ALL JOBS ARE EQUAL, BUT SOME JOBS ARE MORE EQUAL THAN OTHERS: WHAT A CLUSTERING ALGORITHM REVEALS ABOUT LABOUR MARKET SEGMENTATION IN SOUTH AFRICA

Jonathan Matthew Kensett

A minor dissertation submitted in partial fulfilment of the degree of Master of Commerce specialising in Economics

School of Economics

Faculty of Commerce

University of Cape Town

March 2021

Supervisors: Prof. Vimal Ranchhod & Prof. Murray Leibbrandt

1. Introduction

It is now a truism that South Africa has one of the highest levels of measured income inequality in the world, with recent evidence suggesting that it has increased in the post-apartheid period (Wittenberg, 2017). Despite well-designed and well-targeted policies, this is alarming and suggests that the project to create a more equal society seems to be stalling (Leibbrandt, Ranchhod & Green, 2018). An important question therefore arises: why has South Africa's income inequality persisted?

The most dominant contributor to income inequality is inequality within the South African labour market, which is the most prominent source of income for South African households (Leibbrandt *et al*., 2018). A possible explanation, therefore, for the persistence of income inequality is deep structural issues within the labour market which give rise to significant rigidities that prevent real wages from adjusting to market clearing levels (Fedderke, 2012). These rigidities indicate that South Africa's labour market is segmented. Labour market segmentation arises when jobs for individuals with a given skillset differ in terms of observable characteristics, while access to more attractive jobs is limited (Fields, 2011). This segmentation gives rise to significant wage differentials between sectors as frictions prevent labour market participants from transitioning between segments and wages equalising (Heintz & Posel, 2008).

A standard example of segmentation is that between the formal and informal labour markets. In this paper I make use of a clustering algorithm to cluster employed individuals into distinct groups based on their similarities with regards to several key characteristics identified in the literature that are related to formal and informal employment, such as labour market income, hours worked, whether an employer deducts tax, whether an employee receives medical aid or pension benefits, and an employees' trade union status.

Using data from the National Income Dynamics Study (NIDS), which is the first nationally representative panel survey conducted in South Africa, the clustering algorithm identifies three distinct groups of employed individuals. Characteristics of two of the groups resemble individuals who would typically be classified as informally or formally employed. A third group is also identified that includes individuals who can neither be strictly classified as informally nor formally employed. The characteristics of this group

seems to straddle formal and informal employment and is therefore referred to what has been identified by some researchers as semiformal employment (Cobb *et al*., 2009). This suggests the existence of a spectrum of informality in the South African labour market, as opposed to a clear dichotomy between informal and formal employment.

I find that informal employment is associated with elementary occupations mostly in private households, retail, and agriculture. Workers in this category are also more likely to have no or little schooling. Semiformal employment is associated with elementary occupations in retail, manufacturing and community and social services. These workers are more likely to have a matric and much of the employment growth observed post 2008 was skewed towards semiformal employment. While formal employment is associated with professional employment mostly in mining, manufacturing, transport and community and social services. Formally employed workers are more likely to have some form of tertiary education. Furthermore, those who are formally employed earn considerably more than those who are semiformally or informally employed. On average, formal employees earn 4.5 and 2.5 times more than informal and semiformal workers, respectively. Semiformal workers earn roughly 1.5 times more than informal workers.

There are also gendered and racial differences between these three groups. Women and black individuals are more likely to be informally and semiformally employed. Individuals in formal employment are disproportionately male and are more likely to be white.

I also analyse the transition between labour market states over time. Overall, I find that out of all three groups, informally employed individuals are least likely to remain employed and formally employed individuals are most likely to stay employed. I also find a significant degree of churning between informal and semiformal employment, however relatively few individuals enter formal employment from informal or semiformal employment. This suggests that much of the churning observed in the South African labour market is between informal and semiformal jobs, rather than between informal and formal jobs.

The rest of the paper is structured as follows: I begin with an overview of the literature in section 2. Section 3 describes the clustering methodology used to identify labour

market segments; section 4 describes the data used. Results are presented in section 5 and discussed in section 6.

2. Literature Review

2.1. Labour market segmentation in South Africa

Inequality within the South African labour market is the largest contributor to overall income inequality. Therefore, any attempt to reduce inequality needs to have a core focus on understanding the structure of our labour market (Leibbrandt *et al*., 2018). To do so, it is useful to understand how wages are determined in labour markets.

Standard neo-classical theory assumes that firms compete for labour in a single labour market and profit maximising firms in a perfectly competitive market hire workers based on their individual observed characteristics and offer wages equal to a worker's productivity. A prediction of this theory is that that over time differences between groups of workers will disappear due to competitive mechanisms and assuming workings can invest in their productivity (Reich et al., 1973). One would therefore expect wage inequality to decrease as worker productivity converges.

However, South Africa is characterised by a wide range of rigidities which prevent wages from converging, contributing to a poor functioning labour market, evidenced by our high unemployment levels (Fedderke, 2012). Additionally, income inequality has not decreased and there is evidence to suggest that it may have in fact increased (Wittenberg, 2017). Given that wage differences between groups of workers is not disappearing, conventional neo-classical theory does not provide a suitable explanation for the levels of inequality we observe in and the structure of our labour market.

A useful alternative is the theory of labour market segmentation, which suggests that not all workers and firms in an economy participate in the same labour market. Instead, they participate in several smaller homogenous labour markets with different working conditions and wages. Additionally, there are frictions between segments which prevent workers from entering the segment which offers higher wages, which prevents wages converging and labour markets clearing (Fields, 2011).

Although labour market segmentation is not unusual, what is unusual in the South African context is the degree to which frictions between segments give rise to high wage

differentials, high unemployment, and consequently high wage inequality (Fedderke, 2012). Therefore, to enhance our understanding of inequality it is helpful to understand the source of these rigidities, so that they can be lowered with the goal of reducing inequality.

Perhaps the most standard form of segmentation is that between the formal and informal labour market (Fedderke, 2012). Therefore, this paper will focus on identifying formal informal segments within the South African labour market to try and understand the source of the rigidities that resulted in this segmentation.

2.2. The link between labour market segmentation and inequality

According to Rubery (2003), labour markets are socially constructed phenomena. They are shaped by social conditions, which give rise to specific institutional arrangements. These institutional arrangements shape the way labour markets are structured, which, in turn, shapes how income is distributed in a society. Income inequality is then perpetuated by the institutions, which gave rise to the segmentation of labour markets. Therefore, given that labour markets in an economy shape income inequality, it is important to understand the structure of labour markets to reduce inequality.

According to Reich et al. (1973), in line with Rubery's (2003) explanation of labour markets as a social phenomenon, segmentation arises through historical processes whereby political forces give rise to different labour market segments, distinguished by different characteristics. Some initial segregation occurs, which separates a population into groups which resulted in a segmented labour market. This segregation can be thought to arise from discrimination (Bergmann, 1974).

To explain how discrimination results in labour market segmentation, Bergmann (1974) developed the occupational crowding model. This model attempts to explain how initial segregation in a population, which arises from discrimination, gives rise to segmentation in the labour market. This segregation then becomes embedded in the economy's price and productivity structure, resulting in wage differentials between those who work in different segments. This model is a useful lens to understand the nature of segmentation in South Africa, given our history of discrimination in which the apartheid government actively sought to create productivity differentials between race groups (Pellicer & Ranchhod, 2020).

Bergmann (1974) posits that discrimination restricts workers to a limited number of sectors. This results in an over-supply of labour to (or crowding into) those sectors which pushes down the equilibrium wage. These low wages, in turn, disincentivise attempts by discriminated workers to invest in their productivity. Meanwhile, some workers are not discriminated against and can compete in any sector. The firms in sectors for whom only the non-discriminated group can enter compete for a limited supply of workers.

Consequently, this pushes wages up. This increase in wages incentivises this group of workers to invest in their productivity. Over time, as this labour market structure persists, the wage gap between those who are and who are not discriminated against arising increases. Bergmann (1974) therefore concludes that wage inequality is not the consequence of differences in potential productivity (as conventional theory would suggest) but rather arises from the discriminatory process of crowding workers into certain sectors and crowding them out of other sectors, which might be reserved expressly for the group that is not discriminated against.

This initial segregation maintains income inequality, even if discrimination no longer persists. According to Bergmann (1974), segregation creates the initial market conditions leading to lower pay in sectors dominated by previously discriminated workers. If these workers choose to enter in other sectors, which previously they were unable to, potential employers treat the low wage as the opportunity wage for this worker type and adjust the wage offer downwards. Through this mechanism, wage inequality persists, even in the absence of discriminatory practices.

Applying this model to the South African context suggests that wage inequality in the labour market can be explained by the historical racial discrimination of African, Coloured, and Indian workers, which crowded them into certain sectors while crowding them out of sectors reserved for Whites. The mechanisms explained above resulted in a divergence of wages, which has been maintained well into democratic South Africa, given the persistence of our income inequality.

2.3. Identifying informally and formally employed workers.

The previous section proposes an explanation for how segmentation might have arisen in South Africa. I focus on the segmentation between the informal and formal labour markets since racial disparities exist between these two sectors. For example, Africans

and Coloureds are disproportionately more likely to be employed in the informal sector (Essop & Yu, 2008). Using a segmented labour market approach relies on differentiating between these segments. Thus, it is necessary to distinguish between formal and informal employees, which is not a straightforward exercise.

Distinguishing between these two sectors is difficult since there are many dimensions to informality. For example, in South Africa, research has mostly distinguished between these two groups using either an enterprise-based definition or an employee-based definition depending on the researcher's goals and unit of analysis. The former focuses on informal enterprises and defines an enterprise as informal if they are not registered for tax. Sometimes this definition is restricted to unregistered enterprises with fewer than five employees (See, for example, Fourie (2018b) and Rogan and Skinner (2018)).

An alternative approach to identifying an individual as employed in the informal sector is using an employee-based definition. If an employee does not have tax deducted from their salary or does not receive any benefits, or does not have a written contract, they are classified as informally employed. Otherwise, they are regarded as formally employed. This definition has been used by Nackerdien and Yu (2019) in their analysis of the formal-informal sector labour market linkages in South Africa. They also regard casually employed individuals as employed in the informal sector.

However, using a single observed variable to identify segments can limit policy's robustness and success for two reasons. Firstly, using a single variable requires the researcher to specify an arbitrary threshold for that variable (Makaluza and Burger, 2018). Doing so risks making the analysis too dependent on the researcher's discretion, and, secondly, by using a single factor, one risks ignoring the multidimensionality of jobs in the informal sector.

To help overcome these issues, I employ a clustering technique performed by Makaluza and Burger (2018), who use a clustering algorithm to identify distinct groups in datasets. It identifies which points in a dataset are related and then automatically assigns these data points to some number of distinct groups. This technique is favourable because it can detect and identify homogenous groups within the informal sector without identifying thresholds to distinguish between and uses multiple factors to do so.

3. Methodology

This paper aims to improve our understanding of labour market segmentation in South Africa, with a particular focus on formal and informal employment. To achieve this, it is necessary to identify the segments to which participants belong within the South African labour market—identifying informal and formal employees, as outlined in the previous section. To circumvent some of these issues, I make use of a data-driven approach to identify distinct segments of individuals within the labour market. This is a useful data-driven method of finding groups of individuals that are alike (VanderPlas, 2016) without having to impose an arbitrary cut off point between individuals (Makaluza & Burger, 2018).

A *K*-means clustering algorithm is a common technique used to perform cluster analysis. This algorithm divides *M* observational units in *N* dimensions into *K* clusters such that the sum of squares from the observational units to the assigned cluster centres is minimised. At this minimum, all cluster centres are at the arithmetic mean of the set of observational units, which are at the nearest to their respective cluster centres, and each point is closer to its cluster centre than to other cluster centres (Hartigan & Wong, 1979). In other words, given a set of *N* explanatory variables[1], the algorithm will assign *M* observational units to a cluster and units within this cluster are alike based on the explanatory variables that are used in the analysis.

To implement the *K*-means algorithm, the researcher first needs to identify the number of clusters *(K)*. The algorithm then randomly initialises *K* cluster centres, and then assigns each observational unit to its nearest cluster centre[2]. Once all observational units have been assigned to a cluster, the cluster centres are then updated to the mean of all the observational units contained within them. The algorithm then assigns each observational unit to its nearest updated cluster centre. This step is repeated until the sum of squared distance between each observational unit and cluster centre is minimised (VanderPlas, 2016). In other words, at the minimum, updating the cluster centres and

[1] The similarity of individuals is based on the researcher's set of explanatory variables. To avoid confusion and emphasise that these variables were the set of explanatory variables used to perform the cluster analysis, as opposed to say the explanatory variables used for imputation (where I use traditional econometric techniques), I will refer to these variables as cluster variables.

[2] The distance between an observational unit and any given cluster centre is measured using the Euclidean distance between the two points. The nearest cluster to an observational unit is the cluster that minimises the Euclidean distance (Hartigan & Wong, 1979).

then reassigning observational units to their nearest clusters should not change the clusters to which the observational units belong.

A drawback of the *K*-means algorithm is that it is sensitive to outliers (Arora & Varshney, 2016). I address this issue by identifying outliers and replace outliers with their expected value conditional on observable characteristics[3]. Another limitation is that K-means clustering may perform poorly when clustering binary variables since it assumes data are continuous, however, according to Finch (2005) clustering solutions can become more accurate with the more clustering variables and the larger the sample size.

4. Data

4.1. Data description

The National Income Dynamic Survey (NIDS) is a panel study that has been conducted over five waves in two to three-year intervals between 2008 and 2017 by the Southern African Labour and Development Research Unit (SALDRU). It is the first nationally representative panel study conducted in South Africa. The survey covers a wide range of topics related to individual, household, labour market characteristics, and other socioeconomic well-being topics. The survey repeatedly sampled the same household members from Wave 1. Waves 1, 2, 3, 4 and 5 successfully sampled approximately 28 000, 27 000, 29 00, 32 000 and 30 000 individuals, respectively (Brophy *et al*., 2018). Weights provided in the public release of the NIDS data are used to correct for attrition of individuals between waves and ensure that estimates are unbiased and nationally representative of the South African population (Branson & Wittenberg, 2018).

While the data allows for a detailed analysis of the link between informal earnings and poverty reduction in South Africa, there are limitations to accurately identifying employees in the informal sector. Firstly, the questionnaire does not explicitly distinguish between individuals employed in the informal sector and those in the formal sector. This makes it difficult to estimate the sizes of these two sectors accurately. Secondly, it is impossible to distinguish between own-account workers and employers (Cichello & Rogan, 2018). It does, however, identify casual workers, and most workers in this

[3] Following Wittenberg (2014) treatment of wages and hours worked, I regard wage observations with an absolute studentised residual great than five as outliers. I then impute the value of these outliers to their expected value conditional on age, race, gender, industry, occupation, province, and geographic location. I restrict hours worked to 98 per week.

category would likely be informal employees, but again, it is not possible to discern with absolute certainty whether these employees are in the formal or informal sector. Given the possible complexity of trying to discern, who is formally and informally employed based on a single explanatory variable might not be appropriate.

Using a set of variables might better identify formal and informal workers and factors in the multidimensionality of informality (Makaluza & Burger, 2018). It is, therefore, necessary to choose the variables that will be used for clustering – which I will refer to as the cluster variables. Clustering variables are used to distinguish between different segments in the informal sector. It is important to acknowledge that these variables are chosen by the researcher and, therefore, may introduce a degree of subjectivity to the analysis; however, it is not possible to perform our analysis without doing so.

For guidance, I refer to Makaluza and Burger (2018). They chose their clustering variables based on a review of the literature on the characteristics of the informal sector jobs. It is important not to choose variables that will be variables of interest when analysing individuals' characteristics in various segments, as emphasised in Makaluza and Burger (2018)[4]. Based on their choice of variable and data availability, I use the following as clustering variables to identify different segments: hours worked per week, monthly labour income, whether an individual has a pension or medical aid deducted from their salary (which proxy for employment benefits), whether an individual contributes to the Unemployment Insurance Fund, whether an individual is unionised, and whether they work for a company that is registered for income or value-added tax.

I restrict the sample to employed individuals between the ages of 15 and 65 in each wave. This yields a sample size of 37 617 working-age employed individuals across all five waves. In Table 1 below, I summarise the clustering variables by wave.

[4] For example, gender and education would not be good clustering variables because we want to know how gender and education influence an individual's choice to enter or exit the informal sector.

Table 1: Summary of clustering variables

Clustering variables	Wave 1	Wave 2	Wave 3	Wave 4	Wave 5
Weekly hours worked	36,68	36,00	39,06	39,67	38,99
Labour income*	3 012	3 603	3 904	3 669	3 835
Hourly labour income*	23,20	32,27	26,21	25,69	25,84
Medical aid	0,25	0,27	0,25	0,21	0,22
Pension	0,48	0,44	0,44	0,44	0,42
UIF	0,65	0,64	0,62	0,66	0,64
Unionised	0,33	0,31	0,30	0,30	0,29
Tax and/or VAT	0,25	0,27	0,25	0,15	0,24
Total no. employed (N)	5 781	5 125	7 099	9 118	9 661

Note: All values are proportions unless otherwise indicated. * Indicates a median instead of an average. All earnings data have been inflated to March 2017 prices. Data are weighted using post-stratified weights.

Source: Author's calculations using NIDS Waves 1 to 5.

4.2. Dealing with missing data

Cluster analysis, like standard regression analysis, is only performed on observations for which there is missing data for any of the clustering variables (or explanatory variables in the regression analysis case). This section deals with missing data and how to treat it when performing cluster analysis. Table 2a below shows a non-trivial amount of missing data for the clustering variables. Labour market income and hours worked to have the least missing data, missing about 3 and 12 percent, respectively. The variables related to benefits, UIF and unionisation, have approximately 30 percent of their observations missing, while almost 90 percent of the responses to income tax are missing. If one were to cluster this data, a significant amount of information would be lost, resulting in bias and a significant drop in the sample size. This is evident in table 2b, which shows the averages of the clustering variables for employed individuals who respond to all related questions. For wave 1, the potential sample drops from 5 781 employed individuals to only 73. A similar pattern holds for the rest of the waves. Suppose we do not account for the fact that less than 1 percent of the observations have responses for all seven clustering variables. Performing a cluster analysis might yield less efficient and biased results (Donders et al., 2006).

Table 2a: Percent of missing values for each cluster variable

	No. Missing	N	Percent Missing
Weekly hours	4 555	36 784	12,4
Labour income	1 247	36 784	3,4
Medical aid	11 212	36 784	30,5
Pension	11 406	36 784	31,0
UIF	11 525	36 784	31,3
Unionised	11 245	36 784	30,6
Tax and/or VAT	32 504	36 784	88,4

Table 2b: Average for only complete cases of the cluster variables

Clustering variables	Wave 1	Wave 2	Wave 3	Wave 4	Wave 5
Weekly hours	55	70	59	55	48
Labour income*	4 439	8 612	11 322	7 399	11 000
Medical aid	0,24	0,34	0,19	0,30	0,28
Pension	0,36	0,33	0,50	0,53	0,33
UIF	0,72	0,44	0,79	0,58	0,63
Unionised	0,19	0,24	0,22	0,35	0,23
Tax and/or VAT	0,17	0,32	0,18	0,14	0,32
Total no. employed (N)	73	55	50	103	96

Note: * indicates a median instead of an average. The values are weighted using post-stratified weights.

Source: Own calculations using NIDS Waves 1 to 5.

I deal with missing data by imputing the missing values using a single imputation procedure to estimate the expected value of the missing values conditional on an individual's observable characteristics[5]. One can then perform their analysis as if all the imputed values were observed; however, this imputation procedure can underestimate the standard errors and an overestimation of the precision of the estimates for our population parameters (Donders et al., 2006). However, for cluster analysis, there is no well-characterised measure of uncertainty that can be presented, and it is unclear how to incorporate uncertainty due to imputations into the results of the cluster assignment (Basagaña et al., 2013). Unfortunately, in the absence of an uncertainty estimate, it is

[5] For the regression imputation, I include the following as explanatory variables: age, race, gender, industry, occupation, province, and geographic location

difficult to ascertain how well the clustering algorithm performed. This is an important limitation to using this technique[6].

Table 3: Average for only complete cases of the cluster variables after missing values have been imputed

Clustering variables	Wave 1	Wave 2	Wave 3	Wave 4	Wave 5
Weekly hours	37	36	39	40	39
Labour income*	2 853	3 617	3 904	3 655	3 841
Medical aid	0,23	0,25	0,23	0,20	0,21
Pension	0,43	0,41	0,41	0,40	0,41
UIF	0,60	0,61	0,60	0,62	0,62
Unionised	0,32	0,30	0,31	0,29	0,28
Tax and/or VAT	0,25	0,25	0,26	0,24	0,23
Total no. employed (N)	5 751	5 050	7 050	9 095	9 570

Note: * indicates a median instead of an average. The values are weighted using post-stratified weights.

Source: Own calculations using NIDS Waves 1 to 5.

Table 3 presents each clustering variable's average for all complete cases after missing values are imputed. It can be observed that the sample size has only marginally declined. For example, the total number of those employed in wave one after imputation is 5 751, as opposed to 73 shown in Table 2b. The values are also significantly different from those in table 2b and indicate that it is likely that the results would be biased without imputing missing values.

5. Results

5.1. A description of the clusters

The clustering algorithm[7] is applied to all employed individuals in Wave 1 of NIDS. It was identified that the optimal number of clusters was three[8]. Each wave was then clustered separately. The result was five cross-sectional datasets where all employed individuals

[6] Some techniques have been used to address this. See Basagaña et al. (2013), who use a multiple imputation approach to create a pooled estimate of uncertainty. Unfortunately, due to limited computation capacity, I could not perform this check.

[7] I performed the k-means cluster analysis in R using the "miclust" package created by Barrera-Gómez & Basagaña (2014). I only used one imputed dataset and did not use multiple imputed datasets due to insufficient computing power.

[8] See section 8.1 in the appendix for a more detailed explanation for choosing the optimal number of clusters.

were assigned to one of three clusters. I present averages of each cluster variable for each cluster by wave in table 4 below.

Table 4: Summary statistics of cluster variables by cluster and wave

	Cluster	Mean Wages	Median Wages	Mean Hours	Median Hours	Medical aid	Pension	UIF	Union	Tax or VAT
Wave 1	1	2 622	1 585	33	38	0,02	0,08	0,00	0,11	0,16
	2	4 391	2 470	38	40	0,03	0,31	1,00	0,12	0,19
	3	11 875	7 266	40	40	0,67	0,94	0,86	0,77	0,41
Wave 2	1	3 910	2 009	33	38	0,03	0,05	0,00	0,10	0,16
	2	5 586	3 445	36	40	0,04	0,31	1,00	0,12	0,18
	3	13 354	8 612	39	40	0,73	0,94	0,85	0,73	0,43
Wave 3	1	3 699	2 342	36	40	0,02	0,07	0,00	0,10	0,16
	2	4 894	3 268	39	40	0,02	0,31	1,00	0,12	0,19
	3	14 026	9 110	42	40	0,71	0,91	0,78	0,78	0,44
Wave 4	1	3 095	2 130	35	40	0,03	0,07	0,00	0,10	0,16
	2	5 048	3 550	41	42	0,02	0,30	1,00	0,11	0,19
	3	14 655	10 651	44	44	0,67	0,93	0,80	0,77	0,42
Wave 5	1	3 495	2 400	36	40	0,02	0,07	0,00	0,10	0,15
	2	5 432	3 500	40	42	0,02	0,31	1,00	0,10	0,16
	3	14 234	10 000	43	41	0,70	0,94	0,81	0,76	0,42

Note: Includes imputed observations. Following Wittenberg (2014) wage observations with an absolute studentised residual great than 5 are regarded as outliers. These values have been imputed to their expected value conditional on age, race, gender, industry, occupation, province, and geographic location. Wages are real with a base year of March 2017. Hour observations greater than 98 were truncated to 98 following Wittenberg (2014).

Source: Author's calculations using NIDS waves 1 to 5.

Focusing on wave 1, it can be observed that both mean and median wages are different for each cluster, and they appear to be increasing, with the lowest wages for cluster 1 and the highest for cluster 3. Median wages for cluster 2 and 3 are 1.6 and 4.6 times higher than median wages for cluster 1. There also appears to be a positive relationship between average weekly hours and cluster number, with those in cluster 1 working fewer hours on average than those in clusters 2 and 3. Median hours for clusters 2 and 3 are the same at 40 hours per week, while median hours worked per week for cluster 1 is slightly less.

There is also a pattern in the proportion of those who receive work benefits, proxied by medical aid and pension. Roughly 2 and 3 percent of those in cluster 1 and cluster 2 have medical aid deducted from their salary, while about two-thirds of those in cluster 3 have

medical aid deductions. About 8 percent of individuals in cluster 1 have either a pension or provident fund deducted from their salaries, while 31 percent of individuals in cluster 2 have a pension deducted from their salaries. Almost all individuals in cluster 3, 94 percent, have a pension deducted from their salaries.

Interestingly no individuals in cluster 1 contribute towards UIF, while all individuals in cluster 2 contribute and 86 percent of individuals in cluster 3 contribute. Furthermore, about the same proportion, just over 10 percent, of individuals in clusters 1 and 2 are unionised, while just over three-quarters of those in cluster 3 are unionised. Lastly, the proportion of employees who work for a registered firm for income tax or VAT is low across all three clusters but highest for cluster 3 at 41 percent.

The distributions of each clustering variable over the three clusters seem to be relatively stable over the five waves, with no noticeable change in each cluster's characteristics. Therefore, for the rest of this paper, unless otherwise stated, I will focus the discussion on wave 1.

5.2. Informal, semiformal, and formal employment

The clusters described in the previous section appear to characterise, to various degrees, types of formal and informal employment. Cluster 1, on average, is characterised by low wages, fewer hours worked, few benefits, no unemployment insurance and little trade union representation. Considering that many of these are informal employment features outlined in section 2.2, I, therefore, refer to individuals who fall into cluster 1 as informally employed.

Cluster 3, on the other hand, is characterised by high wages, a 'typical' 40-hour workweek, benefits such as medical aid and pension, unemployment insurance and a higher likelihood of trade union representation. These characteristics resemble those of formal employment. I refer to these individuals as formally employed.

Those in cluster 2 seem to straddle the formally and informally employed. Employees in cluster 2 resemble those formally employed in terms of hours worked and unemployment insurance payments; however, they do not receive the same coverage of benefits or trade union representation as those who are formally employed. To describe this group, I will use the term semiformal employment proposed by Cobb et al. (2009). Cobb *et al.* (2009) argue that informality exists on a spectrum of experiences ranging from

informal to almost entirely formal. The traditional dichotomy of formal/informal employment is inadequate to describe employment that cannot adequately be regarded as strictly formal or informal. This term is appropriate since those in cluster 2 do not fit neatly into either of these two groups but rather somewhere in between.

It is also worth noting that there is some overlap between these segments. For example, there is a small proportion of informal workers who receive a pension. However, it is likely that their employment can more appropriately be classified as informal when one looks at their wages, benefits, or absence of unemployment insurance. This is one potential benefit of using a clustering algorithm, since it classifies workers based on multiple dimensions as opposed to one variable, which when viewed in isolation might misrepresent a worker's true degree of formality.

Interestingly, whether someone contributes to unemployment insurance is the only clustering variable that neatly classifies employees as either informal or semiformal/formal. This supports the suggestion that unemployment insurance is a useful indicator of whether someone is informally or formally employed (see Cichello & Rogan, 2018). Cichello and Rogan note that UIF can be used as a proxy for formal sector employment in South Africa, since all workers are required by law to contribute towards unemployment insurance. It is also not surprising that not all formal employees contribute to UIF since some workers are exempt from contributing, such as public servants (Cichello & Rogan, 2018). It will be shown later that public servants are formally employed.

In summary, employed individuals in this analysis fall into three distinct segments along a spectrum that differ in degrees of formality, ranging from informal, semiformal, and formal employment. The next section examines the characteristics of these groups of workers.

5.3. Labour market, individual and household characteristics

Table 5: Distribution of employment group across clusters

Wave 1	**Employment Type**					**Total (%)**
	Wage	Casual	Self	Subsistence agriculture	Unclassified	
Informal	59	16	15	9	1	100
Semi-formal	67	12	14	6	1	100
Formal	81	5	11	3	0	100
Total	68,46	11,6	13,34	6,12	0,48	100

Source: Author's calculations using NIDS waves 1 to 5.

Table 5 tells us that wage employment is the most common employment type across the job clusters. Casual employment is the second most common employment type for informal workers, whereas self-employment is the second most common type of employment for both semi-formal and formal workers. The least common employment type is agricultural employment for all three worker segments.

Formal workers are disproportionately more likely than the informally, and semi-formally employed to be wage employed.

Semi-formal and informal workers are disproportionately more likely to be casually and self-employed. Informal workers are disproportionately more likely to be employed in subsistence agriculture.

Table 6 indicates that formal workers make up the largest proportion of those who are wage-employed, while informal workers make up most of those who are casually and self-employed and engaged in subsistence agriculture.

Table 6: Distribution occupations within clusters – Wave 1 (2008)

	Informal	**Semiformal**	**Formal**	**Total**
Armed forces occupations	0	0	0	0
Managers	1	5	9	5
Professionals	8	7	24	13
Technicians and associate profession	2	6	8	6
Clerical support workers	4	6	8	6
Service and sales workers	25	25	17	22
Skilled agriculture	2	1	0	1
Craft and related trades workers	17	16	12	15
Plant and machine operators	8	11	11	10
Elementary occupations	34	23	10	22
Total	100	100	100	100

Source: Author's calculations using NIDS waves 1 to 5.

The most common occupations for informal and semi-formal workers were elementary occupations, service and sales workers and craft and related trades workers, while the most common for formal workers were professionals, service and sales workers and craft and related trades workers.

Table 7: Distribution of sectors across clusters (2008)

	Informal	Semi-formal	Formal	Total
Private households	22	8	1	10
Agriculture	14	10	1	8
Mining and Quarrying	1	4	11	6
Manufacturing	11	19	16	15
Electricity, gas, and water supply	1	0	1	1
Construction	8	6	2	5
Wholesale and Retail Trade	15	22	8	15
Transport, storage, and communication	4	3	5	4
Financial intermediation, insurance	7	14	9	10
Community, social and personal services	17	15	45	27
Total	100	100	100	100

Source: Author's calculations using NIDS wave 1.

The most common industries for informal workers were private households, wholesale and retail trade and agriculture. The most common for semi-formal workers were wholesale and retail trade, manufacturing, and community, social and personal services. The most common industries for formal workers were community, social and personal services, wholesale and retail trade, and manufacturing.

Formal workers are more likely to work in mining and quarrying; manufacturing; transport, storage and communication, and communication, social and personal services.

Semi-formal workers are more likely to be in manufacturing, construction, wholesale and retail trade, and financial intermediation.

Informal workers are more likely to work in private households, agriculture, wholesale and retail trade, and construction.

The largest industries in wave 1 (not shown), in order, are community, social and personal services which employ mostly formal workers; private households which

employ mostly informal workers, and wholesale and retail trade, which employs mostly semi-formal workers.

Table 8: Individual Characteristics Wave 1 (2008)

	Informal	Semi-formal	Formal	Overall
Population group				
Black	0,79	0,75	0,70	0,75
Coloured	0,09	0,11	0,11	0,10
Asian/Indian	0,03	0,03	0,03	0,03
White	0,09	0,11	0,16	0,12
Female	0,47	0,43	0,39	0,43
Age	37	37	38	37,37
Married	0,42	0,43	0,51	0,45
Education categories				
No schooling	0,05	0,03	0,02	0,03
Less than matric	0,56	0,55	0,42	0,51
Matric	0,18	0,21	0,21	0,20
More than matric	0,21	0,21	0,35	0,26
Geographic area				
Traditional	0,26	0,20	0,16	0,21
Urban	0,67	0,74	0,78	0,73
Farms	0,07	0,06	0,06	0,06
Total no. employed (N)	2 362	1 954	1 435	5 751

Source: Author's calculations using NIDS wave 1.

Table 8 presents the proportion of employees in various demographic by segment. It can be observed while black employees make up 75 percent of all employees; they make up 79 percent of informally employed individuals. They are, therefore, disproportionately more likely to be informally employed.

Coloured and Asian/Indian employees are proportionally distributed across clusters. White employees make up 16 percent of formally employed individuals while making up 12 percent of overall employees. They are also disproportionately less likely to be informally, or semi-formally employed.

There also appears to be a gendered nature to informality, where women are more likely than men to work in informal employment and less likely to work in formal employment.

In terms of education, those with no schooling are disproportionately more likely to work in informal employment. Those with some schooling, but less than a matric, are more

likely to be informally and semi-formally employed, while those with a matric are more likely to be semi-formally and formally employed. Those with more than matric are much more likely to be formally employed. This suggests that average levels of education are lower for those informally employed and increase with work formality. This is in line with research on the informal sector (Stats SA, 2015).

The degree of formality does also appear to be related to geographic location. Those living in traditional areas and farms are more likely to be informally employed, while those who are semi-formally and formally employed appear to be more likely to live in urban areas.

Table 9: Household Characteristics Wave 1 (2008)

	Informal	Semi-formal	Formal	Overall
Household size	2,85	2,77	2,64	2,75
Num. of children	0,80	0,77	0,66	0,74
Num. of youth	0,83	0,80	0,70	0,77
Num. of adults	1,17	1,16	1,24	1,19
Num. of elderly	0,08	0,07	0,07	0,07
Median household income	5 024	5 741	9 110	6 481
Median household income per capita	2 000	2 500	3 883	2 720

Source: Author's calculations using NIDS wave 1.

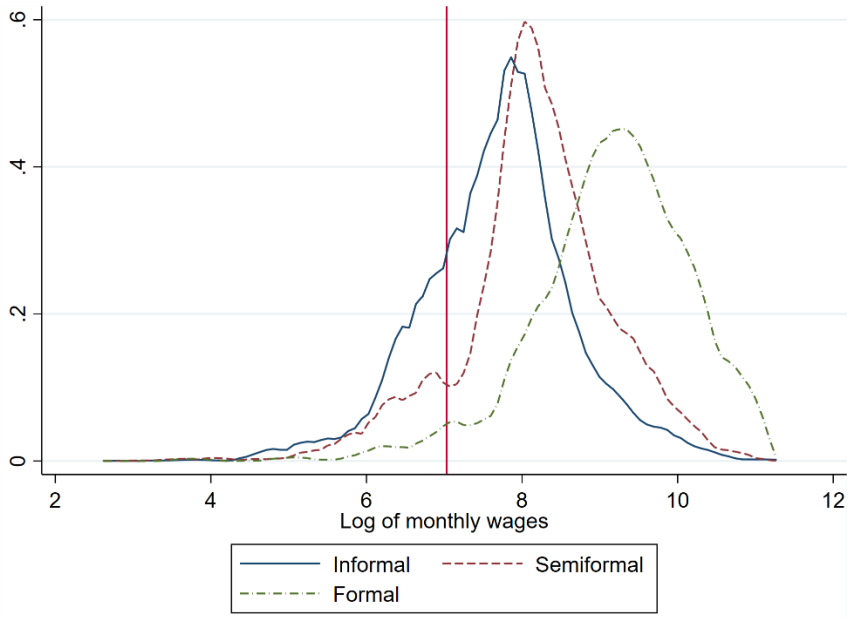
The table above displays details on household size and structure and household income. Although none of the differences between groups is statistically different (not shown), we can observe that mean household size appears to somewhat decline with the degree of formality – larger for those informally employed and smaller for those who are formally employed.

The patterns of household income seem to follow those of labour income reported in table 9 above, in that household income for those who are informally employed is lowest of the three clusters, household income is the highest for those who are formally employed and those who fall into the semi-formal group fall in between. The same pattern holds for household income per capita. Interestingly, median household income for those who are semi-formally and formally employed is 1.1 and 1.8 times higher than those who are informally employed, respectively. These figures are much lower than the ratios of semi-formal and formal individual labour income to informal labour income provided above, which was 1.6 and 4.6 times, respectively. Additionally, the ratio of labour market income for those formally employed to labour income for those who are semi-formally

employed is 2.7; however, this ratio drops to 1.6 times when comparing household income. In the next section, I investigate the relationship between the degree of formality and all income an individual receives.

5.4. Income differences between informal, semiformal, and formal employees

Figure 1: Density of log of monthly wages in 2017 by degree of formality



Notes: Includes imputed observations for missing observations and outliers. The vertical red line corresponds to the upper-bound poverty line in March 2017 of approximately R 1130 per month (Stats SA, 2018).

Source: Authors' calculations using NIDS wave 5.

Figure 1 displays the density of the log of monthly wages by the degree of formality for 2017. The density plots show a positive relationship between the degree of formality and earnings, whereby informally employed individuals earn lower wages and formally employed individuals earn higher wages, with the distribution of formal wages much further to the right than the densities of informal and semiformal wages. This is in line with general findings in South Africa regarding formal and informal wage differentials. For example, Fedderke (2012) finds that formal wages are approximately 2 to 4 times higher than informal wages. Although not directly comparable, given the introduction of the semiformal segment, formally employed individuals in this analysis earn, on average, between 4 to 4.5 times the wages of informal workers. Semiformal workers earn on average, 1.5 times the wages of informal workers.

The red line corresponds to the upper-bound poverty line defined by Stats SA (2018) at approximately R1 130 per month. In can be observed that the highest proportion of the employed whose annual wages are less than the poverty line are those who are informally employed. Approximately 27 percent of informal workers can be classified as the working poor, as defined by the International Labour Organisation (ILO) (2019). The proportion of semiformal workers classified as working poor is about half of that for informal workers, at about 13 percent. A comparatively small fraction, just over 4 percent, of the formally employed, can be regarded as working poor. This suggests that earning a labour income, especially for those who are semiformally or informally employed, and in particular for informal employment, is no guarantee of lifting one out of poverty (ILO, 2019).

We can also compare labour market income to non-labour market income by breaking down total income into five groups: labour market income, government grant income, other income from government, investment income, and remittance income. I exclude subsistence agriculture since it is only reported at an individual level in Wave 2 (Brophy *et al.*, 2018).

Table 10: Summary statistics of different income sources by cluster (2008)

		Informal	Semi-formal	Formal	Overall
Labour market income	Mean	2 603	4 500	11 779	6 244
	Median	1 585	2 536	7 266	3 012
	Std. dev.	3 234	6 237	12 700	9 210
	N	2 031	1 775	1 361	5 197
Investment income	Mean	3 826	4 309	6 654	5 743
	Median	872	1 214	2 219	1 902
	Std. dev.	4 778	5 898	14 888	12 424
	N	35	37	84	157
Remittance income	Mean	872	1 799	1 854	1 418
	Median	634	634	1 585	793
	Std. dev.	873	4 173	1 121	2 630
	N	148	110	48	307
Other government income	Mean	2 696	3 967	736	2 257
	Median	1 332	238	317	317
	Std. dev.	2 475	9 477	983	6 210
	N	11	25	24	62
Grant income	Mean	756	643	994	755
	Median	634	333	1 015	634
	Std. dev.	548	510	601	555
	N	645	404	131	1 183
Total income	Mean	**2 806**	**4 727**	**12 325**	**6 507**
	Median	1 902	2 600	7 404	3 052
	Std. dev.	3 291	6 618	14 702	10 212
	N	2 182	1 865	1 404	5 481

Note: Values are in March 2017 prices and are weighted.

Source: Authors' calculations using NIDS waves 1.

The above table provides summary statistics for different income sources by the degree of formality. For labour market income, investment income and remittance income, there is a positive relationship between the degree of formality and both the mean and median income. Semiformal employees, on average, have the highest government income, while grant income is higher for informal workers compared to semiformal workers. Interestingly, and somewhat counterintuitively, the average grant income for formal workers is higher than that of informal workers, although the sample is small.

In the table below, I include only the 24 percent of individuals in wave one who reported earning at least two sources of income to understand how different income sources contribute to total earnings for employees by the degree of formality.

Table 11: Average percent contribution to total income by cluster for individuals with two or more income sources (2008)

Income Source	Informal	Semiformal	Formal	Overall
Labour market income	0,56	0,63	0,72	0,63
Investment income	0,02	0,04	0,08	0,04
Remittance income	0,06	0,06	0,07	0,06
Other government income	0,01	0,02	0,02	0,01
Grant income	0,35	0,25	0,11	0,25

Note: Values are in March 2017 prices and are weighted.

Source: Authors' calculations using NIDS wave 1.

On average, labour market income makes up the largest share of total income for all three groups. This share increases as the degree of formality increases, while the non-labour market income share decreases commensurately. Non-labour market income makes up 41 percent of monthly income for informally employed individuals, 37 percent for semiformally employed individuals and 28 percent for formally employed individuals. The majority of non-labour market income is made up of grant income and suggests that social assistance plays an important role in total income, in line with Leibbrandt *et al.* (2010). Furthermore, when we include all income, the working poverty rates drop for the informally, semiformally and formally employed considerably to 16, 9 and 3 percent, respectively. This highlights the importance that grant income plays in reducing poverty. This is in good agreement with other findings that grant income plays an important role in reducing poverty and inequality in South Africa (Leibbrandt *et al.*, 2018)

5.5. Transition between labour market states

Previous sections established that there do exist observable racial, gendered and skills differences between informal, semiformal, and formal workers and that labour market earnings are highest for formal workers and lowest for informal workers. According to labour market segmentation theory, this suggests informal workers might face barriers to entry into semiformal or formal employment, or semiformal workers might face rigidities into formal employment.

To investigate possible rigidities, I exploit the panel component of NIDS to describe the rates at with individuals within the working-age population enter, exit, and move between 5 mutually exclusive labour market states. I start by first looking at individuals' distribution across these states for each wave in table 12 below.

Table 12: Proportion of individuals in each labour market state by wave

Labour Market State	Wave 1	Wave 2	Wave 3	Wave 4	Wave 5
Not economically active	34,4	44,7	37,8	36,6	37,2
Unemployed	20,2	15,1	18,5	13,8	13,4
Informally employed	16,3	13,8	14,7	16,0	16,1
Semi-formally employed	14,8	14,3	15,8	20,0	19,8
Formally employed	14,3	12,2	13,2	13,6	13,7
N	13 823	15 801	19 348	22 079	23 147

Note: proportions are weighted using post-stratified weights. N indicates the total number of non-missing observations for labour market status – this figure is unweighted.
Source: Author's calculations using NIDS, Waves 1 to 5.

It can be observed that the proportion of informally and formally employed individuals remained somewhat constant over the period between 2008 and 2017, hovered around 16 and 14 percent of the working-age population, respectively. The proportion of semiformal workers increased from 15 to 20 percent. In 2008 the most common semiformal occupation was service workers, and the most common industry was retail (discussed in Section 5.3.). Considering that employment growth during this period was skewed towards the services sector (Tregenna, 2008), with employment flows out of manufacturing and into services (Cichello, 2013), an increase in the proportion of semiformal jobs could be attributed to the increase in service occupations driven by growth in the retail industry.

Table 13 allows us to look at how individuals transitioned between states over time. The first column indicates an individual's state in the current wave, while the first row indicates a person's state in the next period. For example, 64 percent of individuals who are not economically active in one period are likely to be economically inactive in the next. There is also a significant degree of persistence, which is in line with findings by Ranchhod and Dinkelman (2007). Individuals are most likely to remain in the state they are currently in, as opposed to moving into another state. Unemployed people, however, are more likely to transition into economic inactivity and leave the labour market.

Table 13: Transition rates (%) across labour market states

		Wave t + 1				
		NEA	**Unemployed**	**Informal**	**Semiformal**	**Formal**
Wave t	**NEA**	**64**	18	10	7	2
	Unemployed	36	**29**	17	15	4
	Informal	24	13	**33**	23	8
	Semi-formal	17	11	21	**39**	13
	Formal	10	5	12	15	**58**

Note: Unemployed is refers to broad unemployment, which includes the discouraged unemployed in addition to those who are actively searching for jobs.

Source: Own calculations using NIDS, Waves 1 to 5.

About one in every three informally employed individuals are likely to remain informally employed in the next period. Semi-formally employed individuals are more likely than informally employed individuals to retain employment in their current state, while formally employed individuals are most likely to stay formally employed. Furthermore, informal employees are the most likely to become unemployed in the next period, while formal employees are least like, with a 5 percent chance of transitioning into unemployment.

The informal employed have the highest probability of leaving employment with a combing probability of unemployment or becoming not economically at 37 percent. However, those employed semiformally fair somewhat better than the informally employed still face higher probabilities of leaving employment at 28 percent. The formally employed are likely to remain employed in some employment type with a probability of 85 percent.

Additionally, those in informal employment in the current period and who move to another type of employment in the next period are almost three times more likely to enter semiformal employment than formal employment (23 as opposed to 8 percent). In contrast, those in semiformal employment are 1.6 times more likely to enter informal employment than formal employment. Therefore, individuals are more likely to enter formal employment if they were semiformally employed in the previous period than if they were informally employed. This suggests that it is easier to enter formal employment from semiformal employment than informal employment.

Furthermore, formal employees have the greatest job stability with an 85 percent chance of remaining in some time of employment, followed by those in semiformal jobs at 73 percent, and finally, informal employees have the least stable employment with a 64 percent chance of remaining employed. Additionally, an informal worker is almost three times more likely to enter semiformal employment than formal employment, and a semiformal worker is 1.6 times more likely to enter informal as opposed to formal employment. This suggests that much of the movement observed between informal and formal employment in traditional analysis, which only includes these formal and informal employment (for example, see Cichello et al., 2013), is, in fact, between informal and semiformal employment, as opposed to informal and formal employment.

The instability of informal and semiformal jobs, and the relative stability of formal sector jobs, in addition to their corresponding earnings seems to correspond to the schema of social classification suggested by Schotte *et al*. (2017). Schotte *et al*. define five classes in which South Africans can belong: the chronic poor, transitory poor, the vulnerable, the middle class and the elite. Although not the focus of this paper, further research into the relationship between degree of formality and social class would be interesting.

6. Discussion and conclusion

This paper aimed to identify distinct segments within the South African labour market. A clustering technique was employed, which grouped employed individuals into one of three groups based on a set of characteristics that are typically used when defining or analysing the informal sector and informal employment. This technique allows for a multidimensional approach to defining informality and is not restricted by dichotomies that are often imposed on data due to using only one variable, say tax status, to define informality. This technique also identified a third group of semiformal employees, which allows for a more nuanced analysis of the labour market.

There are, however, limitations to this approach. Firstly, it is difficult to assess the accuracy of the clustering. There are techniques one can perform however these techniques are computationally prohibitive. The clustering results in this paper should therefore be treated with caution. Future research implementing this clustering technique on additional labour force data to ascertain if there is indeed an additional sector that resembles semiformal employed, such as Stats SA's Quarterly Labour Force

Survey, would be valuable. Secondly, the algorithm chosen for clustering may not have been appropriate for our choice of cluster variables, most of which were binary. Although the large sample size might reduce the impact of this, it would be valuable in future to implement several different clustering algorithms to assess if the clusters identified change significantly. Despite these limitations, the descriptive analysis of each cluster seems to be in line with much of the literature in South Africa, which suggests that the clustering my have performed reasonably well.

The evidence presented in this paper suggests that there is quite a bit of movement between informal and semiformal employment, while there is much less movement from either of these states into formal employment. This suggests significant frictions between informal or semiformal employment into formal employment. This is an interesting finding. In his paper on labour market rigidities, Fedderke (2012) points out that, given evidence of a high degree of churning in the labour market, segmentation may be less of a binding constraint on finding employment. This statement is based on the observation that there is a high degree of churning in the labour market between the informal and formal sectors. Future research on labour market rigidities in South Africa would benefit from looking beyond binaries. Although this is analytically convenient and sometimes necessary given data limitations, it has been shown here that imposing dichotomies onto data could potentially mask important information which might be valuable from a policy perspective.

A proposed policy response to reduce the wage differential between segments, when initial segregation is a causal factor of this segmentation, is the establishment of a wage floor set above the prevailing rate in the crowded sector. This wage floor must extend to the crowded sector (Rubery, 2003). This implies that in South Africa, a wage floor should be established in all labour markets above the prevailing rate in the semiformal labour market. Although informal and semiformal labour markets' informality makes it potentially difficult to enforce a wage floor, Dinkelman and Ranchhod (2012) find evidence to suggest that labour legislation in developing countries could potentially impact the informal sector, at least in the short run. An alternative mechanism for establishing a wage floor could be implementing a basic income grant. A basic income grant could increase the productivity of informal and semiformal workers (Samson *et al*.,

2002) and increase their reservation wages if the balance of bargaining power shifts in their favour (Wright, 2002).

Overall, it is suggested that policy responses aimed at reducing segmentation should be focussing on lowering the rigidities of wages that contribute to segmentation and income inequality (Fedderke, 2012). Additionally, given that the inequality reducing the role of social grants is plateauing (Maboshe & Woolard, 2018) and that the labour market is the largest contributor to inequality, the labour market needs to be the focus of policy aimed at reducing inequality. Establishing a wage floor covering all three distinct labour markets identified in this paper would increase informal and semiformal workers' reservation wage. When interpreted within the context of Bergmann's (1974) occupational crowding model, wages are low not necessarily due to the productivity of workers but rather historical discrimination, which created the forces that pushed their wages lower than what they would have been in the absence of discrimination, and in the presence of few rigidities in the labour market.

In summary, this paper used applied a clustering algorithm to identify segments within the South African labour market to shed light on the persistence of inequality. The results suggest that there are three distinct labour market segments in South Africa: the informally employed, the semiformally employed and the formally employed. Further analysis suggests that there is limited mobility between the people who are either informally or semiformally employed into formal employment. This, as well as wage differentials between segments, indicates the presence of rigidities that prevent labour market participants from entering formal employment. Given that the racial and gendered composition of the informal and semiformal labour markets is skewed towards historically disadvantaged individuals, this is concerning. Further research aimed at understanding what drives transitions between informal, semiformal and formal employment would be valuable.

Based on the conclusions from Bergmann's (1974) occupational crowding model, policies aimed at reducing inequality should aim at increasing the wage floor for individuals whose wages were pushed down due to historical discrimination. This also helps explain why wage inequality has persisted in South Africa. The initial discrimination that creates the wage differential, as discussed in the beginning of the paper (Rubery, 2003), is perpetuated by the structure of the labour market long after the initial acts of

discrimination have ended. This supports findings by other researchers (Leibbrandt *et al.*, 2010) that meaningfully reducing income inequality requires a restructuring of the South African labour market.

7. References

Arora, P. & Varshney, S., 2016. Analysis of k-means and k-medoids algorithm for big data. *Procedia Computer Science*, *78*:507-512. DOI:10.1016/j.procs.2016.02.095

Banerjee, A., Galiani, S., Levinsohn, J., McLaren, Z. & Woolard, I. 2008. Why has unemployment risen in the New South Africa? 1. *The Economics of Transition.* 16(4):715-740. DOI:10.1111/j.1468-0351.2008.00340.x Available: https://search.datacite.org/works/10.1111/j.1468-0351.2008.00340.x.

Barrera-Gómez, J & Basagaña, X. 2014. *Multiple imputation in cluster analysis: Package 'miclust'* [R package]. Version 1.2.5. Available: https://www.isglobal.org/documents/10179/5641200/miclust-manual.pdf/a5cca51f-077b-4d19-8abf-408265a63f65 [Dec 4, 2020].

Basagaña, X., Barrera-Gómez, J., Benet, M., Antó, J.M. & Garcia-Aymerich, J. 2013. A Framework for Multiple Imputation in Cluster Analysis. *American Journal of Epidemiology.* 177(7):718-725. DOI:10.1093/aje/kws289 Available: https://doi.org/10.1093/aje/kws289 [Dec 10, 2020].

Bergmann, B.R. 1974. Occupational segregation, wages and profits when employers discriminate by race or sex. *Eastern Economic Journal*. *1*(2):103-110. DOI:

Branson, N. & Wittenberg, M., 2018. Longitudinal and Cross-Sectional Weights in the NIDS Data 1-5. *NIDS Technical Paper no. 9.* Cape Town: SALDRU. Available: http://www.nids.uct.ac.za/publications/technical-papers/230-nids-technical-paper-no9-longitudinal-and-cross-sectional-weights-in-the-nids-data-1-5/file [Dec 4, 2020].

Brophy, T., Branson, N., Daniels, R.C., Leibbrandt, M., Mlatsheni, C. and Woolard, I., 2018. National income dynamics study panel user manual. *Technical Note Release*. Available: http://www.nids.uct.ac.za/images/documents/20180831-NIDS-W5PanelUserManual-V1.0.pdf [Nov 15, 2020].

Cichello, P., Leibbrandt, M. & Woolard, I. 2013. Winners and losers: South African labour-market dynamics between 2008 and 2010. *Development Southern Africa (Sandton, South Africa).* 31(1):65-84. DOI:10.1080/0376835x.2013.853612.

Cichello, P. and Rogan, M., 2017. Informal sector employment and poverty in South Africa: Identifying the contribution of 'informal' sources of income on aggregate poverty measures Working Paper No. 34. *Southern Africa Labour and Development Research Unit (SALDRU), University of Cape Town, Cape Town, South Africa*.

Cobb, C.L., King, M.C. & Rodriguez, L. 2009. Betwixt and Between: The Spectrum of Formality Revealed in the Labor Market Experiences of Mexican Migrant Workers in the United States. *Review of Radical Political Economics.* 41(3):365-371. DOI:10.1177/0486613409336351 Available: https://doi.org/10.1177/0486613409336351 [Mar 7, 2021].

Dinkelman, T. and Ranchhod, V. 2012. Evidence on the impact of minimum wage laws in an informal sector: Domestic workers in South Africa. *Journal of Development Economics*, *99*(1):27-45. DOI: 10.1016/j.jdeveco.2011.12.006

Donders, A.R.T., van der Heijden, Geert J. M. G., Stijnen, T. & Moons, K.G.M. 2006. Review: A gentle introduction to imputation of missing values. *Journal of Clinical Epidemiology.* 59(10):1087-1091. DOI:10.1016/j.jclinepi.2006.01.014 Available: http://www.sciencedirect.com/science/article/pii/S0895435606001971 [Dec 10, 2020].

Essop, H. and Yu, D., 2008. The South African informal sector (1997–2006). *The Regulatory Environment and its Impact on the Nature and Level of Economic Growth and Development in South Africa*, pp.1-58.

Fedderke, J. 2012. The Cost of Rigidity: the case of the South African labor market. *Comparative Economic Studies.* 54(4):809-842.

Fields, G.S. 2011. Labor market analysis for developing countries. *Labour Economics.* 18(1):S16-S22. DOI:10.1016/j.labeco.2011.09.005 Available: http://dx.doi.org/10.1016/j.labeco.2011.09.005.

Finch, H. 2005. Comparison of distance measures in cluster analysis with dichotomous data. *Journal of Data Science.* 3(1):85-100.

Fourie, F., 2018a. Chapter 1: Analysing the informal sector in South Africa: Knowledge and policy gaps, conceptual and data challenges. In *The South African informal sector: Creating jobs, reducing poverty*. F. Fourie & C. Skinner, Eds. Cape Town: HSRC Press.

Fourie, F., 2018b. Chapter 5: Informal-sector employment in South Africa: An enterprise analysis using the SESE survey. In *The South African informal sector: Creating jobs, reducing poverty*. F. Fourie & C. Skinner, Eds. Cape Town: HSRC Press.

Hartigan, J.A. & Wong, M.A. 1979. Algorithm AS 136: A K-Means Clustering Algorithm . *Journal of the Royal Statistical Society.* 28(1):100-108. DOI:10.2307/2346830 Available: https://www.jstor.org/stable/2346830 [10 December 2020].

Heintz, J. & Posel, D. 2008. Revisiting Informal Employment and Segmentation in the South African Labour Market. *South African Journal of Economics.* 76(1):26-44. DOI:https://doi.org/10.1111/j.1813-6982.2008.00153.x Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1813-6982.2008.00153.x [Mar 8, 2021].

International Labour Organisation. 2019. The working poor or how a job is no guarantee of decent living conditions. *Spotlight on Work Statistics No. 6.* Available: https://ilo.org/wcmsp5/groups/public/---dgreports/---stat/documents/publication/wcms_696387.pdf [Mar 7, 2021]

Kassambara, A. No date. *Determining The Optimal Number Of Clusters: 3 Must Know Methods* Available: https://www.datanovia.com/en/lessons/determining-the-optimal-number-of-clusters-3-must-know-methods/ [Dec 18, 2020].

Leibbrandt, M., Woolard, I., Finn, A. and Argen, J., 2010. Trends in South African income distribution and poverty since the fall of apartheid. *OECD Social, Employment and Migration Working Papers No. 101*. Available: https://opendocs.ids.ac.uk/opendocs/bitstream/handle/20.500.12413/11861/Trends_in_South_African.pdf [Mar 7, 2021]

Leibbrandt, M.V., Ranchhod, V. and Green, P. 2018. *Taking stock of South African income inequality* (No. 2018/184). UNU-WIDER Working Paper. Available: https://doi.org/10.35188/UNU-WIDER/2018/626-5 [Mar 8, 2021]

Makaluza, N. & Burger, R. 2018. Chapter 7: Job-seeker entry into the two-tiered informal sector in South Africa. In *The South African informal sector: Creating jobs, reducing poverty*. F. Fourie & C. Skinner, Eds. Cape Town: HSRC Press.

Maboshe, M. and Woolard, I.D., 2018. *Revisiting the impact of direct taxes and transfers on poverty and inequality in South Africa* (No. 2018/79). UNU-WIDER Working Paper.Available: https://www.wider.unu.edu/sites/default/files/Publications/Working-paper/PDF/wp2018-79.pdf [Mar 7, 2021]

Nackerdien, F. & Yu, D. 2019. A panel data analysis of the formal-informal sector labour market linkages in South Africa. *Development Southern Africa (Sandton, South Africa).* 36(3):329-350. DOI:10.1080/0376835X.2018.1487830 Available: http://www.tandfonline.com/doi/abs/10.1080/0376835X.2018.1487830.

Pellicer, M. & Ranchhod, V., 2020. Estimating the effect of racial classification on labour market outcomes: A case study from Apartheid South Africa. *SALDRU Working Paper Series 251*. Version 1. Available: http://www.caps.uct.ac.za/bitstream/handle/11090/975/2020_259_Saldruwp.pdf?sequence=1 [Mar 8, 2021]

Ranchhod, V. & Dinkelman, T. 2007. *Labour Market Transitions in South Africa: What can we learn from matched Labour Force Survey data?* Available: http://econpapers.repec.org/paper/ldrwpaper/14.htm.

Reich, M., Gordon, D.M. & Edwards, R.C. 1973. A Theory of Labor Market Segmentation. *The American Economic Review.* 63(2):359-365. Available: https://www.jstor.org/stable/1817097 [Mar 7, 2021].

Rogan, M. & Skinner, C., 2018. Chapter 4: The size and structure of the South African informal sector 2008-2014: A labour-force analysis. In *The South African informal sector: Creating jobs, reducing poverty*. F. Fourie & C. Skinner, Eds. Cape Town: HSRC Press.

Rubery, J., 2003. *Pay equity, minimum wage, and equality at work*. Geneva: International Labour Organisation. Available: https://www.ilo.org/wcmsp5/groups/public/---ed_norm/---declaration/documents/publication/wcms_decl_wp_20_en.pdf [Mar 6, 2021]

Samson, M., Babson, M.O., Haarmann, C., Haarmann, D., Khathi, M.G., Mac Quene, K. and van Niekerk, I., 2002. Social security reform and the basic income grant for South Africa. *Report commissioned by the International Labour Organization (ILO) and produced by the Economic Policy Research Institute (EPRI)*. Available: https://basicincome.org/bien/pdf/rp31.pdf [Mar 8, 2021]

Schotte, S., Zizzamia, R. and Leibbrandt, M., 2017. *Social stratification, life chances and vulnerability to poverty in South Africa.* Available: http://opensaldru.uct.ac.za/bitstream/handle/11090/883/2017_208_Saldruwp.pdf?sequence=1 [Mar 7, 2021]

Southern Africa Labour and Development Research Unit (SALDRU). National Income Dynamics Study (NIDS) Wave 1, 2008 [dataset]. Version 7.0.0. Pretoria: SA Presidency [funding agency]. Cape Town: SALDRU [implementer], 2018. Cape Town: DataFirst [distributor], 2018. https://doi.org/10.25828/e7w9-m033

SALDRU. NIDS Wave 2, 2010-2011 [dataset]. Version 4.0.0. Pretoria: SA Presidency [funding agency]. Cape Town: SALDRU [implementer], 2018. Cape Town: DataFirst [distributor], 2018. https://doi.org/10.25828/j1h1-5m16

SALDRU. NIDS Wave 3, 2012 [dataset]. Version 3.0.0. Pretoria: SA Presidency [funding agency]. Cape Town: SALDRU [implementer], 2018. Cape Town: DataFirst [distributor], 2018. https://doi.org/10.25828/7pgq-q106

SALDRU. NIDS 2014-2015, Wave 4 [dataset]. Version 2.0.0. Pretoria: Department of Planning, Monitoring, and Evaluation [funding agency]. Cape Town: SALDRU [implementer], 2018. Cape Town: DataFirst [distributor], 2018. https://doi.org/10.25828/f4ws-8a78

SALDRU. NIDS 2017, Wave 5 [dataset]. Version 1.0.0 Pretoria: Department of Planning, Monitoring, and Evaluation [funding agency]. Cape Town SALDRU [implementer], 2018. Cape Town: DataFirst [distributor], 2018. https://doi.org/10.25828/fw3h-v708

Stats SA. 2015. *Labour statistics: Labour market dynamics in South Africa, 2015.* Available: https://www.statssa.gov.za/publications/Report-02-11-02/Report-02-11-022015.pdf [Aug 2, 2020]

Stats SA. 2018. *National Poverty Lines.* Available: http://www.statssa.gov.za/publications/P03101/P031012018.pdf [Mar 7, 2021]

Tregenna, F. 2008. The Contributions of Manufacturing and Services to Employment Creation and Growth in South Africa. *South African Journal of Economics.* 76(s2):S175-S204. DOI:https://doi.org/10.1111/j.1813-6982.2008.00187.x Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1813-6982.2008.00187.x [Mar 8, 2021].

VanderPlas, J. 2016. Chapter 5: Machine Learning. In *Python Data Science Handbook.* California, United States of America: O'Reilly. 331-515.

Wittenberg, M. 2014. Analysis of employment, real wage, and productivity trends in South Africa since 1994. *Conditions of Work and Employment Series No. 45.* Geneva: International Labour Office. Available: http://wcmstraining2.ilo.org/wcmsp5/groups/public/---ed_protect/---protrav/---travail/documents/publication/wcms_237808.pdf [Nov 14, 2020]

Wittenberg, M. 2017. Wages and wage inequality in South Africa 1994–2011: Part 1–Wage measurement and trends. *South African Journal of Economics*, *85*(2): 279-297. DOI: 10.1111/saje.12148

Wright, E.O., 2002. The shadow of exploitation in Weber's class analysis. *American Sociological Review*: 832-853. DOI: 10.2307/3088972

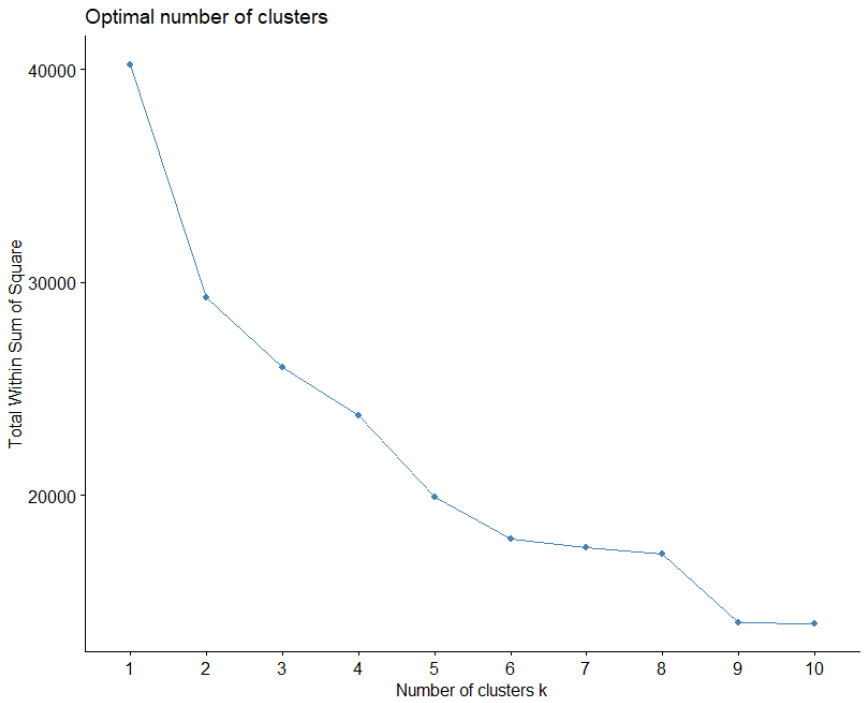
8. Appendix

8.1. Choice of optimal number of clusters

Choosing the optimal number of clusters is a non-trivial exercise since the results of clustering varies as the number of clusters parameter changes – which is unknown and must be determined *a priori* by the researcher. This, however, can be overcome by using several methods direct methods and statistical testing methods. In this section, I apply some of these methods to estimate what the optimal number of clusters in the data may be. The methods described here are explained in more depth in Kassambara (No date)

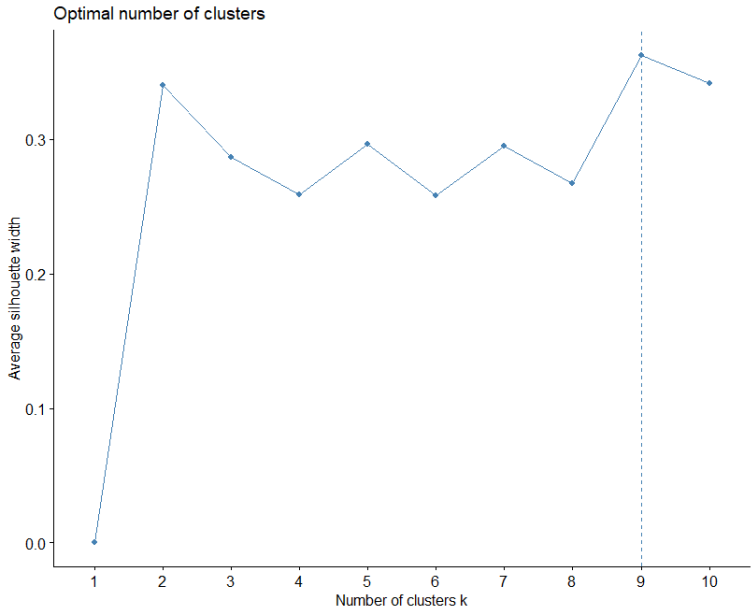
Direct method – Elbow method

This method looks at the Within-Cluster sum of Squares as a function of the number of clusters. The optimal number of clusters is the number whereby increasing it does not improve the total WSS. If it is unambiguous as to what the optimal number of clusters is, there would be a clear bend at this point K*. Unfortunately, in this case there is no clear bend or "elbow" – it could be 5 or 9. In this case clustering beyond 4 or 5 cluster will add little value – however this is ambiguous. Therefore, using this method does not help us identify the optimal number of clusters.



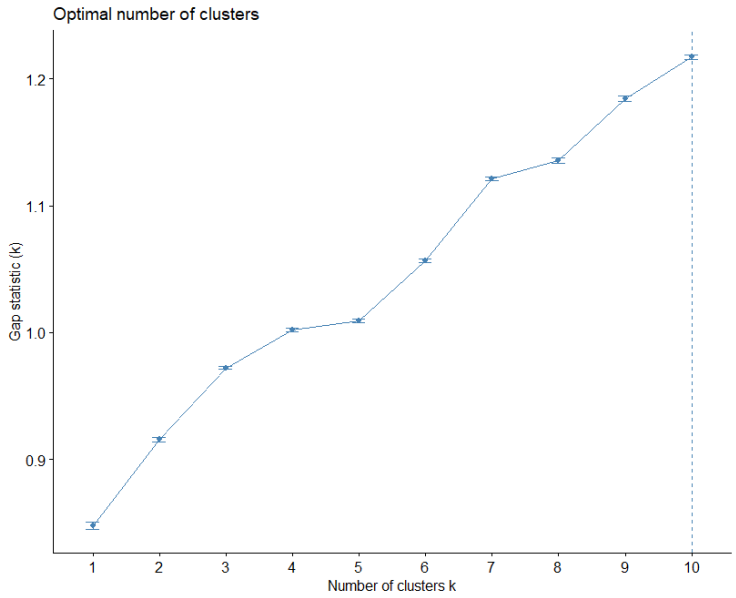
Direct method – Average Silhouette method

This method measures the quality of clustering. A high average silhouette width indicates good clustering. In this case, the graph below indicates that 9 is the appropriate number of clusters. At this point the average silhouette width is maximised.



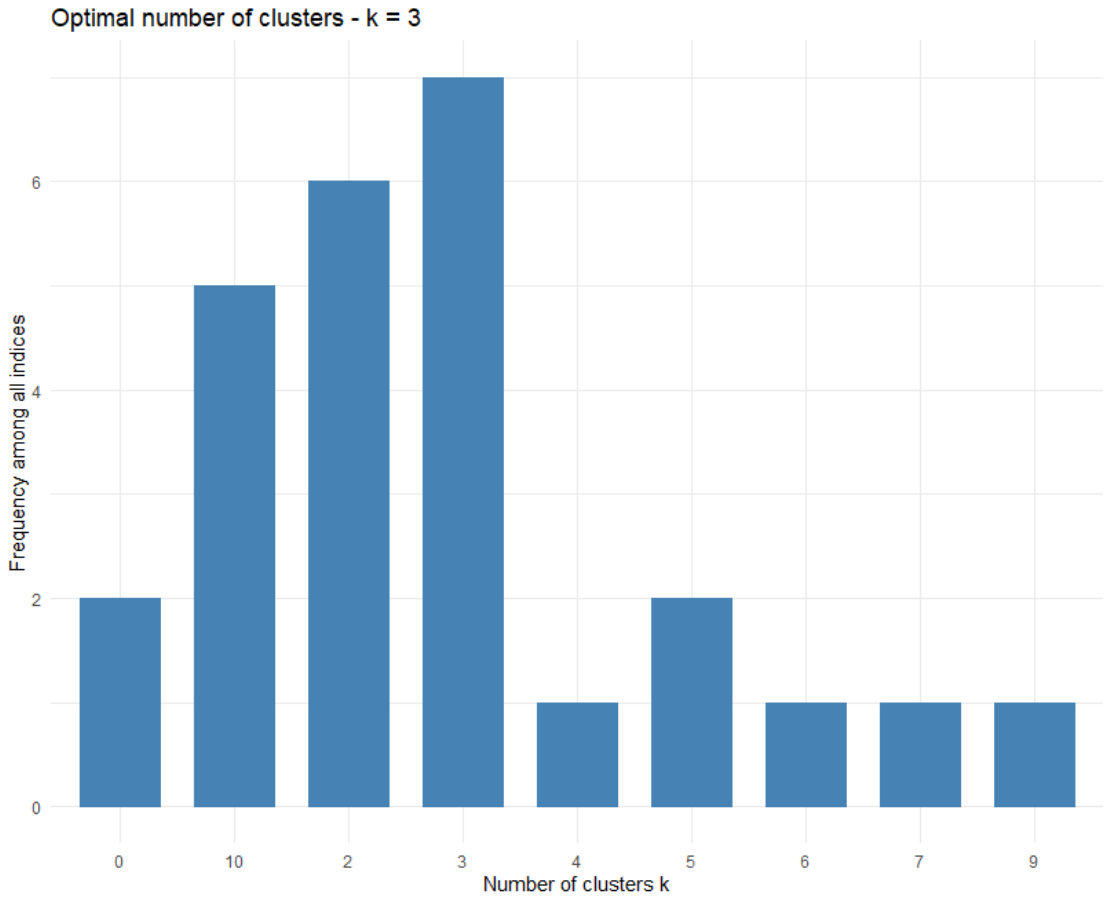
Statistical method – Gap statistic

The gap statistic applies compares the total within intra-cluster variation for different values of k with their expected values under the null reference distribution of the data. The estimate of the optimal clusters is the value that maximises this statistic. The figure below suggests that the optimal number of clusters is 10.



30 indices for choosing the best number of clusters

The previous methods have given different estimates for the appropriate number of clusters. However, in R the *NbClust* package provides 30 indices for determining the optimal number of clusters and suggests to a user the best choice from the different results by varying all combinations of clusters, distance measures and clustering methods. The graph below is the output of this function. It suggests the optimal number of clusters is 3. Out of the 30 indices, 3 was chosen as the optimal number of clusters 7 times, followed by 2 and then 10. I therefore choose 3 as the optimal number of clusters.



8.2. Additional tables

Table A1: Proportion of employed individuals in each cluster (%)

	Wave 1	Wave 2	Wave 3	Wave 4	Wave 5	Average
Cluster 1	41	41	39	38	36	39
Cluster 2	34	36	37	40	39	37
Cluster 3	25	24	24	22	24	24

Note: Data are weighted.
Source: Author's calculations using NIDS waves 1 to 5

Table A2: Distribution of type of employment across clusters

Wave 1	**Employment Type**					
	Wage	Casual	Self	Agri	Unclass.	Total
Informal	31	51	39	55	52	36
Semi-formal	32	35	35	30	42	33
Formal	37	14	25	15	6	31
Total	100	100	100	100	100	100

Source: Author's calculations using NIDS wave 1

Table A3: Overall average percent contribution to total income by cluster

		Informal	**Semi-formal**	**Formal**	**Overall**
Labour market income	Mean	0,77	0,85	0,92	0,85
Investment income	Mean	0,01	0,01	0,02	0,01
Remittance income	Mean	0,02	0,02	0,01	0,02
Other government income	Mean	0,00	0,01	0,00	0,00
Grant income	Mean	0,12	0,07	0,03	0,07

Source: Author's calculations using NIDS wave 1

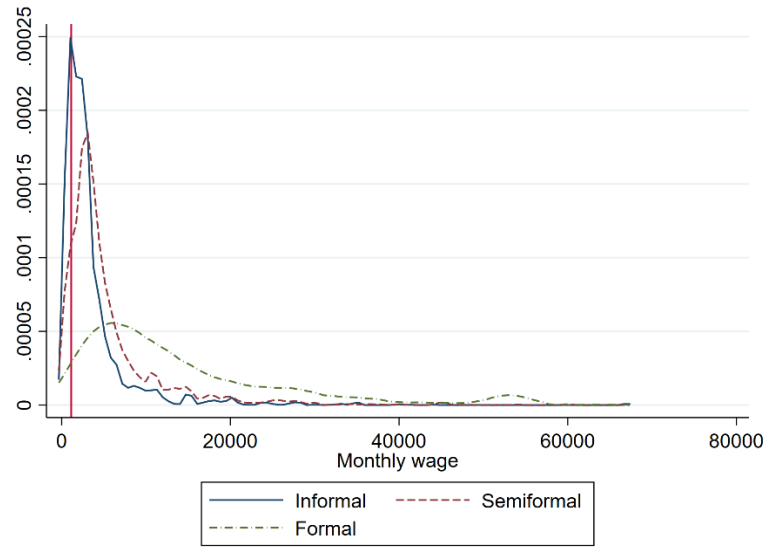


Table A4: Proportion of individuals in each labour market state by wave

Labour Market State	Wave 1	Wave 2	Wave 3	Wave 4	Wave 5
Not economically active	34,4	44,7	37,8	36,6	37,2
Discouraged unemployed	5,5	4,4	3,4	1,7	1,6
Searching unemployed	14,7	10,7	15,1	12,1	11,8
Informally employed	16,3	13,8	14,7	16,0	16,1
Semi-formally employed	14,8	14,3	15,8	20,0	19,8
Formally employed	14,3	12,2	13,2	13,6	13,7
N	13 823	15 801	19 348	22 079	23 147

Note: proportions are weighted using post-stratified weights. N indicates the total number of non-missing observations for labour market status – this figure is unweighted.

Source: Author's calculations using NIDS, Waves 1 to 5.