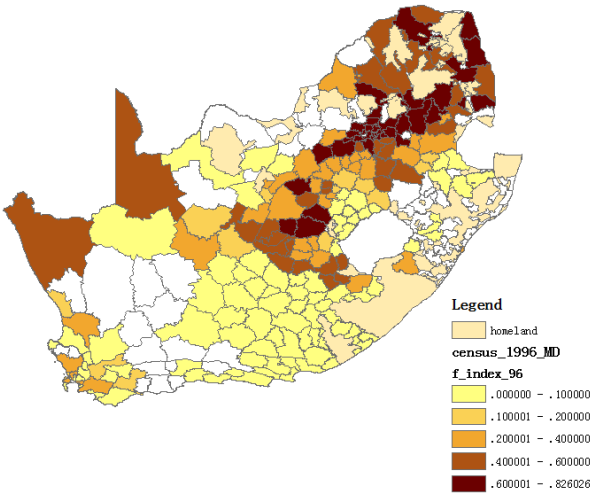
**Figure 1.** Timeline of Bantu migration, historical development and Apartheid regime in South Africa



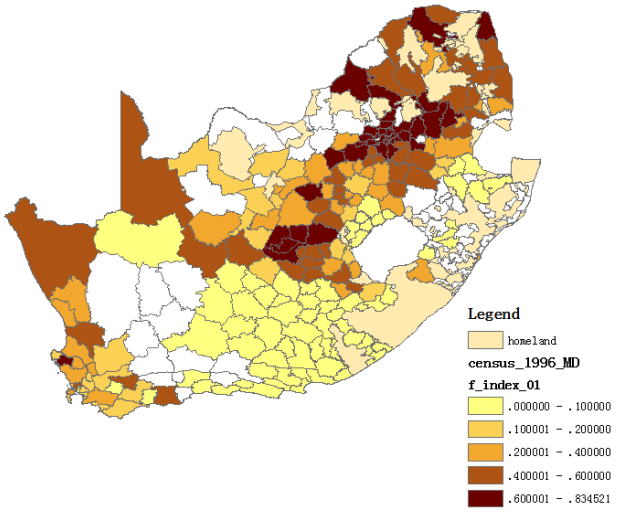

**(a)** *ELF* 1980




**(b)** *ELF* 1985



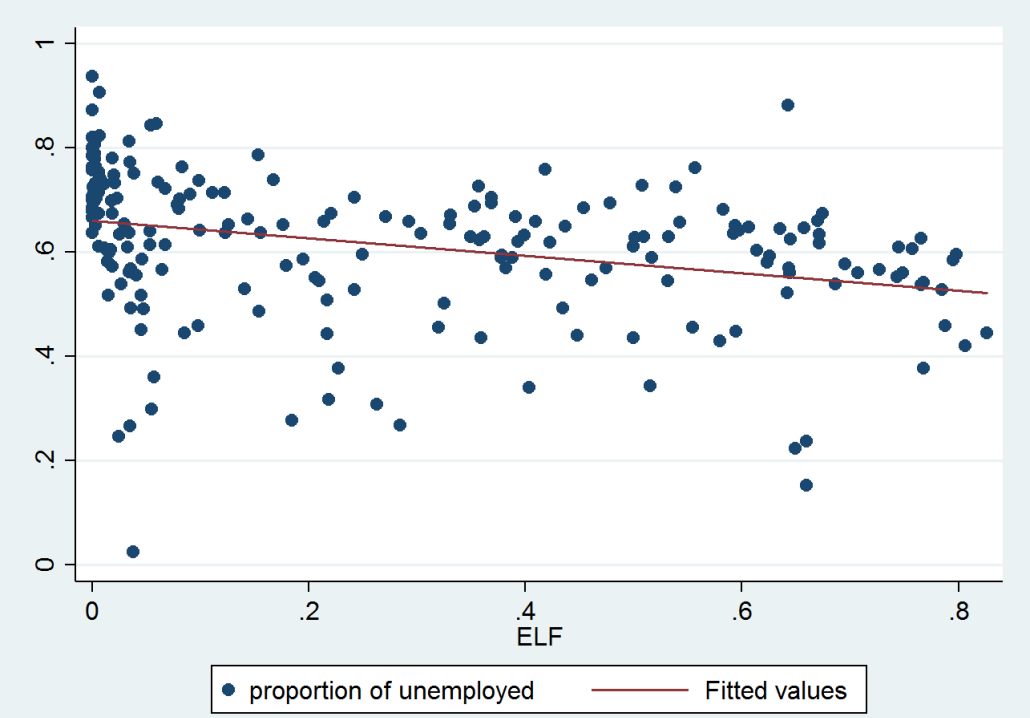

**(c)** *ELF* 1996



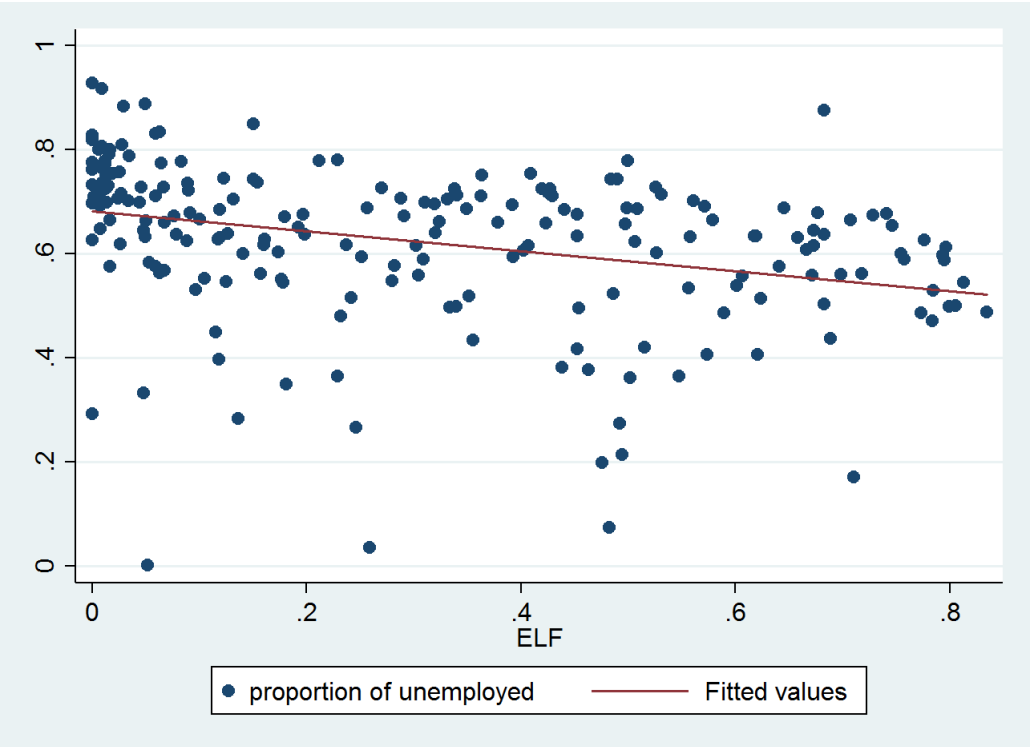

**(d)** *ELF* 2001

Notes: The figure presents the geographical pattern of ethnic diversity across South African districts in 1980, 1985, 1996 and 2001 for the "white areas". Within-black ethnic diversity is measured with Fractionalisation Index analogue to Herfindahl Index. The results are calculated by the authors based on the corresponding census data.

**Figure 2.** Distribution of ethnic fractionalisation index: 1980. 1985, 1996, 2001



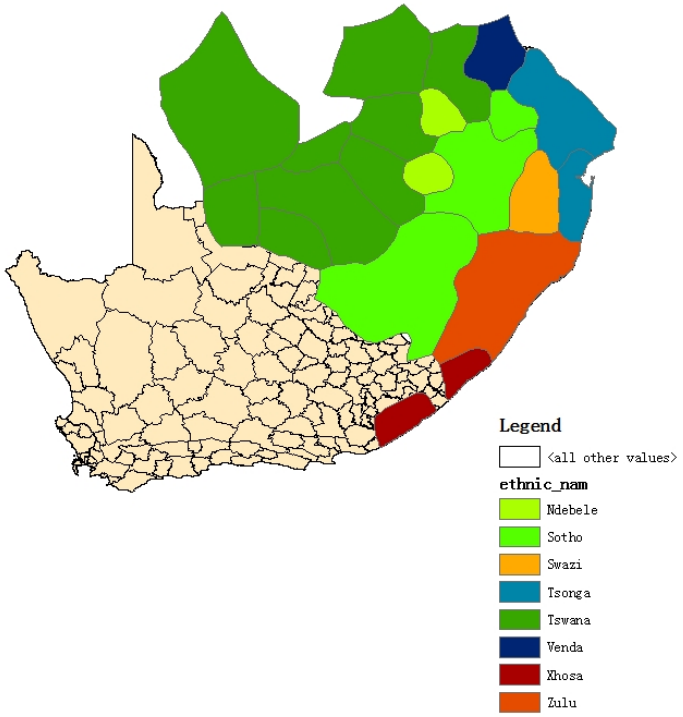

**(a)** Unemployment and ethnic diversity (ELF) 1996



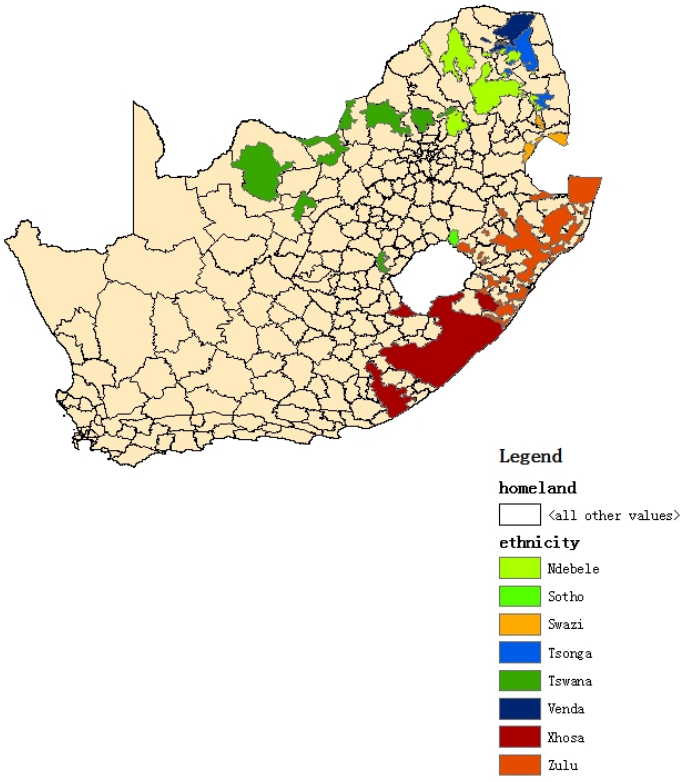

**(b)** Unemployment and ethnic diversity (ELF) 2001

Notes: The figures present the results on the correlation between ethnic diversity and unemployment rate. Both are measured at the magisterial district level (therefore unemployment rate is calculated as the proportion of unemployed people over the whole working-age black population in a district). The results are calculated by the authors based on 1996 and 2001 census data.

**Figure 3.** The relationship between ethnic diversity and unemployment in South Africa in 1996 and 2001



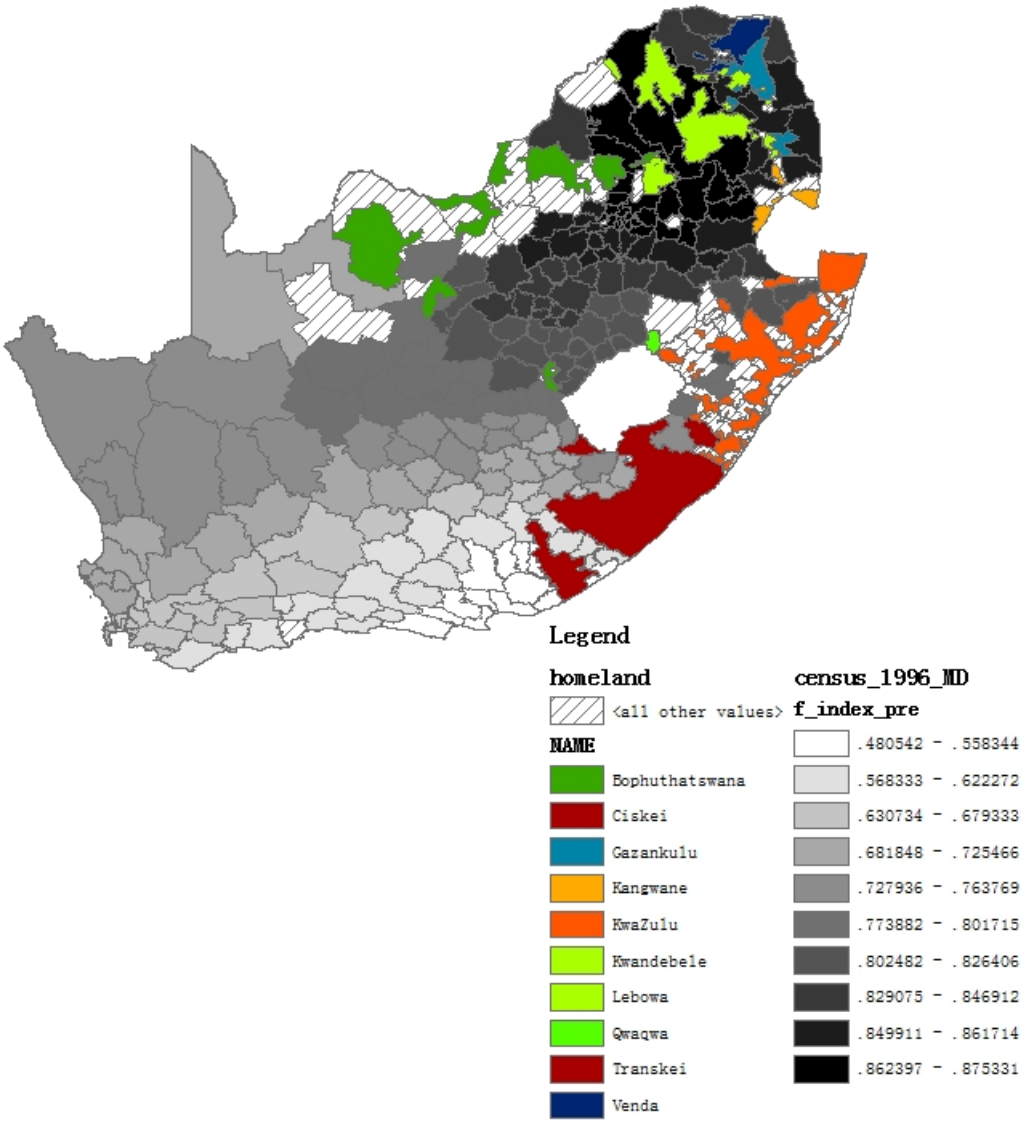

**(a)** Murdock's map




**(b)** Bantustan

Notes: The figures compare the distribution of ethnic groups in South Africa in Murdock map and the location of Bantustans as proxies for ethnic homelands. Murdock map comes from George Murdock's 1959 work which illustrates the dominant ethnic group in each geographical unit, which is highly consistent with the Bantustans for these ethnic groups assigned by the Apartheid government. This confirms that the location of these Bantustans can well reflect the spatial distribution of original homelands for those ethnic groups.

**Figure 4.** Comparison between the historical settlements of the black ethnic groups and Bantustans




Notes: The figure shows the spatial distribution of our instrumental variable for ethnic diversity - the predicted ethnic fractionalisation index. Following the idea that districts more (less) equidistant to multiple homelands are more (less) diverse, we first calculate the stock of each ethnic group in each district based on the distance between the district and the corresponding homeland with a gravity model. The instrumental variable is a predicted fractionalisation index calculated based on the predicted stock of black migrants.

**Figure 5.** Distribution of predicted ethnic fractionalisation index

**Table 1.** Summary statistics of demographics and employment among black ethnic groups in 2001

| | Population size | Share of the black population | Self employed | Wage employee | Unemployed+inactive | ELF |
|---|---|---|---|---|---|---|
| Xhosa | 3105625 | 0.249 | 0.017 | 0.299 | 0.684 | 0.251 |
| | | | [0.130] | [0.458] | [0.465] | [0.298] |
| Zulu | 2798132 | 0.224 | 0.025 | 0.331 | 0.643 | 0.558 |
| | | | [0.156] | [0.471] | [0.479] | [0.264] |
| South Sotho | 2531013 | 0.203 | 0.020 | 0.324 | 0.657 | 0.500 |
| | | | [0.139] | [0.468] | [0.475] | [0.221] |
| Tswana | 1373413 | 0.110 | 0.018 | 0.373 | 0.610 | 0.578 |
| | | | [0.132] | [0.484] | [0.488] | [0.225] |
| North Sotho | 1341608 | 0.107 | 0.027 | 0.396 | 0.577 | 0.689 |
| | | | [0.163] | [0.490] | [0.494] | [0.148] |
| Tsonga | 552403.3 | 0.044 | 0.048 | 0.421 | 0.531 | 0.714 |
| | | | [0.214] | [0.494] | [0.50] | [0.128] |
| Ndebele | 292188.3 | 0.023 | 0.029 | 0.370 | 0.601 | 0.708 |
| | | | [0.168] | [0.483] | [0.490] | [0.131] |
| Swazi | 324071.7 | 0.026 | 0.028 | 0.376 | 0.597 | 0.579 |
| | | | [0.164] | [0.484] | [0.491] | [0.189] |
| Venda | 172927.4 | 0.014 | 0.034 | 0.457 | 0.509 | 0.724 |
| | | | [0.183] | [0.498] | [0.500] | [0.119] |
| Overall | 12491382 | 1.000 | 0.023 | 0.341 | 0.636 | 0.302 |
| | | | [0.149] | [0.474] | [0.481] | [0.259] |

Note: The number and proportion of each ethnic group in the whole black population are calculated in the "white" magisterial districts which can be matched to 1985 census and whose black population accounts for more than 1% of the overall population. Population size is not always an integer because it is weighted by each person's weight in the census data. Employment outcomes are calculated from individual-level 2001 census data among the working-age blacks."Self-employed" refers to the proportion of self-employed people in each ethnic group over the whole working-age population of the corresponding ethnic group. Other labour market outcomes are calculated in similar ways. The 2001 census data does not distinguish unemployed and economically inactive people. The mean degree of ethnic diversity index is calculated at the district level. All other statistics are calculated at the individual level.

**Table 2.** Summary statistics of ethnic fragmentation and labour market outcomes in 2001

| | High ELF | | | Low ELF | | | |
|---|---|---|---|---|---|---|---|
| | Mean | S.d | Obs | Mean | S.d. | Obs | ttest |
| ELF | 0.527 | 0.126 | 105 | 0.077 | 0.084 | 105 | *** |
| Self employment | 0.022 | 0.055 | 105 | 0.019 | 0.045 | 105 | |
| Wage employee | 0.396 | 0.114 | 105 | 0.315 | 0.118 | 105 | *** |
| Unemployed | 0.582 | 0.118 | 105 | 0.667 | 0.118 | 105 | *** |
| Agriculture | 0.338 | 0.155 | 105 | 0.376 | 0.152 | 105 | |
| Manufacture | 0.183 | 0.130 | 105 | 0.096 | 0.089 | 105 | *** |
| Service | 0.478 | 0.138 | 105 | 0.527 | 0.152 | 105 | * |
| Manager | 0.017 | 0.051 | 105 | 0.017 | 0.068 | 105 | |
| Profession | 0.082 | 0.075 | 105 | 0.080 | 0.076 | 105 | |
| Clerk | 0.056 | 0.055 | 105 | 0.054 | 0.084 | 105 | |
| Serve | 0.081 | 0.071 | 105 | 0.076 | 0.069 | 105 | |
| Craft | 0.059 | 0.071 | 105 | 0.084 | 0.093 | 105 | *** |
| Skilld agriculture | 0.117 | 0.084 | 105 | 0.074 | 0.071 | 105 | *** |
| Operator | 0.108 | 0.075 | 105 | 0.088 | 0.071 | 105 | *** |
| Unskill | 0.480 | 0.120 | 105 | 0.527 | 0.121 | 105 | ** |

Note: This table compares labour market outcomes in districts with relatively high (i.e. above the median value) and low levels of ethnic diversity. The sample is only for the "white" magisterial districts which can be matched to 1985 census and whose black population accounts for more than 1% of the overall population. Employment outcomes are calculated from individual-level 2001 census data among the working-age black population. "Self-employed" refers to the proportion of self-employed people in each ethnic group over the whole working-age population of the corresponding ethnic group."Wage employee" and "unemployed" are calculated in similar ways. We only focus on people who are employed when comparing the allocation of workers across industrial sectors and occupations.

**Table 3.** Validity of the instrumental variable

| Dependent variable | 2001 |
|---|---|
| **Panel A: Job opportunities** | |
| Distance to the closest economic centre | -258,241.449 |
| | (264,611.697) |
| **Panel B: Economic activities of the white** | |
| Share of white who are self employed contemporarily | 0.041 |
| | (0.122) |
| Share of white who are wage employed contemporarily | 0.199 |
| | (0.164) |
| Proportion of white | 0.139 |
| | (0.148) |
| **Panel C: Path dependence** | |
| Share of white who are wage employed in 1980 | -0.192 |
| | (0.223) |
| Proportion of white in 1980 | -0.264 |
| | (0.273) |
| **Panel D: Contemporary migration** | |
| Number of migrants | 64,956.38 |
| | (66,800.72) |
| District controls | YES |
| Individual controls (district average) | YES |
| Province fixed effect | YES |
| Obs | 210 |

Note: This table conducts validity test of the instrumental variable based on 2001 census data. The sample is only for the "white" magisterial districts which can be matched to 1985 census and whose black population accounts for more than 1% of the overall population. All regressions are at the district level. We control for district-level variables especially geographical features, individual-level controls aggregated at district average and province fixed effects. *** p<0.01, ** p<0.05, * p<0.1.

**Table 4.** First-stage regression results

| | [1] Age 15-64 | [2] Age 25-64 |
|---|---|---|
| Predicted ELF | 1.569*** | 1.540*** |
| | (0.291) | (0.293) |
| Male | 0.001 | 0.001* |
| | (0.001) | (0.001) |
| Age | -0.000 | -0.000*** |
| | (0.000) | (0.000) |
| Married | 0.004*** | 0.003*** |
| | (0.001) | (0.001) |
| Father alive | 0.001** | 0.002** |
| | (0.001) | (0.001) |
| Pop density | 0.000** | 0.000** |
| | (0.000) | (0.000) |
| Urban | 0.005 | 0.004 |
| | (0.010) | (0.010) |
| River | 0.062** | 0.061** |
| | (0.028) | (0.029) |
| Density mine | 0.888 | 0.810 |
| | (0.728) | (0.708) |
| Prop black | -0.338*** | -0.338*** |
| | (0.068) | (0.067) |
| Distance closest | -0.000*** | -0.000*** |
| | (0.000) | (0.000) |
| Ruggedness | -0.006 | -0.005 |
| | (0.007) | (0.007) |
| Soil quality | 0.059** | 0.059** |
| | (0.027) | (0.027) |
| Per capita light | 0.672* | 0.651* |
| | (0.367) | (0.360) |
| Road | 0.017 | 0.021 |
| | (0.029) | (0.028) |
| Proportion manu | 0.275*** | 0.285*** |
| | (0.076) | (0.074) |
| Proportion service | 0.199** | 0.192** |
| | (0.089) | (0.086) |
| Ethnicity fixed effect | YES | YES |
| Province fixed effect | YES | YES |
| F-statistics of the instrument | 29.01 | 27.65 |
| R-squared | 0.108 | 0.103 |
| Observations | 697,369 | 484,639 |

Note: This table reports the first-stage results of the instrumental variable based on 2001 census data and report the F-statistics of the instrumental variable. The sample is only for the "white" magisterial districts which can be matched to 1985 census and whose black population accounts for more than 1% of the overall population. All regressions are at the individual level. We report all the control variables, both district-level variables especially geographical features and individual-level controls for socio-economic status. We control for ethnicity and province fixed effects. *** p<0.01, ** p<0.05, * p<0.1.

**Table 5.** Ethnic diversity and employment opportunities

| | [1] | [2] | [3] | [4] | [5] | [6] |
|---|---|---|---|---|---|---|
| | Unemployed + inactive | | Wage employment | | Self/wage | |
| | Age 15-64 | Age 25-64 | Age 15-64 | Age 25-64 | Age 15-64 | Age 25-64 |
| **Panel A: OLS estimates** | | | | | | |
| Ethnic fractionalisation ELF | -0.143*** | -0.144*** | 0.142*** | 0.143*** | 0.009 | 0.009 |
| | (0.037) | (0.043) | (0.038) | (0.044) | (0.013) | (0.013) |
| Individual controls | YES | YES | YES | YES | YES | YES |
| District controls | YES | YES | YES | YES | YES | YES |
| Province FE | YES | YES | YES | YES | YES | YES |
| R-squared | 0.161 | 0.093 | 0.159 | 0.092 | 0.008 | 0.007 |
| Observations | 697,369 | 484,639 | 681,529 | 470,552 | 253,809 | 228,519 |
| **Panel B: IV estimates** | | | | | | |
| Ethnic fractionalisation ELF | -0.206** | -0.251** | 0.208** | 0.257** | -0.032 | -0.031 |
| | (0.092) | (0.104) | (0.091) | (0.104) | (0.039) | (0.036) |
| Individual controls | YES | YES | YES | YES | YES | YES |
| District controls | YES | YES | YES | YES | YES | YES |
| Province FE | YES | YES | YES | YES | YES | YES |
| F statistics of the instrument | 29.01 | 27.65 | 29.36 | 27.98 | 23.92 | 24.68 |
| R-squared | 0.161 | 0.093 | 0.158 | 0.092 | 0.007 | 0.007 |
| Observations | 697,369 | 484,639 | 681,529 | 470,552 | 253,809 | 228,519 |

Note: This table reports results about the effect of ethnic diversity on unemployment rate, wage employment rate and the ratio of self-employment to wage employment at individual-level regressions based on 2001 census data. The sample is only for the "white" magisterial districts which can be matched to 1985 census and whose black population accounts for more than 1% of the overall population. We control for district-level variables especially geographical features, individual-level controls aggregated at district average and ethnicity fixed effects. We also control for province fixed effects. Ethnic diversity is measured with fractionalisation index. We include both unemployed and economically inactive groups as these two categories are combined in 2001 census. "Unemployed + inactive" is a dummy variable which equals 1 if one is unemployed or inactive and 0 if one is employed. In columns 3 and 4 we drop self-employed people as they are a very small proportion of the whole working-age population. Columns 5 and 6 are based only on the employed black people. *** p<0.01, ** p<0.05, * p<0.1.

**Table 6.** Decomposing ethnic diversity into number of group and dispersion of group size

| | [1] Wage employee | [2] College |
|---|---|---|
| **Panel A: Instrument dispersion of size** | | |
| Dispersion of size | -0.243*** | -0.089** |
| | (0.093) | (0.035) |
| 1/No. of groups | -0.405*** | -0.098** |
| | (0.132) | (0.045) |
| F statistics of the instrument | 28.17 | 28.17 |
| Observations | 681,529 | 681,529 |
| **Panel B: Instrument number of groups** | | |
| Dispersion of size | -3.203 | -2.718 |
| | (37.091) | (33.415) |
| 1/No. of groups | -16.991 | -14.826 |
| | (203.864) | (138.402) |
| F statistics of the instrument | 0.00652 | 0.00652 |
| Observations | 681,529 | 681,529 |
| Individual controls | YES | YES |
| District controls | YES | YES |
| Province FE | YES | YES |

Note: This table reports results based on the decomposition of ethnic diversity index into items relating to number of ethnic groups and group share, and how these two items are associated with employment rate and educational attainment at individual-level IV regressions based on 2001 census data. In Panel A we use our IV to instrument dispersion of group size while in panel B we use our IV to instrument number of groups. The sample is only for the "white" magisterial districts which can be matched to 1985 census and whose black population accounts for more than 1% of the overall population. We control for district-level variables especially geographical features, individual-level controls aggregated at district average and ethnicity fixed effects. We also control for province fixed effects. Ethnic diversity is measured with fractionalisation index. We look at the sub-samples split by group size. *** p<0.01, ** p<0.05, * p<0.1.

**Table 7.** Ethnic diversity, education and labour market outcomes

| | [1] |
|---|---|
| **Panel A: Ethnic diversity and education (IV)** | |
| Ethnic fractionalisation ELF | 0.087** |
| | (0.034) |
| F statistics of the instrument | 29.362 |
| R-squared | 0.059 |
| Obs | 681,529 |
| | |
| **Panel B: Education and wage employment, conditional on ethnic diversity (OLS)** | |
| College | 0.278*** |
| | (0.009) |
| R-squared | 0.174 |
| Obs | 681,529 |
| | |
| **Panel C: Ethnic diversity and wage employment by occupation (IV)** | |
| Manager | 0.029*** |
| | (0.011) |
| Profession | 0.104 |
| | (0.069) |
| Clerk | 0.040 |
| | (0.029) |
| Serve | -0.002 |
| | (0.034) |
| Craft | 0.026 |
| | (0.052) |
| Skilled agriculture | -0.066* |
| | (0.038) |
| Operator | -0.072* |
| | (0.039) |
| Unskilled | -0.059 |
| | (0.086) |
| Obs | 224,942 |
| | |
| Individual controls | YES |
| District controls | YES |
| Province FE | YES |

Note: This table reports the link between ethnic diversity, educational attainment and labour market outcomes based on 2001 census data. Panel A reports the effect of ethnic diversity on the probability of achieving a college degree. Panel B reports the link between a college degree and wage employment rate. Panel C only focus on employees to illustrate the allocation of employed workers across different occupations as a response to ethnic diversity. The sample is only for the "white" magisterial districts which can be matched to 1985 census and whose black population accounts for more than 1% of the overall population. We control for district-level variables especially geographical features, individual-level controls aggregated at district average and ethnicity fixed effects. We also control for province fixed effects. Ethnic diversity is measured with fractionalisation index. *** p<0.01, ** p<0.05, * p<0.1.

**Table 8.** Employment and education as a response to the proportion of own ethnic group

| | [1] OLS | [2] IV |
|---|---|---|
| **Panel A: Wage employment as dependent variable** | | |
| Proportion of own group | -0.096*** | -0.063** |
| | (0.013) | (0.028) |
| F statistics of the instrument | | 183.826 |
| R-squared | 0.159 | 0.159 |
| **Panel B: Education as dependent variable** | | |
| Proportion of own group | -0.025*** | -0.034*** |
| | (0.004) | (0.009) |
| F statistics of the instrument | | 183.826 |
| R-squared | 0.014 | 0.014 |
| Obs | 681,529 | 681,529 |
| Individual controls | YES | YES |
| District controls | YES | YES |
| Province FE | YES | YES |

Note: This table reports how wage employment rate and the chance of achieving a college degree (Panel A and Panel B, respectively) responds to the proportion of own ethnic group in the district. We use predicted proportion of own ethnic group as an instrumental variable for the corresponding proportion in reality. The sample is only for the "white" magisterial districts which can be matched to 1985 census and whose black population accounts for more than 1% of the overall population. We control for district-level variables especially geographical features, individual-level controls aggregated at district average and ethnicity fixed effects. We control for province fixed effects. *** p<0.01, ** p<0.05, * p<0.1.

# A Appendix. Bantu migration and the formation of ethnic diversity from historical narratives

Below we provide a summary of the history of the Bantu migration from central and eastern Africa and the settlement of these groups in South Africa for each ethnic groups in details. Narrative evidence is summarised from Mwakikagile (2010) and Gradin (2014).

| **Ethnicity** | **Time of migration into SA** | **Homelands** | **Time of moving into white areas** | **Bantustan** |
|---|---|---|---|---|
| Xhosa | Before 1400s | Today's Eastern Cape | After conflicts with the native Khoisan | Ciskei and Transkei |
| Zulu | 16th century | Eastern part, today's Kwazulu-Natal | Early 1800s | KwaZulu |
| Swazi | 15th and 16th centuries | Southern part of Tongaland in what is now Mozambique | 17th and 18th centuries into the Pongola River | KaNgwane |
| Ndebele | Before 1835 | Today's Northern Province, Mpumalanga and Gauteng | By 1835 towards Swaziland and Northern Transvaal | KwaNdebele, Lebowa |
| North Sotho | 1500s | Today's Limpopo and Northwest | After the war with Boers and Ndebele | Qwawa |
| South Sotho | 1500s | Today's Limpopo and Northwest | After the war with Boers and Ndebele | Qwawa |
| Tswana | 1500s | Today's Limpopo and Northwest | After the war with Boers and Ndebele | Bophuthatswana |
| Tsonga | Before the early 1500s | Close to today's Mozambique | After conflicts with Zulu | Gazankulu |
| Venda | Before 800s A.D. | A mountainous area in the northern part close to Limpopo River | 800s A.D. to Matopo Hills | Venda |

# B Appendix. Data source and construction of district-level variables

In this section we present data sources and the construction of our district-level control variables in detail. Emphasis has been given on those geographical measures.

| Variable | Data source | Construction of variable |
|---|---|---|
| **Panel A: From census** | | |
| Area of the district | Census 1996 and 2001 district-level shape file. | Calculated from the shape file directly in ArcGIS. |
| Population density | Census 1996 and 2001. | Calcualte the total number of black in each district in census data and divide it by area. |
| Proportion of the black | Census 1996 and 2001. | Calculate the number of black over the whole population. |
| Proportion of manufacturing | Census 1996 and 2001. | Calculate the number of people working in manufacturing sector over the whole employed black people. |
| Proportion of service | Census 1996 and 2001. | Calculate the number of people working in service sector over the whole employed black people. |
| Urban/rural | Census 1996 and 2001. | Information on whether one lives in an urban or rural settlement is explicit in census data. |

| Variable | Data source | Construction of variable |
|---|---|---|
| **Panel B: Sources on geography** | | |
| Overlap of district and homeland | A map (shape file) of homeland provided by Tim Brophy and Adrian Frith. | Intersecting the boundary of districts with that of homelands and seeing the overlap in ArcGIS. |
| River | Census 2001 river shape file. | Overlapping shape file of districts and river and directly calculating in ArcGIS. |
| Road | Census 2001 major road shape file. | Overlapping shape file of districts and road and directly calculating in ArcGIS. |
| Ruggedness | From Nunn and Puga (2012). We also tried the measure of slope from the same data source with similar results. | Same as Nunn and Puga (2012). |
| Soil quality | Harmonized World Soil Database. http://www.fao.org/soils-portal/soil-survey/soil-maps-and-databases/harmonized-world-soil-database-v12/en/. | Calculating average soil quality measures in a district (average of the index over grids in a district). |
| Density of mine | Mineral Resources Data System (MRDS) https://mrdata.usgs.gov/mrds/. | Overlapping shape file of districts and mines. Calculating number of mines in each district and dividing it by area. |
| Nighlight per capita | The National Oceanic and Atmospheric Administration night-time light satellite images. www.noaa.gov/stories/our-earth-at-night. | Calculating nightlight measures in a district (summation of the index over grids). Dividing it by the whole population in the district obtained from census data. |
| Distance from district to homeland | A map (shape file) of homeland provided by Tim Brophy and Adrian Frith. | Calculating Euclidean between centroid of districts and the border of homelands. |
| Distance to closest homeland | A map (shape file) of homeland provided by Tim Brophy and Adrian Frith. | Choosing the mininum value of the distance to all homelands. |

# C Appendix. Tables and figures

**Table A0.** Gravity model predicting the stock of black population in white districts: PPML estimator

| | Dependent variable: ethnic population $N_{kd}$ | | |
|---|---|---|---|
| | Coef. | Std. Err. | t-stat |
| Distance $Dis_{kd}$ | -.0039 | ( .0007) | -5.17 |
| *Ethnic group fixed effects:* | | | |
| Group 1 | .9750 | ( .2139) | 4.56 |
| Group 2 | .6133 | (.1769 ) | 3.47 |
| Group 3 | .1778 | (.2248 ) | 0.79 |
| Group 4 | -.4604 | (.2311 ) | -1.99 |
| Group 5 | .2220 | (.2259) | 0.98 |
| Group 6 | .8940 | (.1803) | 4.96 |
| Group 8 | .0469 | (.1833) | 0.26 |
| Group 9 | -.8184 | (.2776) | -2.95 |
| Constant | 9.157 | ( .2176) | 42.08 |
| R-squared | .092 | | |
| Observations | 1989 | | |

Note: This table reports results about the gravity model which helps estimate the stock of each ethnic group in each "white" district based on 1985 census data. The sample is for all the "white" magisterial districts which can be matched to 1996 and 2001 census. We control for homeland fixed effects and run a regression of the stock of ethnic groups on the distance between their corresponding homelands and each district using PPML models. *** p<0.01, ** p<0.05, * p<0.1.

**Table A1.** Summary statistics of demographics and employment among black ethnic groups in 1996

| | Population size | Share of the black population | Self employed | Wage employee | Unemployed | Inactive | Unemployed +inactive | ELF |
|---|---|---|---|---|---|---|---|---|
| Xhosa | 2229452 | 0.252 | 0.027 | 0.312 | 0.273 | 0.388 | 0.661 | 0.230 |
| | | | [0.162] | [0.463] | [0.445] | [0.487] | [0.474] | [0.301] |
| Zulu | 2073036 | 0.234 | 0.037 | 0.349 | 0.264 | 0.350 | 0.614 | 0.521 |
| | | | [0.189] | [0.477] | [0.441] | [0.477] | [0.487] | [0.273] |
| South Sotho | 2009582 | 0.227 | 0.026 | 0.349 | 0.242 | 0.383 | 0.625 | 0.467 |
| | | | [0.159] | [0.477] | [0.428] | [0.486] | [0.484] | [0.23] |
| Tswana | 1039138 | 0.117 | 0.026 | 0.384 | 0.224 | 0.367 | 0.590 | 0.525 |
| | | | [0.158] | [0.486] | [0.417] | [0.482] | [0.492] | [0.243] |
| North Sotho | 770110.7 | 0.087 | 0.038 | 0.412 | 0.236 | 0.314 | 0.549 | 0.700 |
| | | | [0.192] | [0.492] | [0.424] | [0.464] | [0.498] | [0.116] |
| Tsonga | 295688.6 | 0.033 | 0.075 | 0.434 | 0.247 | 0.244 | 0.491 | 0.715 |
| | | | [0.263] | [0.496] | [0.431] | [0.430] | [0.500] | [0.109] |
| Ndebele | 185065.8 | 0.021 | 0.040 | 0.358 | 0.227 | 0.375 | 0.602 | 0.689 |
| | | | [0.195] | [0.480] | [0.419] | [0.484] | [0.490] | [0.111] |
| Swazi | 181467.1 | 0.020 | 0.041 | 0.405 | 0.224 | 0.330 | 0.554 | 0.604 |
| | | | [0.198] | [0.491] | [0.417] | [0.470] | [0.497] | [0.166] |
| Venda | 80189.34 | 0.009 | 0.048 | 0.497 | 0.202 | 0.252 | 0.455 | 0.714 |
| | | | [0.214] | [0.500] | [0.402] | [0.434] | [0.498] | [0.113] |
| Overall | 8863729.5 | 1.000 | 0.032 | 0.355 | 0.251 | 0.361 | 0.613 | 0.274 |
| | | | [0.177] | [0.479] | [0.434] | [0.48] | [0.487] | [0.266] |

Note: The number and proportion of each ethnic group in the whole black population are calculated in the "white" magisterial districts which can be matched to 1985 census and whose black population accounts for more than 1% of the overall population. Population size is not always an integer because it is weighted by each person's weight in the census data. Employment outcomes are calculated from individual-level 1996 census data among the working-age blacks. "Self-employed" refers to the proportion of self-employed people in each ethnic group over the whole working-age population of the corresponding ethnic group. Other labour market outcomes are calculated in similar ways. The mean degree of ethnic diversity index is calculated at the district level. All other statistics are calculated at the individual level.

**Table A2.** Summary statistics of ethnic fragmentation and labour market outcomes in 1996

| | | High ELF | | | Low ELF | | |
|---|---|---|---|---|---|---|---|
| | Mean | S.d | Obs | Mean | S.d. | Obs | ttest |
| ELF | 0.507 | 0.134 | 102 | 0.044 | 0.071 | 103 | *** |
| Self employment | 0.028 | 0.044 | 102 | 0.021 | 0.044 | 103 | *** |
| Wage employee | 0.400 | 0.110 | 102 | 0.320 | 0.120 | 103 | *** |
| Unemployed | 0.570 | 0.110 | 102 | 0.658 | 0.118 | 103 | *** |
| Agriculture | 0.466 | 0.138 | 102 | 0.454 | 0.134 | 103 | |
| Manufacture | 0.115 | 0.105 | 102 | 0.090 | 0.105 | 103 | * |
| Service | 0.419 | 0.114 | 102 | 0.455 | 0.130 | 103 | ** |
| Manager | 0.014 | 0.032 | 102 | 0.012 | 0.045 | 103 | |
| Profession | 0.070 | 0.063 | 102 | 0.082 | 0.081 | 103 | * |
| Clerk | 0.032 | 0.056 | 102 | 0.020 | 0.049 | 103 | *** |
| Serve | 0.073 | 0.059 | 102 | 0.063 | 0.069 | 103 | * |
| Craft | 0.107 | 0.092 | 102 | 0.125 | 0.108 | 103 | |
| Skilld agriculture | 0.121 | 0.079 | 102 | 0.107 | 0.079 | 103 | * |
| Operator | 0.088 | 0.071 | 102 | 0.062 | 0.063 | 103 | *** |
| Unskill | 0.495 | 0.118 | 102 | 0.529 | 0.111 | 103 | ** |

Note: This table compares labour market outcomes in districts with relatively high (i.e. above the median value) and low levels of ethnic diversity. The sample is only for the "white" magisterial districts which can be matched to 1985 census and whose black population accounts for more than 1% of the overall population. Employment outcomes are calculated from individual-level 1996 census data among the working-age population. "Self-employed" refers to the proportion of self-employed people in each ethnic group over the whole working-age population of the corresponding ethnic group."Wage employee" and "unemployed" are calculated in similar ways. We only focus on people who are employed when comparing the allocation of workers across industrial sectors and occupations.

**Table A3.** Ethnic diversity and employment opportunities in 1996

| | [1] | [2] | [3] | [4] | [5] | [6] | [7] | [8] |
|---|---|---|---|---|---|---|---|---|
| | Unemployed | | Inactive | | Wage employment | | Self/wage | |
| | Age 15-64 | Age 25-64 | Age 15-64 | Age 25-64 | Age 15-64 | Age 25-64 | Age 15-64 | Age 25-64 |
| **Panel A: OLS estimates** | | | | | | | | |
| Ethnic fractionalisation ELF | -0.026 | -0.033 | -0.056** | -0.050** | 0.086** | 0.087** | -0.025 | -0.021 |
| | (0.018) | (0.027) | (0.024) | (0.023) | (0.033) | (0.042) | (0.018) | (0.018) |
| Individual controls | YES | YES | YES | YES | YES | YES | YES | YES |
| District controls | YES | YES | YES | YES | YES | YES | YES | YES |
| Province FE | YES | YES | YES | YES | YES | YES | YES | YES |
| R-squared | 0.033 | 0.072 | 0.147 | 0.09 | 0.184 | 0.114 | 0.01 | 0.009 |
| Observations | 473,258 | 325,539 | 473,258 | 325,539 | 457,915 | 311,644 | 184,385 | 165,874 |
| **Panel B: IV estimates** | | | | | | | | |
| Ethnic fractionalisation ELF | -0.146*** | -0.123* | 0.046 | -0.036 | 0.114 | 0.170* | -0.056 | -0.029 |
| | (0.047) | (0.064) | (0.064) | (0.066) | (0.079) | (0.096) | (0.044) | (0.042) |
| Individual controls | YES | YES | YES | YES | YES | YES | YES | YES |
| District controls | YES | YES | YES | YES | YES | YES | YES | YES |
| Province FE | YES | YES | YES | YES | YES | YES | YES | YES |
| F statistics of the instrument | 21.35 | 19.84 | 21.35 | 19.84 | 21.46 | 19.96 | 19.43 | 19.47 |
| R-squared | 0.032 | 0.071 | 0.147 | 0.090 | 0.184 | 0.114 | 0.010 | 0.009 |
| Observations | 473,258 | 325,539 | 473,258 | 325,539 | 457,915 | 311,644 | 184,385 | 165,874 |

Note: This table reports results about the effect of ethnic diversity on unemployment rate, the proportion of people out of labour force, wage employment rate and the ratio of self-employment to wage employment at individual-level regressions based on 1996 census data. The sample is only for the "white" magisterial districts which can be matched to 1985 census and whose black population accounts for more than 1% of the overall population. We control for district-level variables especially geographical features, individual-level controls aggregated at district average and ethnicity fixed effects. We also control for province fixed effects. Ethnic diversity is measured with fractionalisation index. We separate unemployed and economically inactive groups for 1996 results. In columns 5 and 6 we drop self-employed people as they are a very small proportion of the whole working-age population. Columns 7 and 8 are based only on the employed black people. *** p<0.01, ** p<0.05, * p<0.1.

**Table A4.** Ethnic diversity and different employment status: individual level regressions

| | [1] | [2] | [3] | [4] |
|---|---|---|---|---|
| | Self employment | | Wage employment | |
| | Age 15-64 | Age 25-64 | Age 15-64 | Age 25-64 |
| **Panel A: OLS estimates** | | | | |
| Ethnic fractionalisation ELF | 0.008* | 0.009* | 0.135*** | 0.135*** |
| | (0.004) | (0.005) | (0.037) | (0.043) |
| Individual controls | YES | YES | YES | YES |
| District controls | YES | YES | YES | YES |
| Province FE | YES | YES | YES | YES |
| R-squared | 0.01 | 0.007 | 0.146 | 0.082 |
| Observations | 697,369 | 484,639 | 697,369 | 484,639 |
| **Panel B: IV estimates** | | | | |
| Ethnic fractionalisation ELF | -0.000 | -0.001 | 0.206** | 0.252** |
| | (0.013) | (0.016) | (0.088) | (0.100) |
| Individual controls | YES | YES | YES | YES |
| District controls | YES | YES | YES | YES |
| Province FE | YES | YES | YES | YES |
| F statistics of the instrument | 29.01 | 27.65 | 29.01 | 27.65 |
| R-squared | 0.010 | 0.007 | 0.145 | 0.082 |
| Observations | 697,369 | 484,639 | 697,369 | 484,639 |

Note: This table reports results about the effect of ethnic diversity on self-employment rate and wage-employment rate at individual-level regressions based on 2001 census data using the full sample including self-employed people. The sample is only for the "white" magisterial districts which can be matched to 1985 census and whose black population accounts for more than 1% of the overall population. We control for district-level variables especially geographical features, individual-level controls aggregated at district average and ethnicity fixed effects. We also control for province fixed effects. Ethnic diversity is measured with fractionalisation index. *** p<0.01, ** p<0.05, * p<0.1.

**Table A5.** Ethnic diversity, intensive margin and wage: individual level regressions

| | [1] | [2] | [3] | [4] | [5] | [6] |
|---|---|---|---|---|---|---|
| | OLS | OLS | OLS | IV | IV | IV |
| | Log (monthly income) | Log (hourly wage) | Hour | Log (monthly income) | Log (hourly wage) | Hour |
| **Panel A: Individual level, census data 2001** | | | | | | |
| Ethnic fractionalisation ELF | 0.353*** | 0.392*** | -1.535 | 0.620** | 0.524 | 3.533 |
| | (0.093) | (0.107) | (1.285) | (0.276) | (0.341) | (4.148) |
| Individual controls | YES | YES | YES | YES | YES | YES |
| District controls | YES | YES | YES | YES | YES | YES |
| Province FE | YES | YES | YES | YES | YES | YES |
| F statistics of the instrument | | | | 24.28 | 24.28 | 24.17 |
| R-squared | 0.208 | 0.196 | 0.051 | 0.207 | 0.196 | 0.05 |
| Observations | 228,256 | 228,256 | 232,533 | 228,256 | 228,256 | 232,533 |
| **Panel B: Individual level, LFS data 1996** | | | | | | |
| Ethnic fractionalisation ELF | -0.117 | -0.0893 | 0.909 | 1.435 | 0.952 | 22.72 |
| | (0.349) | (0.347) | (2.960) | (1.484) | (1.245) | (16.85) |
| Individual controls | YES | YES | YES | YES | YES | YES |
| District controls | YES | YES | YES | YES | YES | YES |
| Province FE | YES | YES | YES | YES | YES | YES |
| F statistics of the instrument | | | | 4.823 | 4.667 | 4.914 |
| R-squared | 0.345 | 0.332 | 0.052 | 0.325 | 0.323 | 0.020 |
| Observations | 3,637 | 3,500 | 3,684 | 3,637 | 3,500 | 3,684 |

Note: This table reports results about the effect of ethnic diversity on other labour market outcomes at individual-level regressions, including working hour, hourly wage and monthly earnings. We only report the result in 2001 census as there is no information on hours of working in 1996 census. We report the result in 1996 labour force survey as there is no information on magisterial district in 2001 labour force survey. The sample is only for the "white" magisterial districts which can be matched to 1985 census and whose black population accounts for more than 1% of the overall population. We control for district-level variables especially geographical features, individual-level controls aggregated at district average and ethnicity fixed effects. We control for province fixed effects. Ethnic diversity is measured with fractionalisation index. All the columns only focus on employees. *** p<0.01, ** p<0.05, * p<0.1.

**Table A6.** Ethnic diversity and employment: district fixed effects models

| | [1] Unemploy + inactive | [2] Wage employ | [3] Self/wage |
|---|---|---|---|
| Ethnic fractionalisation ELF | -0.392*** | 0.445*** | -0.134* |
| | (0.075) | (0.075) | (0.070) |
| Individual controls (district average) | YES | YES | YES |
| District controls | YES | YES | YES |
| R-squared | 0.370 | 0.378 | 0.151 |
| Observations | 410 | 410 | 410 |

Note: This table reports results about the effect of ethnic diversity on employment based on the district-level balanced panel. The sample is only for the "white" magisterial districts which can be matched to 1985 census and whose black population accounts for more than 1% of the overall population. We control for district-level variables which vary over time and individual-level controls aggregated at district level and province fixed effects. Ethnic diversity is measured with fractionalisation index. The dependent variable in column 1 is the proportion of unemployed over the whole working-age black population. Column 2 is defined in a similar way but we exclude those who are self-employed. Column 3 has the dependent variable which is the ratio of the number of self-employed to that of employees at district level. *** p<0.01, ** p<0.05, * p<0.1.

**Table A7.** Comparing the importance of ethnic diversity and the overall black population

| | [1] Unemployed + inactive | [2] Wage employee |
|---|---|---|
| Ethnic fractionalisation ELF | -0.155** | 0.156** |
| | (0.075) | (0.075) |
| Black population | 0.340*** | -0.341*** |
| | (0.075) | (0.074) |
| Individual controls | YES | YES |
| District controls | YES | YES |
| Province FE | YES | YES |
| R-squared | 0.127 | 0.122 |
| F statistics | 12.974 | 12.974 |
| Observations | 697,369 | 681,529 |

Note: This table reports the results on the magnitude of the effect of ethnic diversity within the black and the proportion of black people in the district. Ethnic diversity is measured with fractionalisation index and is instrumented with the predicted diversity index. The proportion of black people is instrumented with the total distance to all homelands. The sample is only for the "white" magisterial districts which can be matched to 1985 census and whose black population accounts for more than 1% of the overall population. We control for district-level variables especially geographical features, individual-level controls aggregated at district average and ethnicity fixed effects. We also control for province fixed effects. *** p<0.01, ** p<0.05, * p<0.1.

**Table A8.** Intra-ethnic marriage rate and ethnic diversity: 1996 census

| | Mean | Std. Dev. | Obs |
|---|---|---|---|
| Own generation | 0.966 | 0.18 | 96,031 |
| Parental generation | 0.99 | 0.0966 | 532 |

Note: This table reports intra-ethnic marriage rate (i.e. marriage within ethnic groups within the black population). Ethnicity is identified from the first language spoken by both household head and spouse for the current generation, and household head's parents for the parental generation.

**Table A9.** Ethnic diversity and employment: district level regressions using spatially correlated standard errors

| | [1] Unemployed + inactive | [2] Wage employee | [3] Self/wage |
|---|---|---|---|
| **Panel A: OLS estimates** | | | |
| Ethnic fractionalisation ELF | -0.132*** | 0.136*** | -0.004 |
| | (0.043) | (0.044) | (0.020) |
| Individual controls | YES | YES | YES |
| District controls | YES | YES | YES |
| Province FE | YES | YES | YES |
| **Panel B: IV (GMM) estimates** | | | |
| Ethnic fractionalisation ELF | -0.192* | 0.216* | -0.137** |
| | (0.110) | (0.115) | (0.069) |
| Individual controls | YES | YES | YES |
| District controls | YES | YES | YES |
| Province FE | YES | YES | YES |

Note: This table reports results about the effect of ethnic diversity on employment and the allocation between self- and wage-employment at district-level regressions based on 2001 census data. The sample is only for the "white" magisterial districts which can be matched to 1985 census and whose black population accounts for more than 1% of the overall population. We control for district-level variables especially geographical features and individual-level controls. We also control for province fixed effects. Ethnic diversity is measured with fractionalisation index. Dependent variables are the proportion of people in each employment status over the whole working-age black population. We use Conley's standard errors with spatial correlations for both OLS and GMM analysis. We use 1000km as the cutoff value above which there is no spatial correlation. *** p<0.01, ** p<0.05, * p<0.1.

**Table A10.1.** Estimations based on non-linear econometric models

| | [1] | [2] | [3] | [4] | [5] | [6] |
|---|---|---|---|---|---|---|
| | Unemployed + inactive | | | Wage employment | | |
| | Logit | Probit | IV Probit | Logit | Probit | IV Probit |
| Ethnic fractionalisation ELF | -0.146*** | -0.144*** | -0.192** | 0.144*** | 0.142*** | 0.194** |
| | (0.039) | (0.039) | (0.093) | (0.039) | (0.039) | (0.092) |
| Observations | 697,369 | 697,369 | 697369 | 681,529 | 681,529 | 681,529 |
| Individual controls | YES | YES | YES | YES | YES | YES |
| District controls | YES | YES | YES | YES | YES | YES |
| Province FE | YES | YES | YES | YES | YES | YES |

Note: This table reports results about the effect of ethnic diversity on employment based on non-linear econometric models in 2001. The sample is only for the "white" magisterial districts which can be matched to 1985 census and whose black population accounts for more than 1% of the overall population. We control for district-level variables especially geographical features, individual-level controls and ethnicity fixed effects. We also control for province fixed effects. Ethnic diversity is measured with fractionalisation index. In column 4, 5 and 6 we drop self-employed people as they are a very small proportion of the whole working-age population. *** p<0.01, ** p<0.05, * p<0.1.

**Table A10.2.** Estimations based on multinomial econometric models

| | [1] | [2] |
|---|---|---|
| | Mlogit | |
| | Self employment | Wage employee |
| Ethnic fractionalisation ELF | 0.010* | 0.136*** |
| | (0.005) | (0.039) |
| Observations | 697,369 | 697,369 |
| | | |
| Individual controls | YES | YES |
| District controls | YES | YES |
| Province FE | YES | YES |

Note: This table reports results about the effect of ethnic diversity on employment based on multinomial econometric models in 2001. The sample is only for the "white" magisterial districts which can be matched to 1985 census and whose black population accounts for more than 1% of the overall population. We control for district-level variables especially geographical features, individual-level controls and ethnicity fixed effects. We also control for province fixed effects. Ethnic diversity is measured with fractionalisation index. *** p<0.01, ** p<0.05, * p<0.1.

**Table A11.** Ethnic diversity and the emigration of the white

| | [1] | [2] | [3] |
|---|---|---|---|
| | Number of white in 2001 | Number of white in 1985 | Diffrence: 01 - 85 |
| **Panel A: OLS estimates** | | | |
| Ethnic fractionalisation ELF | -30,372.282 | -53,687.943 | 23,315.661 |
| | (31,090.467) | (39,165.874) | (27,889.385) |
| R-squared | 0.726 | 0.747 | 0.513 |
| Observations | 210 | 210 | 210 |
| | | | |
| **Panel B: IV estimates** | | | |
| Ethnic fractionalisation ELF | 138,705.471 | 226,189.810 | -87,484.339 |
| | (147,383.227) | (154,628.703) | (59,278.370) |
| F statistics of instruments | 25.36 | 25.36 | 25.36 |
| R-squared | 0.691 | 0.672 | 0.445 |
| Observations | 210 | 210 | 210 |
| Individual controls at district level | YES | YES | YES |
| District controls | YES | YES | YES |
| Province FE | YES | YES | YES |

Note: This table looks at whether ethnic diversity is correlated with the number of white population in 2001 and the emigration of the white from the district after the end of Apartheid at district-level regressions. The sample is only for the "white" magisterial districts which can be matched to 1985 census and whose black population accounts for more than 1% of the overall population. We control for district-level variables especially geographical features and individual-level controls aggregated at district average. We also control for province fixed effects. Ethnic diversity is measured with fractionalisation index. *** p<0.01, ** p<0.05, * p<0.1.

**Table A12.** Heterogenous effects of ethnic diversity on wage employment with respect to group sizes

| | [1] Wage employee | [2] College |
|---|---|---|
| ELF * small | -0.211*** | -0.087*** |
| | (0.063) | (0.027) |
| Ethnic fractionalisation ELF | 0.220** | 0.092*** |
| | (0.091) | (0.035) |
| Individual controls | YES | YES |
| District controls | YES | YES |
| Province FE | YES | YES |
| F statistics of the instrument | 14.654 | 14.654 |
| R-squared | 0.159 | 0.011 |
| Obs | 681,529 | 681,529 |

Note: This table reports results on how the effect of ethnic diversity differs across group sizes based on 2001 census data. We interact the ethnic diversity index with whether each individual belongs to the four "small" groups. This interaction term is also instrumented with an interaction between the predicted diversity index and whether one is from a "small" group. The sample is only for the "white" magisterial districts which can be matched to 1985 census and whose black population accounts for more than 1% of the overall population. We control for district-level variables especially geographical features, individual-level controls and ethnicity fixed effects. We also control for province fixed effects. Ethnic diversity is measured with fractionalisation index. *** p<0.01, ** p<0.05, * p<0.1.

**Table A13.** Heterogenous effects of ethnic diversity on wage employment with respect to particular groups

| | [1] Tsonga | [2] Ndebele | [3] Swazi | [4] Venda |
|---|---|---|---|---|
| ELF*Prop of the group | 0.191 | -0.580*** | 0.261* | -0.075 |
| | (0.245) | (0.205) | (0.147) | (0.194) |
| Ethnic fragmentation index ELF | 0.203** | 0.266** | 0.208** | 0.210** |
| | (0.095) | (0.103) | (0.091) | (0.093) |
| F statistics of the instrument | 15.32 | 11.061 | 14.274 | 14.488 |
| R-squared | 0.158 | 0.159 | 0.158 | 0.158 |
| Obs | 681,529 | 681,529 | 681,529 | 681,529 |
| Individual controls | YES | YES | YES | YES |
| District controls | YES | YES | YES | YES |
| Province FE | YES | YES | YES | YES |

Note: This table reports results on how the effect of ethnic diversity differs when the proportions of specific small groups are different based on 2001 census data. We interact the ethnic diversity index with the proportion of each "small" groups in the same district respectively. This interaction term is also instrumented with an interaction between the predicted diversity index and the proportion of each "small" group. The sample is only for the "white" magisterial districts which can be matched to 1985 census and whose black population accounts for more than 1% of the overall population. We control for district-level variables especially geographical features, individual-level controls and ethnicity fixed effects. We also control for province fixed effects. Ethnic diversity is measured with fractionalisation index. *** p<0.01, ** p<0.05, * p<0.1.