Handbook on **Inequality Measurement for Country Studies**

Muna Shifa and Vimal Ranchhod

AFD
AGENCE FRANÇAISE
DE DÉVELOPPEMENT

University of Cape Town • iYunivesithi yaseKapa • Universiteit van Kaapstad

ACEIR
African Centre of Excellence
for Inequality Research

Handbook on **Inequality Measurement for Country Studies**

Muna Shifa[1] *and Vimal Ranchhod*[2]

1. INTRODUCTION

Writing a country level study on inequality can be an intimidating task for researchers. How does one begin to conceptualize such a study? What needs to be included? What sorts of analyses are required, and how does one go about generating the required results? This Handbook is written primarily as a reference document to guide researchers who are about to embark on writing a research report that summarizes inequality within a given country-level context.[3]

We have written the Handbook assuming that the researchers have limited experience with such an endeavour, but that they do have some understanding of data management and the ability to work with the statistical package Stata. Since each context is unique, we are also assuming that the relevant research team has a substantial amount of localised contextual and institutional knowledge. Such knowledge is required in at least two different dimensions.

First, researchers need to have some awareness of the general socio-economic environment that their study is located in. This is necessary so that important drivers of inequality are included in the study, while relatively unimportant drivers of inequality may be excluded. For example, in some contexts the differences in earnings between rural and urban populations may explain a large fraction of national level inequality, while in a different country it may account for only a trivial proportion of aggregate inequality. A similar argument can be made for inequality derived from the returns on financial assets. Researchers are asked to draw on their localised expertise in deciding which subset of inequality drivers are relatively more important in their context.

Second, researchers need to be knowledgeable about the various surveys and alternative data sources that can be used for their study. We discuss what types of data are

1 Post-doctoral research fellow in SALDRU at the University of Cape Town.

2 Professor of Economics in SALDRU at the University of Cape Town.

3 While we are writing this specifically as an intermediate guide for the members of the ACEIR project, the document may nevertheless be useful for researchers who are planning a study of this sort in different contexts. This could include both a narrower or broader focus, such as regional or cross-country analyses respectively.

required in more detail in the Data section below, but at a general level there are two potential decisions that require an awareness of what data is available. The first possibility is that there are multiple surveys that could be used to analyse inequality. In this case, researchers will need to make a choice about which data sources to include in their study, based on their knowledge of the full set of available datasets. A second possibility is that there may be desirable pieces of information that are captured in different datasets.[4] For example, educational data may be best obtained from a Department of Education, while fertility data may be captured by the Department of Health. A study on the interaction between educational outcomes and fertility might best be undertaken by being able to merge these two institutional databases. In order to consciously decide whether or not to include such an analysis in a study, one would have to know that the relevant information is captured in the two datasets, that both of them are available to the researchers, and that it is feasible to combine this information in a way that can be useful.

In addition to enabling researchers to undertake an inequality study within a particular country, the Handbook has also been written with a meta-objective in mind; namely to facilitate the comparability of results and findings across countries. This externality represents one of the major motivations for a multi-country collaboration such as ACEIR. In order to maximize the comparability of results, it helps to coordinate our approaches *a priori* in a deliberate manner. By following the methods and interpretation described in the Handbook, researchers will ensure that data is prepared and analysed in similar ways, and that the results can thus be compared across the different country nodes.

Of course, no two countries will ever be fully comparable. As discussed above, history, context, and institutions all matter in idiosyncratic ways, such that any country is always going to be unique once we include ever finer levels of detail in our analyses. This implies that there is likely to be some trade-off between the comparability of the studies across countries on the one hand, with the specificity and completeness of the study for a given country on the other hand. To balance these somewhat competing objectives, we adopt what we think of as a pragmatic approach. Any dimension of inequality that a researcher believes is important for obtaining a full understanding of inequality within a particular country should be included in that country's study. At the same time, we list in Section 5.2 a minimal set of results that each country study should ideally contain. This minimal set of required results, to the extent that they are feasible to implement in each country, will provide the basis for a meta-study that focusses on the comparisons of inequality across the different countries.

The remainder of this Handbook is structured as follows: In Section 2 we discuss some of the parameters that all empirical inequality studies need to decide on a priori. In Section 3 we discuss relevant data requirements and how one might address the common issues that arise. In Section 4 we describe how to implement the various estimators when focusing on income inequality in particular. Section 5 provides a basic structure of how we imagine each country report will be written. We conclude in Section 6 with a brief summary of this Handbook.

4 Note that these two possibilities are not mutually exclusive. Indeed, it is probable that both of these data-related considerations are relevant simultaneously.

2. PARAMETERS OF INEQUALITY STUDIES

The set of research that can be considered to be relating to inequalities is vast. Any subject relating to the structure of society potentially includes some aspect of inequality. Thus, the disciplines that have contributed to a holistic understanding of inequality includes sociology, history, politics, economics, health, literature, statistics, geography, moral philosophy and psychology; and even this exceptionally broad list may not be complete. Even within disciplines, one needs to determine the methods for investigating the subject matter. These can range from purely abstract theory, to large scale quantitative analyses, to small sample qualitative studies. Once the scope of the study is sufficiently narrowed, one still needs to determine what sources of information are going to be utilised. For example, data can be obtained from surveys, administrative databases, company records, historical archives, maps, legal systems of property rights and registers, or even indirectly using price data and accounting systems. Thus, any researcher working on inequality needs to determine the parameters of their study so that answering the research question becomes feasible.

Within the broad class of empirical studies of inequality in economics, we still need to answer at least three questions that define the scope of our study: Inequality of what, amongst whom, and over what time period?

The country studies, almost by construction, mean that the grouping that we are restricting our analysis to will be the people who reside within a country's borders. Note that this conceptually includes immigrants, regardless of their legal status, and excludes people who have emigrated, regardless of whether such a migration is temporary or permanent. Such a choice may seem trivial but could have a significant impact on our measures of inequality. Without going into any detail about the advantages or disadvantages of such a choice, we note simply that this accords with how Censuses are typically conducted. As such, data availability is likely to make such a decision moot from a practical perspective.

In terms of time horizons, one could quite easily motivate for measuring inequality over a long period of time, or in a different political or historical era. In this Handbook, however, we are more interested in analyses that investigate contemporary inequality and relatively recent trends in said inequality. To be precise, this refers to the time period spanned by the most recent available data that is suitable for the analysis that we wish to undertake.

Determining what dimensions of inequality we wish to investigate is another decision that can substantially change the implementation, and thus findings, of an inequality study. This project is concerned with economic inequalities, although even this class of studies is fairly broad. One could ask about the inequalities in market power, in access to credit, or in terms of rental incomes. We could measure labour market discrimination by race or ethnicity, or by gender, or look at occupational sorting and stratification. A decision is required, and for this set of papers we are focussed primarily on contemporaneous inequality in aggregate income or consumption over the entire population of

individuals. Nonetheless, as already stated in the introduction, researchers are expected to expand on this with the inequality decompositions, where sub-groups are categorised based on prior knowledge and local expertise.

The next section introduces the data requirements for the country studies, as well as some common issues that arise and how to address them.

3. DATA AND MEASURING WELLBEING

In the previous section, we have discussed the different aspects of inequality. Our discussion in this section focuses on data and measurement issues related to analyzing economic inequality. Although the number and frequency of household surveys that collect information on income and consumption are increasing recently, data quality is still a pressing issue in most developing countries. We discuss some of the data-related issues that we should be aware of when measuring inequality.

The underlying welfare measures

Income or consumption are the conventional measures used in the literature to measure individual well-being for analyzing economic inequality. In a developing country context, however, consumption data are widely used to estimate both poverty and inequality. One reason for this is that data on income is not readily available. Most developing and emerging countries have a large informal sector, and it is difficult to collect income information from self-employment and subsistence farming. Furthermore, given that households smooth consumption (via saving and borrowing), consumption is preferred as a measure of current welfare. Thus, while income can be considered as a means to achieve well-being, consumption is a more direct measure of individual well-being.

Nonetheless, there are various problems that we face in measuring consumption using household surveys. For example, it is often difficult to impute a monetary value for goods and services that are consumed from own production (e.g. subsistence farming), or that are provided by the public sector (e.g. access to free education and health services). These measurement issues can bias our inequality estimates. Such issues also create difficulty in making inequality comparisons across countries.

Non-response and under reporting

In most household surveys, households at the higher end of the income distribution are underrepresented due to the high rate of survey non-response amongst these households. In addition, richer households also tend to underreport their income levels. These problems may lead to the underestimation of inequality levels. We use weights (post-stratification weights) to correct for problems related to non-response among the rich. Even if we use consumption data, inequality estimates based on consumption data

may still lead to the underestimation of economic inequality, since the rich tend to save more than the poor. In some cases, tax records have also been used to estimate the extent of inequality. However, data on taxable income is generally only available for income earners exceeding a certain threshold level of income (Wittenberg, 2017).

We may also have item-non-response, and income given in brackets in our data. Unless values in a data set are missing completely at random (MCAR), ignoring missing values can lead to a biased estimate of inequality. If missing data is not MCAR, we may use some imputation methods to impute for missing values. Reporting incomes in brackets is also common in household surveys and using this type of data may be the only option for some studies. For example, income values in the South African censuses are reported only in brackets. In this case, most studies use imputation techniques to convert the values reported in brackets into point estimates. Such approaches may still underestimate inequality if every individual in a given bracket is assigned a single income value, as is often the case.

Survey comparability

Consistency over time or across countries/regions is another key challenge that we face in measuring inequality trends. It is possible that changes in our inequality measures could be due to a real change in living standards, or it could be due to methodological changes in how the data were collected, or due to some combination of these two effects. Changes in data collection (i.e. survey design and instruments used) and variable measurement, changes in prices, and seasonality adjustments are part of changes in methodology. For example, the way that we measure income or consumption should be consistent across survey years. Changes in income or consumption categories (e.g. due to an update of the list of consumption items, net income versus gross income), changes in the time period covered[5], and the seasonality of economic activities could lead to an inconsistent measure of income or consumption data over time. We thus need to be mindful of these issues when making inequality comparisons *over time, regions or countries* using survey data.

Equivalent scales

Data on income or consumption are often collected at a household level. The National Income Dynamics Study (NIDS), which is a nationally representative panel survey in South Africa, is one of the few exceptions that collects income data at both an individual and household level. Analysing poverty or inequality requires welfare information at individual levels. Thus, we may need to make certain assumptions about how income or consumption is distributed within households in order to convert household-level income or consumption into individual-level data. Even if we have income information at the individual level, we have to first aggregate this into household level data since families share income and other resources within a household. One approach is to use

5 For example, the reference periods in collecting consumption or income data could be yearly or monthly. This can create comparability problems because monthly income or consumption data may include transitory fluctuations that may not be the case if we use yearly reference period. Thus, we expect our inequality measure to be higher if we use a monthly reference period instead of a yearly reference period in collecting income or consumption data.

a per capita scale, which is to divide total household consumption or income by household size and assign this average value to all individuals in a household. In this case, we are assuming that household income or consumption is equally distributed across each individual in a household, and we are also ignoring economies of scale.[6] An alternative is to use an adult equivalence scale. The adult equivalent approach adjusts for both economies of scale as well as the cost of children (assuming that children consume less than adults).[7] Note that in both cases the intra-household allocation of resources is ignored, since this requires detailed consumption information for each household member. We are thus not able to disaggregate our inequality estimates by groups such as gender, if group members are typically co-resident within the same households.

4. APPROACHES TO MEASURING ECONOMIC INEQUALITY

4.1 Choosing among inequality measures

In the literature, there are various tools (inequality indices) that are used to measure inequality. Our choice of inequality measure depends partly on the type of question that we want to examine. One question that we might want to investigate, for example, is the extent to which inequality in South Africa is driven by an unequal income distribution within race groups, relative to an unequal income distribution across race groups. In order to answer these types of questions, we need to have an inequality measure that allow us to decompose overall inequality into different groups.

One approach to choose among the various inequality measures (indices) is to follow the axiomatic approach (Cowell,1985). Accordingly, we first specify a set of minimum desirable properties that we would like an inequality measure to satisfy. Then, we use these axioms to choose among inequality indices. We discuss below some of the key desirable properties (axioms) that an inequality index should satisfy. The discussion in this section draws largely from the work by Foster et al (2013) entitled *"A unified approach to measuring poverty and inequality: Theory and practice"*.

Axiom 1: Anonymity (symmetry): *this axiom requires that an inequality measure should not change due to permutation; i.e. an individual's identity is not relevant to the analysis of inequality.* Consider a society of four individuals named A, B, C and D, with incomes

6 For example, the per capita cost of living for a single-family household can be higher than that of a two-family household because the two-family household members can share the cost of rent and other common household costs. If we do not adjust for such economies of scale, we tend to underestimate the welfare of larger households.

7 A common formula used to calculate an adult equivalent is: Adult equivalents=(adults+a×children)^q where a is the child parameter which often ranges between 0.5-0.75 and q is the parameter for economies of scale. For example, q= 0.9 in the case of South Africa.

10, 20, 30, and 40 respectively.

$$Y_1 = (10, 20, 30, 40)$$

$$A, B, C, D$$

Consider a second population with the same set of incomes, but with different recipients.

$$Y_2 = (10, 20, 30, 40)$$

$$D, C, B, A$$

The anonymity /symmetry axiom implies that distributions Y_1 and Y_2 are equally unequal.

Axiom 2: Population invariance: this property requires that the level of inequality within a society is invariant to population size. For instance, if we have Y_3= (10, 10, 20, 20, 30, 30, 40, 40) from Y_1 (by doubling the number of individuals without changing the income distribution), then population independence implies that we regard the two distributions as equally unequal.

Axiom 3: Normalization: this property requires that an inequality index should be zero when all incomes are equally distributed.

Axiom 4: Scale invariance: this axiom requires that inequality should not change if all individuals' income increased by the same proportion. For instance, if we multiply everybody's income in Y_1 by two, we get Y_4 = (20, 40, 60, 80). Note that the income level of the richest person is 4 times that of the poorest person in the case of both Y_1 and Y_4. Scale invariance implies that the level of inequality in Y_1 and Y_4 is the same, and that inequality is a purely relative concept. Thus, the size of the income does not matter. Note that the desirability of this property depends on whether we are interested in absolute or relative inequality measures. If we consider absolute gaps, the absolute gap between the richest and poorest individuals in the case of Y_1 is 30 while this gap is 60 in the case of Y_4. Thus, if we choose absolute inequality measures, we can say that the level of inequality is higher in Y_4 compared to Y_1.

Based on the invariance property we can classify inequality measures as either absolute or relative inequality measures. Absolute inequality measures are translation invariant: Adding/subtracting an absolute amount to/from all individuals' income will not change absolute income inequality measures. Relative inequality measures, on the other hand, are scale invariant: Multiplying all incomes by a positive scalar value will not change relative income inequality measures. Relative inequality measures are not translation invariant while absolute inequality measures are not scale invariant.

From an analytical perspective, the scale invariant property is desirable since it ensures that the value of an inequality measure does not change with the units in which income is measured, while translation-invariant measures violate this property. For example, the variance is one of the simplest absolute inequality measures, but its value depends on the unit of measurement. For this reason, relative inequality measures are preferable

to absolute inequality measures. Thus, our discussion in this paper focuses on relative inequality measures.

Axiom 5: Transfer principle: What happens when income is transferred from someone who is relatively rich to someone who is relatively poor, holding their ranks in the income distribution constant? For example, suppose that the richest person in Y_4 transferred ten rands of income to the poorest person, producing a new income distribution Y_5 = (30, 40, 60, 70). The judgement is that such a transfer should reduce inequality, and therefore that the level of inequality in Y_5 is lower than that of Y_4, commands widespread support. There are two versions of the transfer principle; (i) *Weak transfer principle* - inequality should decrease or remain the same after transferring income from a relatively rich individual to someone who is relatively poor, and (ii) *Strong transfer principle* - inequality should strictly decrease after transferring income from a relatively rich individual to someone who is relatively poor.

Axiom 6: Transfer sensitivity: This property requires that an inequality measure be more sensitive to transfers at the lower end of the distribution (i.e. between two poor individuals rather than between two rich individuals). For example, suppose that we have income distribution Y_6 = (30, 30, 60, 80), which is obtained by transferring 10 rands from the second poorest individual to the poorest individual in Y_4 = (20, 40, 60, 80). Compare this to income distribution Y_7 = (20, 40, 70, 70), which is obtained by transferring the same amount from the richest individual in Y_4 = (20, 40, 60, 80) to the next richest individual. Transfer sensitivity implies that our inequality measure should be more sensitive to the transfers that generated Y6 than those that generated Y_7.

Axiom 7: Decomposability: If we want our inequality measure to be broken down into group contributions, then we want our inequality measure to be decomposable. These groups could be income sources (labour vs non-labour) or other dimensions including race, sex, and locations. There are two desirable properties that a decomposable inequality measure should satisfy:

i. Additive decomposability: Overall inequality is the sum of all within-groups and between-groups inequality. Within-group inequality is a weighted sum of sub-group inequalities (the weights could be population shares or relative incomes) while between-group inequality is inequality between groups (mean group income is assigned to every individual within each group).

ii. Sub-group consistency: This concept relates to the responsiveness of the overall inequality measure to changes in the inequality levels of constituent groups. If there is a rise in inequality for a given population sub-group, and inequality does not fall in the rest of the subgroups, then our overall inequality measure should rise.

4.2 Commonly Used Inequality Measures

In this section, we consider some of the most commonly used inequality measures. In general, based on the approach used to derive them, inequality measures can be broadly classified into two categories: descriptive and normative (Sen, 1973). The descriptive inequality measures are usually mathematical or statistical formulas. Thus, the characteristics of such indices are a function of their mathematical or statistical properties respectively. Most inequality measures are descriptive in nature. The normative inequality measures are derived from a social welfare function based on some *a priori* value judgment about the effects of inequality on social welfare. These inequality measures relate an inequality index to social evaluation and specify whether inequality is bad or not, as well as how much welfare a society loses or gains from inequality. The Atkinson class of inequality indices are among the most cited normative inequality measures. It should be noted that the inequality measures that we discuss here do not necessarily satisfy all of the axioms that we discussed in the previous sub-section. For example, the Atkinson index satisfies almost all of the axioms, but it is not additively decomposable. Thus, if our objective is to decompose overall inequality by population sub-groups, then we can use the entropy class of inequality measures instead.

When discussing the various inequality measures, we use data from NIDS (wave 1 and wave 4), the 1998 South African Demographic and Health Surveys (DHS), which are nationally-representative household surveys, and the 2011 South African Census. We use per capita income data as our measure of individual welfare (i.e. total household income divided by household size). The income variables that we use measure income from all sources (i.e. labour income and non-labour incomes (e.g. social grants)). We use sample weights in all of our income inequality estimations.

We use the DHS when discussing approaches to measuring non-income dimensions of inequality, namely asset inequality. We use variables from our data sets only to provide practical examples of how to estimate and interpret the various inequality measures. Therefore, the results from this exercise should not be used for other purposes.

We will be using the Stata software and the DASP package for estimating most of the inequality indices. Instructions on how to install the DASP package are provided in the appendix (see Araar & Duclos, 2013 for more details).

4.2.1 Quantile ratios/ decile ratios/ percentile ratios

The simplest way to examine income inequality is to divide the population into quantiles or deciles after ranking them from the poorest to the richest. This allows us to calculate the levels or proportions of income that accrue to each quantile or decile. Table 1 below shows the percentile share of income by decile for South Africa in 2008. We use the following Stata command to estimate the percentile shares:[8]

8 Please refer to the Stata help menu to get more details with regards to all the Stata commands that we used in this paper. For example, if you type "*help pshare*" in the Stata command window you will get detailed information about the "*pshare*" command. The variable "*pcminc*" indicates the per-capita income measure, while the "*wgt*" variable indicates weights (post-stratification weights) in the NIDS data set.

pshare estimate pcminc [w=wgt],nquantiles(10) percent

Table 1: percentile shares, 2008

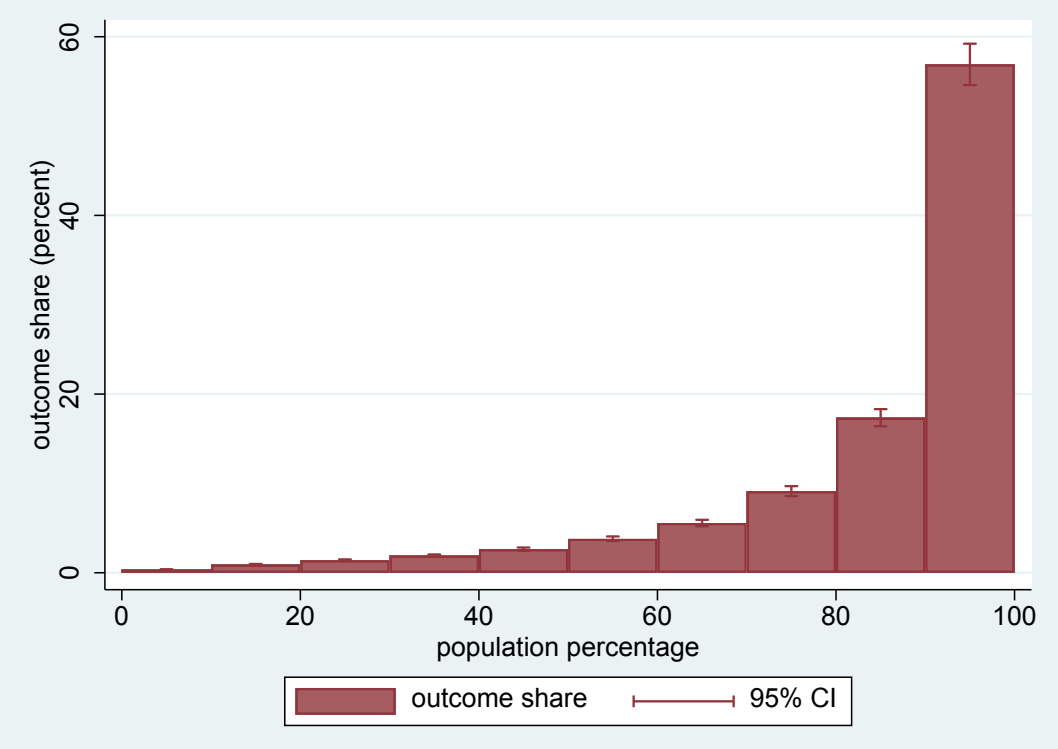
Percentile shares (percent) Number of obs = 17,710

pcminc	Coef.	Std. Err.	[95% Conf.	Interval]
0-10	.3605923	.016485	.3282802	.3929044
10-20	.9175996	.0355134	.84799	.9872093
20-30	1.398974	.0508721	1.299259	1.498688
30-40	1.916194	.0681962	1.782523	2.049866
40-50	2.643722	.0921937	2.463013	2.824431
50-60	3.807283	.1315998	3.549335	4.065232
60-70	5.567746	.1830186	5.209011	5.92648
70-80	9.1327	.2860643	8.571986	9.693415
80-90	17.35443	.4894556	16.39505	18.31381
90-100	56.90076	1.182635	54.58268	59.21884

The results show that the richest 10% of the population received 57% of the total income in 2008 while less than 1% of the total income accrued to the poorest 10% of the population. We can also present the estimates in the above table using a histogram using the following Stata command:

pshare histogram, name(p08, replace)

Figure 1



We can also calculate a quantile (or decile) ratio to compare the incomes of the different quantile groups. For example, we can compare the income earned by the richest 10% of the population to that of the poorest 10% of the population by using the 90/10 ratio. From our table above this ratio is about 158. This means that the income received by the richest 10% of the population is 158 times higher than the income received by the poorest 10% of the population. The magnitude of a quantile ratio ranges from zero to infinity and the higher the magnitude, the higher the level of inequality. However, it is possible to normalize the quantile ratio so that the magnitude ranges from zero to one. We can do this by subtracting the income of the poorest quantile from the richest quantile, and then dividing this quantity by the income of richest quantile. Thus, the normalized 90/10 ratio, for example, captures the difference between the quantile income at the 90th percentile and the quantile income at the 10th percentile, as a proportion of the quantile income at the 90th percentile. The value of a normalized quantile ratio is zero when both the upper and the lower quantile incomes are equal. The value of a normalized quantile ratio reaches its maximum value of one when the lower quantile income is zero. This means that no one in the lower percentile earns any income and that the upper quantile income is positive. The value of the normalized quantile ratio becomes zero if all people in a society have equal incomes. However, a quantile ratio of zero does not necessarily mean that incomes are equally distributed across everyone in a society, as there may still be variation in incomes within the quantiles.

Among the desirable properties discussed above, the quantile ratio satisfies the anonymity, population independence, normalization, and scale invariance properties. However, it does not satisfy the transfer principles (both weak and strong). The quantile ratios are not decomposable as they are not additively decomposable and do not satisfy the sub-group consistency property. Another key limitation of using quantile ratios is that such measures only compare two income quantiles (i.e. compare only the selected percentiles), and therefore do not reflect information from the entire income distribution. Note that there are different quantile ratios used to measure inequality in the literature. Among the most commonly used measures are the proportion of income that goes to the top 1% and the top 10%, and the Palma ratio. The Palma ratio is the income share of the top 10% divided by that of the bottom 40%. The use of the Palma ratio has grown in recent years (see Doyle & Stiglitz, 2014). The motivation for using the Palma ratio as a measure of inequality is based on the empirical observation that the share of income going to the 'middle' deciles (5-9) is relatively stable across countries and over time, and accounts for about half of the gross national income. Thus, changes in income inequality are mainly due to changes in the 'tails' (Cobham, Schlögl & Sumner, 2016). We can use the following Stata command to estimate the Palma ratio:

pshare estimate pcminc [w=wgt], percentiles (40 90)

```
. pshare estimate pcminc [w=wgt], percentiles(40 90)
(sampling weights assumed)

Percentile shares (proportion)     Number of obs    =     17,710
```

pcminc	Coef.	Std. Err.	[95% Conf.	Interval]
0-40	.0459336	.00166	.0426798	.0491874
40-90	.3850588	.0105345	.3644101	.4057076
90-100	.5690076	.0118264	.5458268	.5921884

nlcom (Palma: _b[90-100] / _b[0-40])

```
Palma:  _b[90-100] / _b[0-40]
```

pcminc	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
Palma	12.38761	.6726666	18.42	0.000	11.06921	13.70601

Based on the above estimation, the income share of the top 10% was 12.4 times higher than that of the bottom 40% in South Africa in 2008.

4.2.2 Lorenz curves

A Lorenz curve is a simple graphical representation of an income distribution. A Lorenz curve is a graph of the cumulative proportion of income against the cumulative (ordered) proportion of individuals. To get a Lorenz curve we first order the population from the lowest income to the highest income. Then, on the y-axis we plot the cumulative proportion of income received for each cumulative proportion of the population, where the x-axis reflects the cumulative proportion of the population.

We use the *clorenz* Stata command from DASP. Thus, we need to install the DASP package. First, we use the *svyset* command to adjust for survey design. In the NIDS data sets, the variables "*psu*" and "*strat*" indicate primary sampling units and strata respectively.[9]

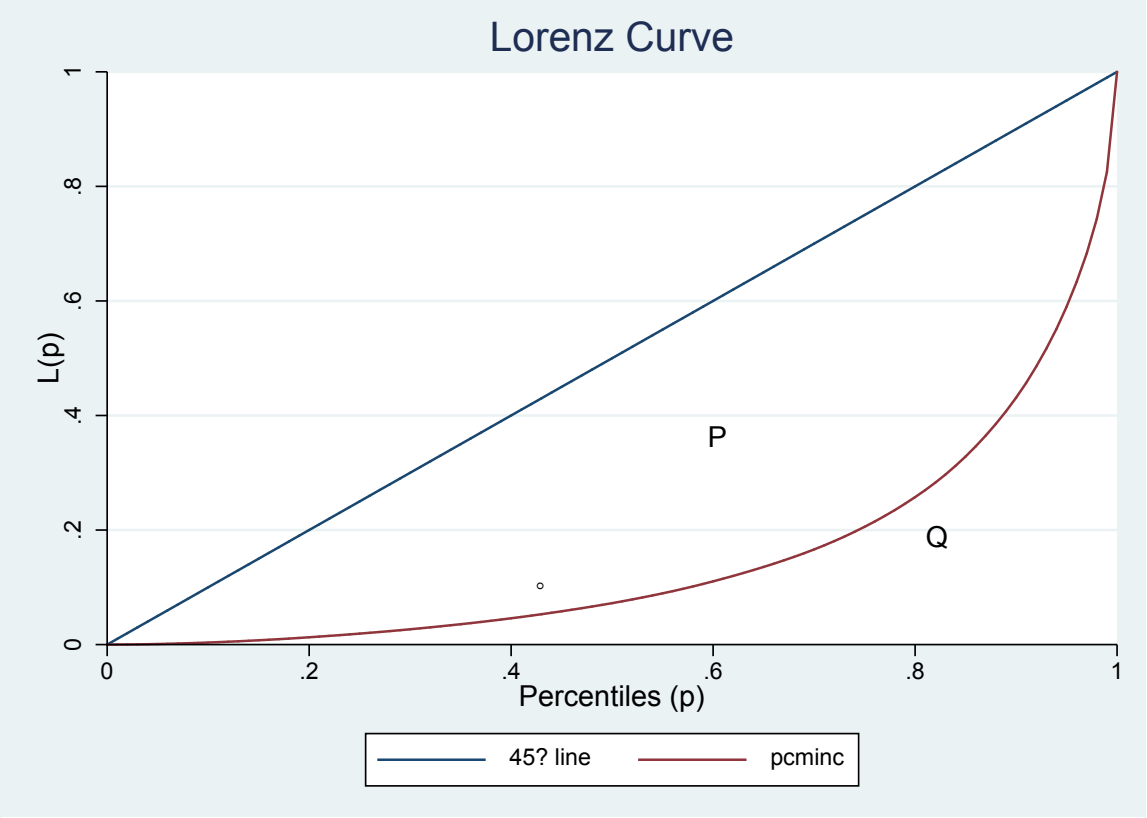
svyset psu [pw=wgt], strata(strat)

9 The sampling design for NIDS was a stratified two-stage cluster sampling (see Leibbrandt, Woolard & de Villiers,2009). Using Stats SA's 2003 Master sample of 3000 Primary Sampling Units (PSUs), 400 PSUs were selected in the first stage. PSUs are defined geographical areas consisting of at least one Enumeration Area (EA) or several EAs from the 2001 Census (Leibbrandt et al.,2009:p.9). The 53 district councils (DCs) were the explicit strata.

We obtain a Lorenz curve using the following command:

clorenz pcminc, type(nor)

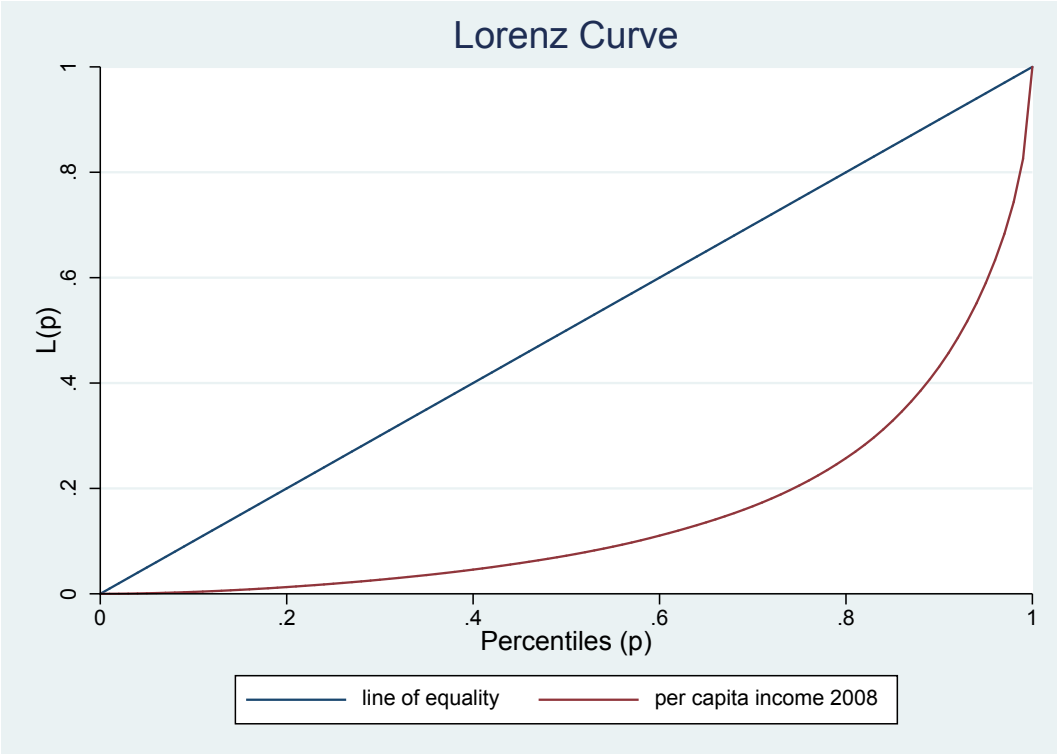
Figure 2a.



Once we obtain the graph after running the *clorenz command*, we can click on the start graph editor tab and edit different parts of the graph as we want. Another alternative is to use additional options in your command. For example, we can use the *legend* option to edit our legend descriptions in the above graph:

clorenz pcminc, type(nor) legend(order(1 “line of equality” 2 “per capita income 2008”))

Figure 2b.



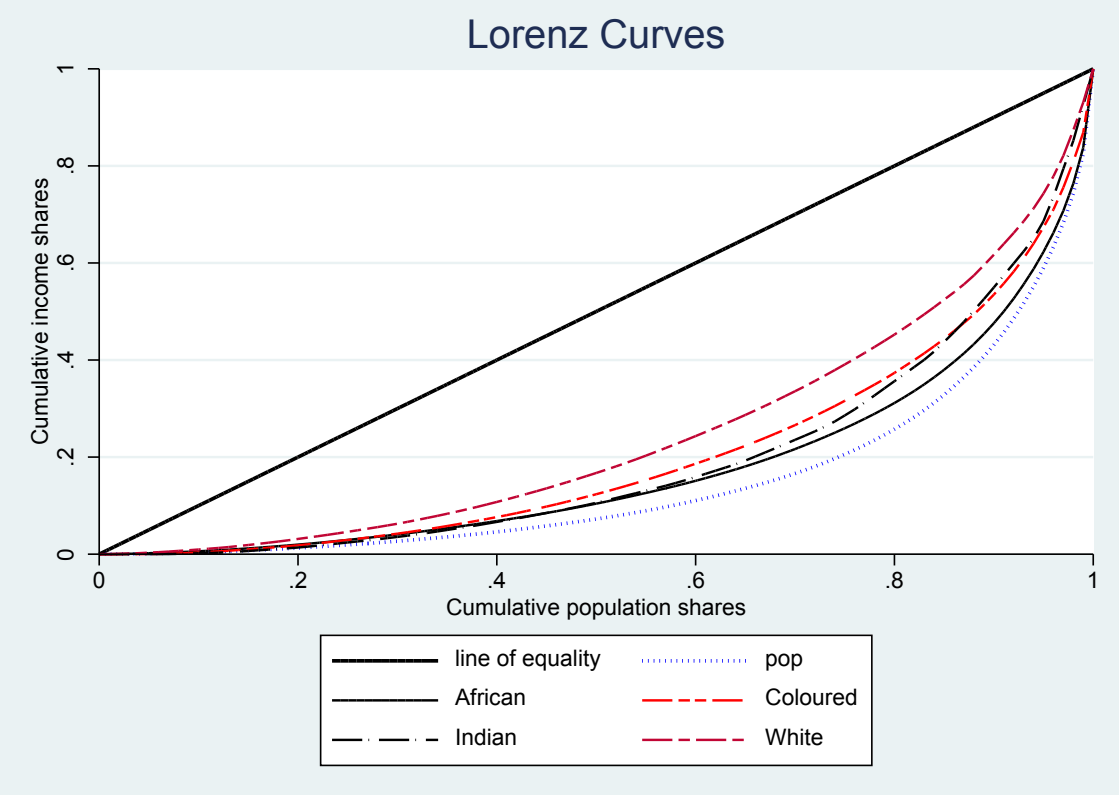
If income is distributed equally across a population size of n, then everyone receives 1/n of the total income. In this case, our Lorenz curve would be the 45° straight line graph. In reality, poor individuals receive less than 1/n of the total income and rich individuals receive more than 1/n of the total income. As a result, a Lorenz curve is a convex curve. The closer a Lorenz curve is to the 45° line, the lower is the level of inequality. We compare different income distributions using the concept of Lorenz dominance. We say distribution A Lorenz dominates distribution B if the Lorenz curve for distribution A is above (i.e. closer to the 45° line) the Lorenz curve for distribution B at all points. In such a case, we can say that the level of inequality in society A is unambiguously lower than the level of inequality in society B. However, if the Lorenz curves for the two distributions cross, we cannot compare the extent of inequality between the two distributions using Lorenz curves. We can use a generalized Lorenz curve or other inequality indices such as the Gini coefficient to compare the inequality between the two distributions.

For example, we can compare the extent of income inequality among the four race groups in South Africa using the following command:

clorenz pcminc, type(nor) hgroup(race)[10]

10 Here you can edit the graph using the graph editor after running this command, or use the following command :
clorenz pcminc, hgroup(race) ///
lpattern("1" "." "1" "_--_#" "_.#" "_-") lc("black" "blue" "black" "red" "black") lwidth("medthick" "medthick") ///
xtitle("Cumulative population shares", size(small)) ytitle("Cumulative income shares", size(small)) ///
legend(order(1 "line of equality" 2 "pop" 3 "African" 4 "Coloured" 5 "Indian" 6 "White")) ///
saving(ineqrace.gph, replace) name(ineqrace, replace)

Figure 3.

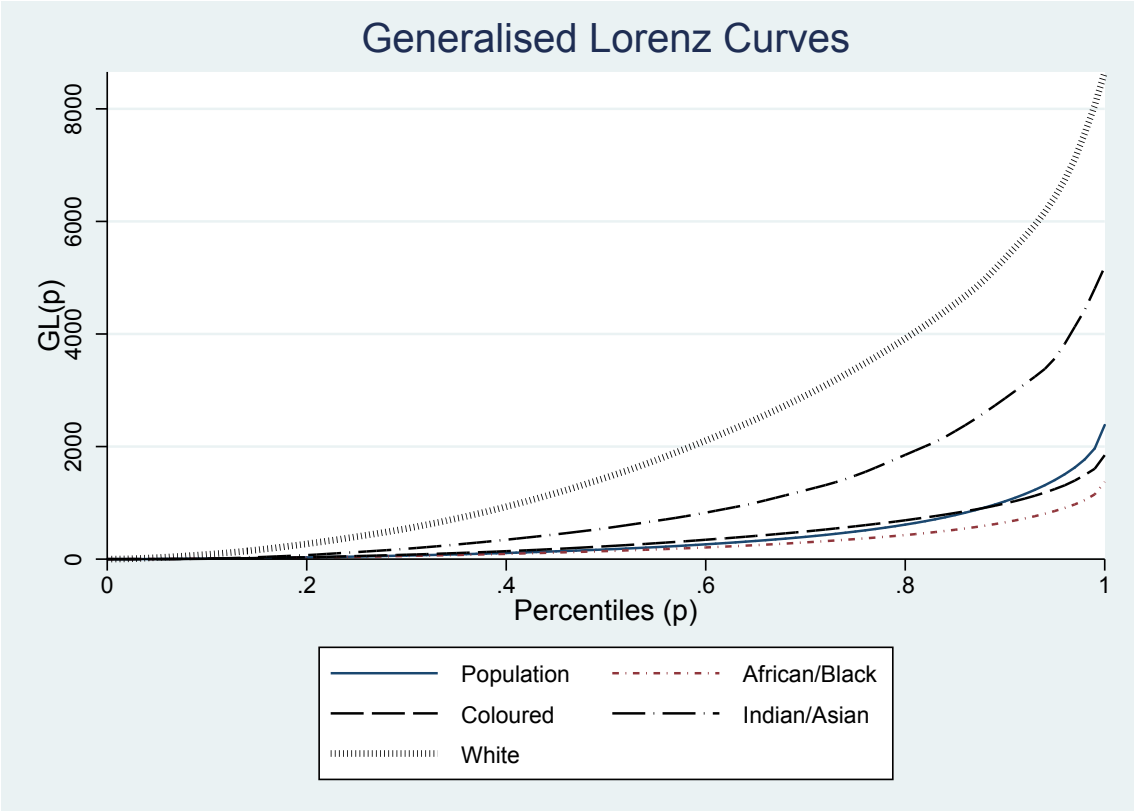


The figure indicates that within race groups, inequality is the lowest amongst Whites and highest amongst Africans. We can say that the distribution for Whites Lorenz dominates the distribution for the rest of the race groups. The distribution for Coloureds Lorenz dominates that of the Africans. However, we cannot compare the level of inequality between Indians and Coloureds since the two curves cross. In this case, we can either use a Generalized Lorenz curve or some other inequality indices such as the Gini coefficient to compare the extent of inequality among race groups. We can get a Generalized Lorenz curve by multiplying the y-coordinates of a Lorenz curve by the mean population income. By explicitly introducing the mean income, we are thus comparing distributions based on welfare grounds. Accordingly, social welfare is higher for income distributions with higher mean income, regardless of the level of inequality. We can use the following Stata command to get Generalized Lorenz curves by race:

clorenz pcminc, type(gen) hgroup(race)

Based on the Generalized Lorenz curves (see Fig. 4), the mean income for Whites is the highest followed by Indians, Coloureds, and Africans. Thus, we can say that welfare is the highest among Whites, followed by Indians, then Coloureds, and it is the lowest amongst Africans. If the Generalized Lorenz curves cross, we cannot rank income distributions using these curves only.

Figure 4.



4.2.3 The Gini coefficient

The Gini coefficient is one of the most widely used inequality measures and can be calculated from a Lorenz curve. It is the ratio of the area between the Lorenz curve and the line of equality, to the entire area below the line of equality. In Figure 2a. this is would be calculated by Area P/ (Area P + Area Q). The mathematical formula for the Gini coefficient can be stated as follows:

$$G = \sum_{i=1}^{N} \sum_{j=1}^{N} \frac{|y_i - y_j|}{2N^2\mu}$$

Where y_i and y_j indicate the income level of individual i and individual j respectively, μ is mean income, and N is population size. The Gini coefficient ranges from zero, a situation of perfect equality where income is equally distributed across everyone in a society, to one, a situation of perfect inequality where one person receives all the income. Unlike the quantile ratio measures, the Gini coefficient uses data from the entire income distribution. We can estimate the Gini coefficient corresponding to the above Lorenz curve using the following Stata command:

igini pcminc

```
. igini pcminc

    Index            :  Gini index
    Sampling weight  :  wgt
```

Variable	Estimate	STE	LB	UB
1: GINI_pcminc	0.698286	0.014947	0.668888	0.727683

Given that the maximum value for the Gini coefficient is one (representing the highest inequality), the Gini coefficient of 0.69 indicates the high level of inequality in South Africa. The Gini coefficient has a readily intuitive interpretation. If we multiply the Gini coefficient estimate by two and the mean income, we will get the expected income difference between two randomly chosen individuals in the population.[11]

We can also calculate Gini coefficients for various population groups. For example, we can use the following Stata command to generate Gini coefficient estimates disaggregated by race groups.

igini pcminc, hgroup(race)

```
    Index            :  Gini index
    Sampling weight  :  wgt
    Group variable   :  race
```

Group	Estimate	STE	LB	UB
1: African/Black	0.643054	0.016474	0.610653	0.675455
2: Coloured	0.592903	0.027139	0.539526	0.646280
3: Indian/Asian	0.610705	0.064400	0.484043	0.737368
4: White	0.506923	0.028801	0.450277	0.563569
Population	0.698286	0.014947	0.668888	0.727683

The Gini coefficient is the lowest for Whites, followed by Coloureds, Indians and Africans. The estimates suggest that inequality is the highest among Africans followed by Indians, Coloureds, and Whites. However, looking at the large confidence intervals around the Indian estimates suggests that the estimate is not very precise, due to the small sample size of the Indian/Asian population group.

We can also calculate Gini coefficients disaggregated by geographic locations. The table below shows income inequality estimates by geographic locations (i.e. rural/urban). Based on the Gini coefficient estimates, income inequality is higher in urban areas relative to rural areas.

11 We can re-write the formula for the Gini coefficient as follows:

$$2\mu G = \sum_{i=1}^{N}\sum_{j=1}^{N}\frac{|y_i - y_j|}{N^2}$$

The right-hand side of the equation represents the expected income difference between two randomly chosen individuals in the population.

igini pcminc, hgroup(rural)

```
Index              :  Gini index
Sampling weight    :  wgt
Group variable     :  rural
```

Group	Estimate	STE	LB	UB
1: 0. Urban	0.669496	0.016767	0.636520	0.702473
2: 1. Rural	0.593243	0.028197	0.537785	0.648700
Population	0.698286	0.014947	0.668888	0.727683

Ideally, it is also possible to disaggregate inequality by gender. However, as our discussion in Section 3 indicated, estimating income inequality disaggregated by gender requires detailed individual-level data on consumption and income for each household member. Almost all household surveys collect consumption data at the household level. Likewise, with few exceptions, income and expenditure surveys collect information on income at the household level. For these reasons, inequality estimates are often disaggregated by the gender of the household head only. For example, in our case, the table below presents income inequality estimates disaggregated by the gender of the household head. The figures indicate only slightly higher inequality among individuals living in male-headed households compared to female-headed households.

igini pcminc, hgroup(hhhead)

```
Index              :  Gini index
Sampling weight    :  wgt
Group variable     :  hhhead
```

Group	Estimate	STE	LB	UB
1: male_head	0.699584	0.017137	0.665879	0.733290
2: female_head	0.681727	0.015050	0.652127	0.711327
Population	0.698286	0.014947	0.668888	0.727683

NIDS is one of the exceptions among income and expenditure household surveys that collect income information at an individual level (for adults only). We can use this information to estimate income inequality disaggregated by gender. The table below shows the income inequality estimates by gender. The variable that we use in this case is not income per capita, but individual-level income for each adult individual in a household. (Note that those not earning any income are assigned a zero income value.) A similar estimation can be done using information on earnings or wages at the individual level from labour force surveys.[12] This allows us to estimate inequality in earnings or wages for employed individuals disaggregated by gender. However, given that households share income or other resources, using such information directly to estimate income inequality or poverty (i.e. using individual-level income) may be problematic (see Section 3).

12 This is one way of estimating labour market outcomes. See Wittenberg (2017) for recent estimates on wage inequality in South Africa.

```
. igini totpinc, hgroup(gender)

    Index            :  Gini index
    Sampling weight  :  w1_wgt
    Group variable   :  gender
```

Group	Estimate	STE	LB	UB
1: 1. Male	0.642041	0.020633	0.601460	0.682621
2: 2. Female	0.650411	0.022168	0.606811	0.694010
Population	0.683918	0.016387	0.651688	0.716149

An alternative is to use welfare indicators that can be measured at individual levels such as educational attainment. For example, we use data on years of schooling completed to estimate educational inequality by gender (for those aged 15 years and above).[13]

```
.  igini educ_yrs, hgroup(gender)

    Index            :  Gini index
    Sampling weight  :  wgt
    Group variable   :  gender
```

Group	Estimate	STE	LB	UB
1: female	0.247800	0.005909	0.236179	0.259421
2: male	0.229272	0.005900	0.217667	0.240877
Population	0.239044	0.005110	0.228993	0.249094

The estimates above show that the Gini coefficient for years of schooling is slightly higher among females than it is for males, suggesting higher educational inequality among females relative to males. However, unlike income or consumption data which are continuous variables, data on years of schooling is a discrete variable and often has a significant amount of zero values (in most poor countries). In such cases, it is suggested to use the user-written command "ineqdec0", which calculates the Gini coefficient for data including a significant amount of zero values.

The Gini coefficient satisfies all of the invariance properties (symmetry, population invariance, scale invariance, and normalization) and the transfer principle. However, the Gini coefficient does not satisfy the transfer sensitivity property. The Gini coefficient can be decomposable, but with an added residual term. Note that the Gini coefficient does not satisfy the subgroup consistency property.

13 Often the age is restricted to those aged 25 or greater, due to the assumption that by age 25 most adults have completed their schooling.

4.2.4 The Generalized entropy measures

If we want an inequality measure that is additively decomposable and satisfies the subgroup consistency property, then we can consider the entropy class of inequality measures. The mathematical formula for this class of inequality measures is given as follows:

$$GE(\alpha) = \frac{1}{\alpha\,(\alpha - 1)} \left[\frac{1}{N} \sum_{i=1}^{N} \left(\frac{y_i}{\mu} \right)^{\alpha} - 1 \right]$$

Where, y_i indicates individual income, μ is mean income, and N is population size .The parameter α in the GE class of inequality measures represents the weight given to differences between incomes at different parts of the income distribution, and it can take any real value. With a positive and large α, the index GE will be more sensitive to changes at the upper tail of the income distribution. The GE index will be more sensitive to changes at the bottom tail of the income distribution for α values closer to zero. The GE measures vary between zero and infinity, with zero representing an equal distribution (incomes in a society are equally distributed across all people) and a higher value representing a higher level of inequality. The value of the upper limit, however, depends on the specific value of α. The most common values of α used are 0,1 and 2.[14] The GE (1) index is called the Theil's T index, and the GE (0) is called the Theil's L index (mean logarithmic deviation).[15]

The formula for the Theil's T index, GE (1), is given by:

$$T_T = \frac{1}{N} \sum_{i=1}^{N} \left(\frac{y_i}{\mu} \right) ln \left(\frac{y_i}{\mu} \right)$$

While the formula for the Theil's L index, GE (0), is given by:

$$T_L = -\frac{1}{N} \sum_{i=1}^{N} ln \left(\frac{y_i}{\mu} \right)$$

The upper limit for GE (1) is ln(N) while the corresponding values for GE (0) is unbounded. In both cases, individuals with zero incomes will automatically be dropped from the

14 For negative values of α the GE (α) class of inequality measures are undefined in the presence of zero income values. Thus, in practice, positive values of α used.

15 We cannot derive the Theil's T and Theil's L indices from the GE (α) equation by directly substituting for α= 0 or α=1. The derivation requires using L'Hôpital's Rule. The procedure requires us to first differentiate the denominator and the numerator separately, and then calculate the ratio of the limits of each of these functions.

calculation given that log zero is undefined. This is problematic if zero incomes indicate a genuine value. One approach often used is to replace individuals with zero incomes with small values such as 1 (Jenkins & Jantti, 2005).

We can use the following Stata commands to calculate the Theil's T and Theil's L inequality indices:

For GE(0)

ientropy pcminc, theta(0)

for GE(1)

ientropy pcminc, theta(1)

```
    Index              :  Entropy index
    Parameter theta    :  0
    Sampling weight    :  wgt
```

Variable	Estimate	STE	LB	UB
1: entropy_pcminc	1.054390	0.057380	0.941535	1.167244

```
    Index              :  Entropy index
    Parameter theta    :  1
    Sampling weight    :  wgt
```

Variable	Estimate	STE	LB	UB
1: entropy_pcminc	1.018017	0.063061	0.893989	1.142045

we can also estimate the GE(α) class of indices disaggregated by groups. For example, the GE(1) index disaggregated by race is given as follows:

ientropy pcminc, hgroup(race) theta(1)

```
    Index              :  Entropy index
    Parameter theta    :  1
    Sampling weight    :  wgt
    Group variable     :  race
```

Group	Estimate	STE	LB	UB
1: African/Black	0.885423	0.073779	0.740315	1.030531
2: Coloured	0.700526	0.079994	0.543194	0.857858
3: Indian/Asian	0.686213	0.175553	0.340935	1.031491
4: White	0.464310	0.056843	0.352510	0.576109
Population	1.018017	0.063061	0.893989	1.142045

The estimates show that inequality is the highest amongst Africans while it is the lowest amongst Whites. This result is consistent with the inequality estimates that we obtained using the Gini coefficient. However, we cannot compare the estimates that we obtained from the GE index with that of the Gini coefficients. The value of the Gini coefficient varies from 0 to 1, while values for the GE class of inequality measures range from zero to infinity. Note that these measures may generate a different inequality ranking for the same distribution, because the sensitivity of the various inequality indices to differences between incomes at different parts of the income distribution varies.

The GE class of inequality measures satisfy all of the invariance axioms: population invariance, scale invariance, normalization, and symmetry. In addition, for $\alpha < 2$, the GE measures are transfer sensitive. One of the key advantages of using the GE class of inequality indices is that, unlike the Gini index, this class of inequality measures are additively decomposable and satisfy the subgroup consistency axiom. Thus, we can use the GE class of inequality measures to decompose overall inequality into between and within group components. We can use different factors as our grouping variable including race, gender, location/regions, and sources of income. For instance, using race as our grouping variable we can decompose overall income inequality in South Africa into "between-race group" and "within-race group" components. We can decompose the GE(1) index into between- and within-group components using the following command:

dentropyg pcminc, hgroup(race) theta(1)

```
Decomposition of the Generalised Entropy Index by Groups
Sampling weight :  wgt
Group variable  :  race
Parameter theta :  1.00
```

Group	Entropy index	Population share	(mu_k/mu)^theta	Absolute contribution	Relative contribution
1: African/Black	0.885423	0.756531	0.576466	0.386146	0.379312
	0.073779	0.025794	0.056654	0.058966	0.066516
2: Coloured	0.700526	0.097209	0.775567	0.052814	0.051880
	0.079994	0.015724	0.114612	0.012392	0.013490
3: Indian/Asian	0.686213	0.029545	2.180964	0.044217	0.043435
	0.175553	0.011521	0.782087	0.017998	0.017887
4: White	0.464310	0.116714	3.633287	0.196893	0.193409
	0.056843	0.018548	0.380633	0.037001	0.028353
Within	---	---	---	0.680071	0.668035
	---	---	---	0.063617	---
Between	---	---	---	0.350333	0.344133
	---	---	---	0.017840	---
Population	1.018017	1.000000	---	1.018017	1.000000
	0.063061	0.000000	---	0.063061	0.000000

All observations with missing data on the income variable should be dropped prior to running the *dentropyg* command. Based on the relative contribution, about 66% of the overall income inequality in South Africa in 2008 was due to inequality within race groups, while 34% of the income inequality is due to inequality between race groups.

We can also decompose income inequality by income sources. There are various methods (regression and non-regression techniques) used to decompose income inequality by income sources (see e.g. Shorrocks, 1982, 2013; Fields, 2003). We can use the "dsineqs" DASP module to decompose income inequality by income sources, which uses the Shapley decomposition method. According to Shorrocks (2013:p.101), the Shapley decomposition procedure involves calculating the marginal effect on inequality of "*eliminating each of the contributory factors in sequence, and then assigns to each factor the average of its marginal contributions in all possible elimination sequences.*" Thus, the method allows for the decomposition of inequality measures without a residual. Using the Shapley decomposition procedure, we can decompose income inequality by income sources using the Gini, Atkinson and Generalized entropy inequality indices. Note that although we can use the Shapley decomposition procedure to decompose inequality using the Gini index, the procedure does not solve the subgroup inconsistency problem associated with the Gini index.

We use the 2008 NIDS data to decompose income inequality by income sources. We consider five income sources: wage income (wage), income from social grants (grants), income from remittances (remittance), income from capital (capital), and income from other sources (other). Use the following command to decompose income inequality by income sources using the GE(1) index.

dsineqs wage other grants remittance capital, index(ge) theta(1)

```
Decomposition of the inequality index by income components (using the Shapley value).
Execution  time :          2.34 second(s)
ineq index      :      1.484097
Sampling weight :  w1_wgt
```

Sources	Income Share	Absolute Contribution	Relative Contribution
1: pwageinc	0.791420	0.998938	0.673094
2: potherinc	0.073475	0.185506	0.124996
3: pgrantinc	0.078498	0.066625	0.044893
4: premitinc	0.044207	0.142528	0.096037
5: pcapitinc	0.012400	0.090501	0.060980
Total	1.000000	1.484097	1.000000

We can also use the Gini index to decompose income inequality by income sources.

dsineqs wage other grants remittance capital, index(gini)

```
Decomposition of the inequality index by income components (using the Shapley value).
Execution  time :          3.44 second(s)
ineq index      :      0.723094
Sampling weight :  w1_wgt
```

Sources	Income Share	Absolute Contribution	Relative Contribution
1: pwageinc	0.791420	0.595506	0.823553
2: potherinc	0.073475	0.062187	0.086001
3: pgrantinc	0.078498	0.022013	0.030442
4: premitinc	0.044207	0.032681	0.045196
5: pcapitinc	0.012400	0.010707	0.014808
Total	1.000000	0.723094	1.000000

From the above table, it is clear that wage income is the main driver of income inequality in South Africa. Based on the Gini index decomposition, about 82% of the income inequality was due to wage income. This is not surprising given that the share of wage income out of total income is close to 80%.

4.2.5 *The Coefficient of variation*

The coefficient of variation (CV) is another commonly used inequality measure. In particular, the coefficient of variation is used in the analyses of spatial and horizontal inequality measures. The basic formula for calculating the coefficient of variation is given as follows:

$$CV = \frac{\sqrt{\sum_i^N (Y_i - \bar{Y})^2 / N}}{\bar{Y}}$$

Where Y_i indicates the income of individuals, $\bar{Y}$ is mean income, and N is the number of individuals. The range of CV values goes from zero to infinity, with higher values representing a more unequal income distribution. Of the desirable properties, the CV satisfies the anonymity, normalization, scale invariance, and the transfer principle. It is also additively decomposable and satisfies the subgroup consistency property. However, it does not satisfy the transfer sensitivity axiom. The CV is also affected by extreme values as it depends on the square of the distance between the mean value and individual values.[16] We will discuss the CV further in sub-section 4.5, when we discuss approaches to measuring spatial inequality.

16 Other inequality measures are also affected by the presence of extreme values in the data. For example, Cowell and Flachaire (2007) showed that the GE class of inequality measures with $\alpha > 1$ are very sensitive to the presence of higher incomes in the data. Likewise, the GE class of indices with $\alpha < 0$ and the Atkinson index with > 1 are very sensitive to the presence of very small incomes in the data. In contrast, the Gini coefficient is less sensitive to the presence of extreme values in the data.

4.2.6 The Atkinson class of measures

Thus far, the inequality indices that we have discussed above are descriptive inequality indices, derived without explicitly incorporating social welfare functions. However, it is argued that these inequality measures are also used for policy formulation with some implicit value judgments (Atkinson, 1970). Atkinson proposed a welfare-based inequity measure called the Atkinson's class of inequality measures. The formula for the Atkinson's class of inequality measures is given by:

$$I = 1 - \left[\frac{1}{N} \sum_{i=1}^{N} \left(\frac{y_i}{\mu} \right)^{(1-\varepsilon)} \right]^{\frac{1}{(1-\varepsilon)}}$$

Where, y_i indicates individual income, μ is mean income, and N is population size. The parameter in the Atkinson class of inequality measures represents an inequality aversion parameter and can take values between zero and infinity. The most commonly used values are 0.5, 1.5,1, or 2. The choice of these parameters is somewhat arbitrary. Higher values of the aversion parameter imply that social welfare is more sensitive to a shift in the income of a poorer individual than it is to the same shift affecting a richer individual.

The values for all indices in the Atkinson class of indices vary from zero (i.e. perfect equality) to one (i.e. maximum inequality). The Atkinson class of inequity measures satisfy all of the invariance axioms (population invariance, scale invariance, symmetry, and normalization). This class of inequality measures also satisfy the transfer principle and transfer sensitivity axioms. Although this class of inequality measures satisfy the subgroup consistency property, they are not additively decomposable. However, as discussed in sub-section 4.2, we can use the Shapley decomposition procedure to decompose the Atkinson indices without the residual effect.

We use the following Stata command to estimate the Atkinson index for a ε value of 1.5.

iatkinson pcminc, epsilon(1.5)

```
Index              :  Atkinson index
Parameter epsilon  :  1.5
Sampling weight    :  wgt
```

Variable	Estimate	STE	LB	UB
1: atk_pcminc	0.809247	0.016304	0.777180	0.841315

Given that the values of the Atkinson index vary from zero to one, the value of 0.809 is close to one, thus indicating a high level of inequality. Like the Gini coefficient, the Atkinson index also has an intuitive interpretation. The Atkinson index measures the welfare loss to a society due to inequality. For example, the index value of 0.809 would mean that about 80% of the current income is lost (wasted) due to inequality. In other words, the society would need only 19.1% of the current national income to achieve the same

level of social welfare if all incomes were distributed equally.

We can also estimate the Atkinson index disaggregated by race as follows:

iatkinson pcminc, hgroup(race) epsilon(1.5)

```
Index               :  Atkinson index
Parameter epsilon   :  1.5
Sampling weight     :  wgt
Group variable      :  race
```

Group	Estimate	STE	LB	UB
1: African/Black	0.713273	0.019869	0.674196	0.752351
2: Coloured	0.795259	0.038163	0.720201	0.870317
3: Indian/Asian	0.926112	0.041680	0.844136	1.008087
4: White	0.697085	0.071384	0.556686	0.837484
Population	0.809247	0.016304	0.777180	0.841315

In this case, income inequality is the highest amongst Indians followed by Coloureds and Africans and it is the lowest among Whites. Thus, the estimation results differ from what we obtained when we used the Gini coefficient. Although the values of both the Gini coefficient and the Atkinson indices vary between zero and one, the two-inequality measures can result in a different ranking of a set of distributions. This is possible because the inequality indices have varying levels of sensitivity to differences in income at different parts of the income distribution. In addition, as mentioned earlier, the large confidence intervals around the Indian estimates indicates that the estimates for this group are not very precise, probably due to the small relevant sample size.

4.3 Inequality dynamics

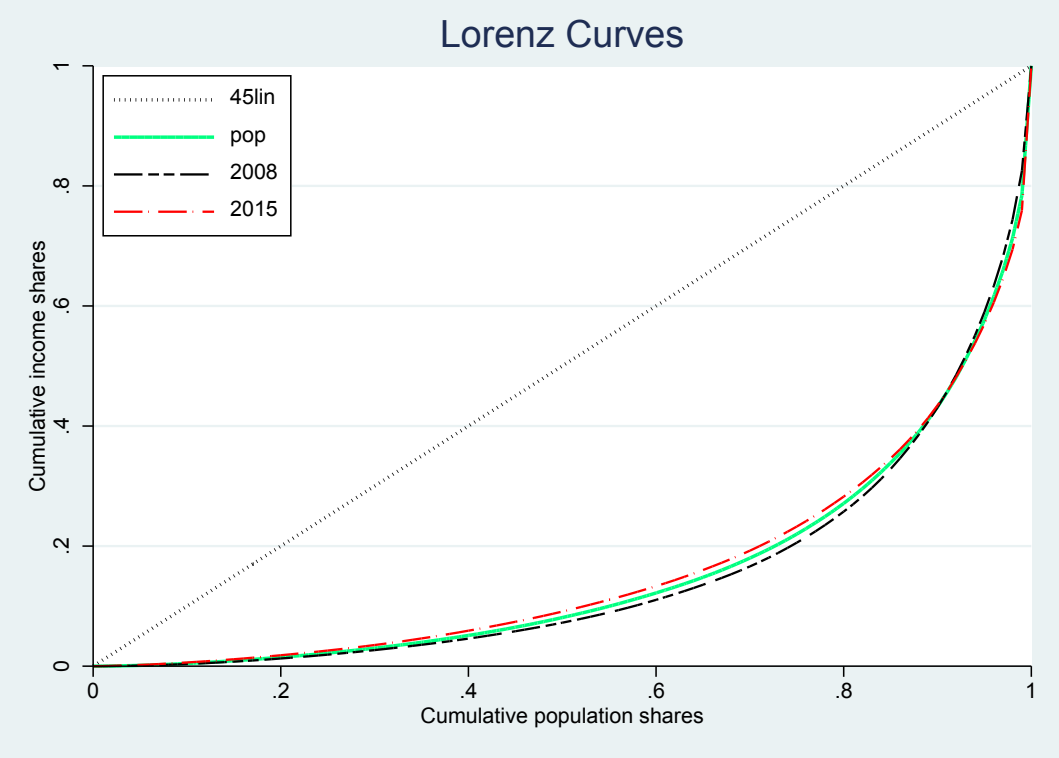
Evaluating the impact of any policy on social welfare requires the analysis of trends in poverty and inequality. In order to estimate trends in inequality, we need data for at least two time points. What is important in this regard is that the data and welfare measure that we are using should be consistent across time. As discussed in Section 3, it is possible that changes in measured inequality over time could be due to real changes in the income distribution, or due to other factors such as changes in data collection, price adjustments or other methodological changes. Once we are comfortable with the data and measurement issues, we can use any of the above inequality measures to compare inequality across time. For example, we can compare income inequality between 2008 and 2015 using Lorenz curves as follows:

First, we need to create a real income variable using the 2008 prices as a base. The CPI value for 2008 was 63.6 (annual average) and it was 92.0 (annual average) in 2015. This means that we multiply our 2015 *pcminc* variable by 63.6/92.0 to get a real per capita

income variable; *real_pcminc*.[17] Then, use this variable to estimate inequality over time. The command to estimate Lorenz curves for multiple years, which in our case are 2008 and 2015, is:

clorenz real_pcminc, hgroup(year)

Figure 5.



Note that because the Lorenz curves for the 2008 and 2015 distributions cross, we cannot draw any conclusion with regard to trends in income inequality across these periods. We can use other inequality indices such as the Gini coefficient and the Atkinson index (we can choose ε ==1.5 for example) to compare inequality over time.

```
. igini real_pcminc,hgroup(year)

    Index            :  Gini index
    Group variable   :  year
```

Group	Estimate	STE	LB	UB
1: 2008	0.661186	0.017534	0.626792	0.695579
2: 2015	0.599787	0.032764	0.535519	0.664055
Population	0.630340	0.020799	0.589544	0.671137

17 The *pcminc* values for 2008 will not need any adjustment, as we are setting the base year for the inflation adjustment to 2008. Thus for 2008, *pcminc* and *real_pcminc* are identical.

```
. iatkinson real_pcminc , hgroup(year) epsilon(1.5)

    Index               :  Atkinson index
    Parameter epsilon   :  1.5
    Sampling weight     :  wgt
    Group variable      :  year
```

Group	Estimate	STE	LB	UB
1: 2008	0.809247	0.016206	0.777460	0.841035
2: 2015	0.718899	0.031568	0.656979	0.780820
Population	0.766667	0.018419	0.730538	0.802796

Results from the Gini coefficient and Atkinson indices suggest that inequality in 2015 is lower than the level in 2008, indicating a decline in income inequality over the relevant time period.

We can also decompose the change in inequality into within-group and between-group components and compare this over time. For example, we can decompose income inequality by race for 2008 and 2015 and examine whether the contribution of the within-group or between-group inequality changed over time. To do this, we have to do the decomposition separately for each year. The relevant code and output are shown below together with the estimation results.

In 2008 the within-race groups inequality contributed about 66% to overall inequality. This percentage increased to 75% in 2015. We conclude that the contribution of within-race groups inequality has thus increased over time, while the contribution of the between-race group inequality has declined.

Inequality decomposition by race for 2008:

```
. preserve

. keep if year == 2008
(27,105 observations deleted)

. dentropyg real_pcminc, hgroup(race) theta(1)

    Decomposition of the Generalised Entropy Index by Groups
    Sampling weight :  wgt
    Group variable  :  race
    Parameter theta :  1.00
```

Group	Entropy index	Population share	(mu_k/mu)^theta	Absolute contribution	Relative contribution
1: African/Black	0.885423	0.756531	0.576466	0.386146	0.379312
	0.073779	0.025794	0.056654	0.058966	0.066516
2: Coloured	0.700526	0.097209	0.775567	0.052814	0.051880
	0.079994	0.015724	0.114612	0.012392	0.013490
3: Indian/Asian	0.686213	0.029545	2.180964	0.044217	0.043435
	0.175553	0.011521	0.782087	0.017998	0.017887
4: White	0.464310	0.116714	3.633287	0.196893	0.193409
	0.056843	0.018548	0.380633	0.037001	0.028353
Within	---	---	---	0.680071	0.668035
	---	---	---	0.063617	---
Between	---	---	---	0.350333	0.344133
	---	---	---	0.017840	---
Population	1.018017	1.000000	---	1.018017	1.000000
	0.063061	0.000000	---	0.063061	0.000000

```
. restore
```

Inequality decomposition by race for 2015:

```
. preserve

. keep if year == 2015
(18,480 observations deleted)

. dentropyg real_pcminc, hgroup(race) theta(1)
```

Decomposition of the Generalised Entropy Index by Groups
Sampling weight : wgt
Group variable : race
Parameter theta : 1.00

Group	Entropy index	Population share	(mu_k/mu)^theta	Absolute contribution	Relative contribution
1: African/Black	0.742994	0.782785	0.632436	0.367828	0.308164
	0.049069	0.024854	0.073726	0.052768	0.109988
2: Coloured	0.659462	0.093881	0.826108	0.051145	0.042849
	0.168030	0.018313	0.165617	0.020127	0.022090
3: Indian/Asian	0.643094	0.026580	2.106360	0.036005	0.030165
	0.071416	0.012041	0.648521	0.015403	0.016512
4: White	1.182039	0.096753	3.838578	0.439003	0.367793
	0.440893	0.015091	0.683283	0.235720	0.109109
Within	---	---	---	0.893981	0.748971
	---	---	---	0.289221	---
Between	---	---	---	0.294447	0.246685
	---	---	---	0.019278	---
Population	1.193613	1.000000	---	1.193613	1.000000
	0.291455	0.000000	---	0.291455	0.000000

4.4 Multidimensional inequality measures (asset indices)

Our discussion thus far has focussed on measuring economic inequality using a uni-dimensional measure of wellbeing, which is per capita income or consumption. However, our discussion in Section 2 indicated that inequality can have many dimensions including education, assets, health, and others, while income may not be an adequate measure of individual wellbeing. Thus, it is well recognized that income inequality is only a proxy measure of either wellbeing inequality or economic inequality (Sen, 1992). Sen argued that there is individual heterogeneity in converting income or other resources into well-being. Thus, the living conditions of individuals should be assessed in terms of actual wellbeing achievements (functionings) and the ability to achieve (capabilities). The actual achievements can include being well-nourished, educated, and being healthy. Based on Sen's capability approach, recent studies try to measure poverty and inequality using a multidimensional approach (see Alkire & Foster, 2011).

With regard to measuring inequality, recent studies have used asset-based living standard indicators to estimate multidimensional inequality (McKenzie, 2005; Wittenberg & Leibbrandt, 2017). Ownership of household assets (e.g. TV, fridge, livestock etc.) and access to basic services (e.g. access to water, sanitation, household building materials etc.) have been used to measure inequality in a multidimensional sense. Before we use some of the inequality indices discussed above, we need to first combine these indicators into

a single index (often called "asset indices"). Statistical approaches such as factor analysis (FA), principal component analysis (PCA), and multiple correspondence analysis (MCA) are the common approaches used to calculate asset indices in the literature (Filmer & Pritchett, 2001; Wittenberg & Leibbrandt, 2017). If we have k living standard indicators (a_1, a_2, …,a_k) we can combine these indicators into a single index using the following formula:

$$Index = w_1a_1+w_2a_2 + \cdots+w_ka_k$$

where $w_1, w_2,\ldots, w_k$ indicates weights associated with each indicator. If we use PCA, the weights are obtained from the first "principal component", which is a linear combination that accounts for the highest variance in the asset distribution. We can write each indicator, a_i, as a linear combination of k factors or components as follows:

$$a_1 = v_{11}A_1 + v_{12}A_2 + \cdots v_{1k}A_k$$

$$a_2 = v_{21}A_1 + v_{22}A_2 + \cdots v_{2k}A_k$$

$$a_k = v_{k1}A_1 + v_{k2}A_2 + \cdots v_{kk}A_k$$

where A_1, A_2,…,A_k are unobserved components that are uncorrelated with each other. Then it can be shown that the solution will be of the form:

$$A_1 = v_{11}\tilde{a}_1 + v_{12}\tilde{a}_2 + \cdots v_{1k}\tilde{a}_k$$

Where $\tilde{a}_{1i}$ indicates a standardised asset variable, $\tilde{a}_{1i}= \frac{a_{1i}- \overline{a_1}}{s_1}$

Where $\overline{a_1}$ and s_1 indicate the mean and standard deviation of the asset variable, respectively. The first principal component, A_1, is the component which explains the largest portion of the common covariance of the asset variables. We can consider "Wealth" as the underlying unobserved variable which is the common factor (A_1).[18] Thus, a higher asset index implies a higher "wealth".

One problem with using PCA and other similar approaches is that some assets such as livestock (mainly owned by rural households) could be assigned negative weights. Thus, we could end up ranking rural households with livestock lower than households with no assets at all (Wittenberg & Leibbrandt, 2017). In addition, asset indices constructed

18 If we want to measure wealth directly, we need to collect detailed information on both financial and non-financial assets and debts. Then, net wealth is calculated subtracting the total value of debts from the total value of assets. We use this variable to estimate wealth inequality. Unfortunately, such information is rarely available in household surveys.

using these approaches have zero mean values by construction (McKenzie, 2005; Wittenberg & Leibbrandt, 2017). In this case, using conventional inequality measures is not appropriate. To solve these problems, Wittenberg and Leibbrandt (2017) suggested the use of the uncentered PCA (UC PCA) approach in calculating the asset indices, adopting a method that was initially proposed by Banerjee (2010). Following Wittenberg and Leibbrandt (2017), we can use the asset indices produced using the UC PCA approach to estimate inequality using this one conventional inequality indices such as the Gini coefficient. To illustrate this point, we use data from the 1998 South African DHS and use the following variables to create an asset index using both the PCA and UC PCA approaches. The asset variables in the DHS or other household surveys are measured at a household level. Thus, we calculate asset indices at a household level. Because there is no standard way to calculate per-capita asset index values, everyone in a household will be assigned the same asset index value calculated at a household level.

Variables used for calculating an asset index (DHS,1998)

Variable	Obs	Mean	Std. Dev.	Min	Max
water_inhouse	12,247	0.353719	0.478143	0	1
electricity	12,247	0.616641	0.486225	0	1
radio	12,247	0.790071	0.407274	0	1
television	12,247	0.54495	0.497996	0	1
refrigerator	12,247	0.465665	0.49884	0	1
car	12,247	0.226831	0.4188	0	1
rooms	12,136	2.201714	1.103846	0	12
telephone	12,247	0.255736	0.436292	0	1
Computer	12,247	0.051278	0.220573	0	1
Washing_machine	12,247	0.186576	0.389587	0	1
Donkey/horse	12,247	0.033559	0.180099	0	1
Sheep/cattle	12,247	0.124765	0.330466	0	1

With the exception of the "rooms" variable, all the variables are dummy variables indicating ownership of the asset in a house. The variable "rooms" measures the number of rooms, which is a count variable. We use the *pca* Stata command for calculating an asset index using the PCA approach as follows:

pca water_inhouse electricity radio television refrigerator car rooms telephone computer washing_machine donkeyhorse sheepcattle [weight=pwt]

predict pcaindex

The variable *pcaindex* is the asset index variable that is created based on the PCA first principal component. If we want to see the coefficient estimates on the asset variables, we need to regress the asset index variable on the asset dummy variables as follows:

reg pcaindex water_inhouse electricity radio television refrigerator car rooms telephone

computer washing_machine donkeyhorse sheepcattle

Unlike the PCA, we do not have a Stata command to calculate an asset index using the UC PCA approach. For this reason, we have to first run an ado file created by Martin Wittenberg (ucpc.ado).[19] Then, run the following command:

ucpc water_inhouse electricity radio television refrigerator car rooms telephone computer washing_machine donkeyhorse sheepcattle [weight=pwt], gen(ucpcindex)

reg ucpcindex water_inhouse electricity radio television refrigerator car rooms telephone computer washing_machine donkeyhorse sheepcattle

Our asset index variable generated using the UC PCA approach is the *ucpcindex* variable. The coefficient estimates on the asset index variables are given in the table below:

Coefficient estimates on asset variables

variables	PCA	UCPCA
water_inhouse	0.729	0.569
electricity	0.690	0.219
radio	0.479	0.138
television	0.699	0.271
refrigerator	0.760	0.369
car	0.770	1.211
rooms	0.096	0.048
telephone	0.832	1.002
computer	0.955	15.048
Washing machine	0.879	1.715
Donkey/horse	-0.344	4.646
Sheep/cattle	-0.408	0.494
_cons	-2.750	0.000

As we can see, the coefficient estimates on the livestock variables are negative in the case of the PCA approach while they are positive in the case of the UC PCA approach. Once we have our asset index variable, we can calculate the Gini coefficient or other standard inequality measures based on the ucpcindex asset index variable. For example, we can calculate the Gini coefficient based on the ucpcindex variable using the following command:

```
svyset cluster_num [pw=pwt]

igini ucpcindex, hsize(hhsize)
```

19 Please see the appendix for the ado file.

```
. igini ucpcindex, hsize(hhsize)

    Index            :  Gini index
    Household size   :  hhsize
    Sampling weight  :  pwt
```

Variable	Estimate	STE	LB	UB
1: GINI_ucpcindex	0.639234	0.006709	0.626067	0.652401

In our command above, we use the hsize(hhsize) option because our observations are at a household level. Thus, in order to get inequality estimates at the individual level, we have to weight household-level observations by household size (hhsize). The estimated coefficient based on the ucpcindex variable is 0.64. Thus, based on the Gini index the multidimensional Gini coefficient was 0.64 in 1998.

Based on the asset index calculated above, we can calculate asset inequality measures disaggregated by groups such as race or regions (the same way we did using the income variable above). We can also compare asset inequality over time or across regions/countries. However, in making comparisons across time or countries care should be taken in generating the asset indices because the weights generated using the statistical approaches discussed above may vary across countries or over time (e.g. asset distributions may vary). For instance, in comparing asset inequality over time, we need to use a common set of assets. Then, we can use two approaches to generate weights: either we can generate an asset index after pooling the data over time, or we can calculate the weights using data from one time period and then use the same set of weights for other time periods.

4.5 Spatial inequality

Spatial inequality is concerned with measuring inequality between geographical units of a country (or a region). As such, the unit of analysis is a geographical unit (i.e. province, municipality,…, etc) and all individuals in a given geographical unit are assigned the same level of income (i.e. per- capita income level of that geographical unit). Following Williamson's (1965) work, the standard approach used to measure spatial/regional inequality is to use the coefficient of variation, which is discussed in Section 5. However, unlike measuring interpersonal inequality, we use the population weighed coefficient of variation, which is calculated using the following formula:

$$CV_w = \sqrt{\frac{\sum_i^m (Y_i - \bar{Y}_w)^2 * p_i}{\bar{Y}_w}}$$

Where CV_w is the population weighted coefficient of variation estimate for a given region or country, Y_i is the per-capita income of sub-region *i* within a region or a country, $\bar{Y}_w$ is the population share weighted average income of sub-regions ($\bar{Y}_w = \sum_{i=1}^{m} p_i Y_i = Y/N$); where Y is total income of a region or a country and N is total population size of the region or the country, p_i is the population share of sub-region *i* (n_i/N, *where* n_i *is the population size of sub-region i)*, and *m* is the number of sub-regions within a region or a country.

Another commonly used inequality measure for spatial inequality analysis is the Theil Index. In the case of measuring spatial inequality we use the following formula of the Theil T index:

$$T_T = \sum_{i=1}^{N} p_i \left(\frac{y_i}{\bar{Y}_w}\right) ln \left(\frac{y_i}{\bar{Y}_w}\right)$$

Where y_i and p_i denote, income per capita and population share of sub-region i, respectively, and $\bar{Y}_w$ is the population share weighted average income of sub-regions in a country or a region ($\bar{Y}_w = \sum_{i=1}^{m} p_i y_i$).

Although both the Theil Index and the coefficient of variation are widely used in the literature measuring spatial inequality, some authors suggest using the coefficient of variation (e.g. Portnov & Felsenstein, 2005; Lessmann, 2014). For example, according to Lessmann (2014), the advantage of using the population weighed coefficient of variation is that the measure is not sensitive to single extreme values, it is independent of the number and the sizes of spatial units, it is mean-independent, and satisfies the transfer principle. The justification for weighting by regional population share, however, has been challenged in recent work by Gluschenko (2018) and the author suggests using the unweighted CV in estimating regional inequalities. In fact, Gluschenko's work shows that the population weighted inequality indices (i.e the CV, Theil index, and the Gini coefficient) violates the three key inequality axioms: population independence, anonymity, and the transfer principle. In addition, Gluschenko (2018: p.40) shows that the population-weighted inequality indices are only a proxy measure of interpersonal inequality in the whole population of a country, instead of being a measure of regional inequality.

Following Gluschenko (2018) we can use the unweighted coefficient of variation to calculate spatial inequality in South Africa at the province level using municipalities as our spatial units. The formula for calculating the unweighted CV is given as follows:

$$CV_p = \frac{\sqrt{\sum_i^m (Y_i - \bar{Y})^2/m}}{\bar{Y}}$$

Where CV_p is the coefficient of variation estimate for province p, Y_i is per-capita income of municipality i within province p, $\bar{Y}$ is the mean of the municipality per-capita incomes within province p ($\bar{Y} = Y_1 + Y_2 +, \ldots, + Y_m/m$), and *m* is the number of municipalities within

province p. We use data from the 2011 census to estimate spatial income inequality in South Africa.[20] We compare spatial inequality across the nine provinces using municipalities as our observation units. Thus, our data should be at the municipality level. We should have the following information: per-capita income for each municipality (Y_i), the average of the municipality per-capita incomes for each province ($\bar{Y}$), and the number of municipalities within each province (m).[21]

Once we have the data, we can calculate the numerator of the CV using the following commands:

$$gen\ gap\_square= (Y_i - y\_bar)^2 / m$$

$$bysort:\ P\_PROVINCE: egen\ gapsq\_sum=sum(gap\_square)$$

$$gen\ numerator=sqrt(gapsq\_sum)$$

Then the Coefficient of variation for each province is calculated using the following command:

$$gen\ CV=\ numerator\ /y\_bar$$

---

20 We use census data given that most household surveys are not representative at lower geographical units.

21 In our case, given that our data is at the individual level, the per-capita income for each municipality (Y_i) can be calculated using the following commands (the variable perincome indicates per capita individual level income and F00_NR is individuals id, P_MUNIC is municipality id, and P_PROVINCE is province id):

Total population size for each municipality:
bysort P_MUNIC: egen Mun_pop= count (F00_NR)

Total municipality level income:
bysort P_MUNIC:egen Y_{mt}=sum(perincome)

Per capita income of each municipality:
gen y_i=Y_{mt}/Mun_pop

If you have a sampling weight variable, you can also use the asgen command to get the weighed per capita income estimate by municipality as follows:
bys P_MUNIC: asgen yi =perincome, w(weight)

We can also calculate the per capita income of each province as follows:
bysort P_PROVINCE: egen Pov_pop= count (F00_NR)
bysort P_PROVINCE:egen Y_{mp}=sum(perincome)
gen Y_bar=Y_{mp}/Pov_pop

Once we get the per capita income for each municipality and province we can have the data at the municipality level only. Use the following command to drop repeated observations:
sort *P_MUNIC*
drop if *P_MUNIC==P_MUNIC[_n-1]*
Then we can count the number of municipalities within a province using the following command:
gen temp=1
bysort P_PROVINCE:egen m=sum(temp)

Given that the number of spatial units (municipalities) varies across provinces, the maximum values of the CV estimates vary correspondingly, thus making it difficult to compare spatial inequality across provinces. One approach to solve this problem is to standardize the CV value by dividing it by its upper bound, which is given by $\sqrt{m-1}$ (Gluschenko, 2018). After this standardization the CV values range from zero to one. Use the following command to do the standardization:

gen upprbond=sqrt(m -1)

gen CV_stand= CV/upprbond

The table below provides the raw and standardized coefficient of variation estimates by province. Using the standardized CV estimates we can say that spatial inequality is the highest in Mpumalanga province followed by, Northwest and Gauteng, while spatial inequality is the lowest in Western Cape and Northern Cape provinces.

```
. tabstat CV CV_stand m,by(P_PROVINCE)

Summary statistics: mean
  by categories of: P_PROVINCE (Province)
```

P_PROVINCE	CV	CV_stand	m
1. Western cape	.2739636	.0559226	25
2. Eastern cape	.4683907	.075983	39
3. Northern cape	.3409061	.0668572	27
4. Free state	.3271644	.0750567	20
5. Kwazulu-Natal	.6304823	.0891637	51
6. North west	.4707559	.1109582	19
7. Gauteng	.3317466	.1105822	10
8. Mpumalanga	.4741826	.1150062	18
9. Limpopo	.4826842	.0985275	25
Total	.452491	.0853092	31.13675

We can also map estimates of spatial inequalities by province. See appendix C regarding basic instructions on how to do mapping using Stata.

Figure 6.

It should be noted that not all censuses collect information on income. However, most censuses collect information on access to basic services (i.e. electricity, water, sanitation), education level, and household asset ownership. In such cases, we can use these variables to construct an asset index and use this measure instead of income to estimate spatial inequality.

The coefficient of variation is also commonly used in the literature analysing horizontal inequality. Horizontal inequalities are inequalities between well-defined identity groups in a society (Stewart *et al*, 2010). In the case of measuring horizontal inequalities, grouping variables such as race and ethnicity are commonly used. Instead of spatial units, grouping variables such as race and ethnicity are used as the units of analysis and all individuals in a given group are assigned the same level of income (i.e. average income of the group). Stewart et al (2010) proposed the use of the population-weighted Group Coefficient of Variation (GCOV) to measure horizontal inequality. The group-based coefficient of variation is given by the ratio of the standard deviation to the mean where the units of analyses are the groups. Formally, we can write GCOV as:

$$\text{GCOV} = \frac{1}{\bar{\bar{Y}}}\left(\sum_{g=1}^{m} n_g(\bar{Y}_g - \bar{\bar{Y}})^2\right)^{1/2}$$

Where m is the number of groups (racial or ethnic); n_m is the population share of group m; $\bar{\bar{Y}}$ is the overall mean of the income variable; and $\bar{Y}_m$ is the average income for group m. The higher the value of GCOV, the larger the inequality between groups (i.e. higher horizontal inequality).

The simplest way to measure horizontal inequality is to look at the mean or median incomes by population groups. For example, we can compare the mean and median income of the different race groups in South Africa in 2015 as follows:

tabstat pcminc [w=wgt], by(race) s(median mean)

```
Summary for variables: pcminc
     by categories of: race (  19 :Population Group : Section 0.0)
```

race	p50	mean
African/Black	1000	2092.421
Coloured	1410	2733.188
Indian/Asian	3223.333	6968.913
White	5610	12699.97
Total	1221.982	3308.51

From the above table, we can say that both the mean and the median income levels are the lowest for Africans and the highest for Whites. Given that the mean income can be influenced by the presence of extreme values it is a good idea to report the median income and make comparisons based on that. The median income of Whites was 5.6 times higher than that of Africans in 2015. Likewise, the median income of the Indian/ Asian group was 3.2 times higher than that of Africans.

We can also use gender as our grouping variable and compare gender gaps in average income levels as follows:

tabstat pcminc [w=wgt], by(gender) s(median mean)

```
(analytic weights assumed)

Summary for variables: pcminc
     by categories of: gender
```

gender	p50	mean
female	1018.571	2774.455
male	1416.667	3880.493
Total	1220	3306.908

Both the mean and the median income levels are larger for males compared to females with the median income for males being 1.4 times higher than that of females.

The mean and median income ratios are the most straightforward measures of horizontal inequality. However, such measures become less convenient when we have a large number of population groups (e.g. ethnicity). In such cases, it is better to use an index that summarizes horizontal inequality between the various population groups. One such measure is the population-weighted Group Coefficient of Variation (GCOV) which is suggested by Stewart *et al* (2010). Using data from NIDS 2008 and 2015 we use the following Stata commands to calculate GCOV:[22]

bys year:gen gap_square= Nm(Ym–Y_bar)^2*

bys year: egen gapsq_sum=sum(gap_square)

gen numerator=sqrt(gapsq_sum)

gen GCOVr= numerator/Y_bar

tabstat GCOVr,by(year)

In order to calculate horizontal inequity by gender groups, we can follow the same procedure we followed for the race group except that we now replace the race variable by the gender variable. The table below presents horizontal inequality estimates for race and gender population groups. Given the debate about whether to use a popula-

22 We estimate a weighted mean income (Ym_bar) and population share (n_m)of each race group by year as follows:
bys year race: asgen Ym_bar = pcminc, w(wgt)
bysort year race: egen pop_race=count(pid)

Then we can get the population share of each race group:
gen Nm=pop_race/27105 if year==2015
replace Nm=pop_race/18480 if year==2008

Here we are just using the samples to calculate pop shares

Alternatively, you can generate weighted pop estimates for each year follows:
gen temp=1
*by race, sort: egen pop_race = total(weight*temp)*

The overall mean for each year can be calculated as follows:
bys year: asgen Y_bar= pcminc, w(wgt)

Keep only observations at year and race level by creating a year race id:
egen raceyer=group (race year)
sort raceyer
drop if raceyer==raceyer[_n-1]

tion-weighted coefficient of variation in measuring spatial inequality, we report both the population-weighed coefficient of variation and population-unweighted coefficient of variation estimates.[23]

	GCOV for Race		GCOV for Gender	
year	pop. unweighted	pop. weighted	pop. unweighted	pop. weighted
2008	1.46	0.781	0.136	0.135
2015	1.54	0.670	0.167	0.167

The coefficient of variation estimates for race groups suggest a divergent trend. The population weighted coefficient of variation estimates indicate that horizontal inequality declined over time, while the population-unweighted coefficient of variation shows a slightly increasing trend. With regard to gender, horizontal inequality (gender gap) shows an increasing trend over time and the estimates are more or less the same whether we use the weighed or unweighted coefficient of variation.

5. STRUCTURE OF A COUNTRY REPORT

The reader that we have in mind for the Country Reports is an interested and well-trained policy maker or economist who does not have detailed knowledge about the respective country's inequality situation. We list the sections of each report below, along with a simple explanation of what each section should include. This is followed by a list of the minimal requirements that each report should ideally contain.

5.1 Sections for the Country Report

1. Introduction and background:

   1.1. Provide a brief overview of the situation of the country.

   1.2. Provide enough background for the reader to meaningfully interpret the results that will be presented.

   1.3. This includes some broad discussion of the overall economic situation in the country.

   1.4. Relevant information would include demographic information such as population size and growth, life expectancy at birth, population pyramids,

23 When calculating the population unweighted CV, the gap_square variable in the case of grouping by race is calculated using the following formula:
bys year:gen gap_square = (Ym_Y_bar)^2/4

education and literacy levels, and the geographic distribution of the population across rural and urban areas.

1.5. Summary statistics on the macro-economic context would include GDP and GDP growth, GDP per capita, major industrial sectors and their contribution to GDP, and which sectors the majority of the workforce are employed in.

1.6. Highlight the context of inequality.

1.7. Provide some key figures on inequality that have been widely used until the time of the present report.

1.8. Highlight the main reports or reviews on the issue of inequality in order to situate the diagnostic report and highlight its importance.

1.9. Mention the overarching project of which the diagnostic report is part (ACEIR).

2. Review of the policy space:

2.1. Review of the main policies that have been designed to have an impact on inequality.

2.2. Review the main social and/or economic frameworks in the country (national development plans etc.).

2.3. Review relevant or impactful policies/actions (not going into details of policy evaluation).

3. Data:

3.1. Review data that will be used in the report, highlighting the diversity of data sources as key for analysing a cross-sectional issue such as inequality.

3.2. We include information about the sampling framework, the representativity of the data, the sample size, when the data were collected, and the survey organisation.

4. Profiling, analysing and mapping inequality:

Present each set of results either graphically or in table form.

Provide a coherent narrative that explains how we interpret the results.

4.1. Consumption inequality and/or income inequality:

- Gini, Lorenz, Theil
- Trends of inequality
- Decomposition of inequality by sources of income
- Decomposition of inequality by population groups

4.2. Labour market:

- Wage inequality
- Earnings distribution
- Access to labour market
- Dynamics, churning & informality

4.3. Wealth Inequality:

- Asset index

- Land
- Return on financial assets
- Wealth index
- Top 1%, 0.1%, 0.01%

4.4. Social Issues:

- Education (distance to school, net enrolment rates, years of schooling)
- Healthcare (distance to health facilities, anthropometric measures, life expectancy)
- Internet
- Transport
- Water
- Electricity
- Sanitation
- Waste removal
- Housing

4.5. Spatial inequality:

- Mapping and the derivation of multidimensional deprivation indices (MDI), income inequality, coefficients of variation and any other measures relevant to each country

4.6. Perceptions/subjective measures of inequality

4.7. Social Mobility:

- Middle class analysis and vulnerability
- Dynamics across the distribution

5. Recommendations

5.1. Highlights of Section 4.

5.2. Specify priorities in terms of groups and regions/geographical units.

6. The way forward

6.1. Challenges with existing data/techniques

6.2. Prioritization of the data wish list

6.3. Harmonisation of inequality measurements and computation within the region and continent

We are assuming that each country has access to nationally representative household survey data, ideally at the individual level. At the same time, it is possible that different countries have different types of data available, and that some types of analyses are not feasible given the data constraints. In these cases, it would be appropriate to state that the relevant analyses were not possible due to data limitations.

6. CONCLUSION

This Handbook has introduced some of the conceptual issues that a researcher is likely to confront when starting out with a study on inequalities. After determining the scope

and methods of the study, one can set about implementing the relevant analyses and interpreting the results. For our purposes, we are interested in contemporary levels and recent trends in income or consumption inequality, at the individual level, within each of the participating countries.

The major part of the Handbook is focussed on the data requirements, data issues, and how to implement the various estimators. These include estimating and interpreting common inequality measures such as the Gini coefficient, the Palma ratio, the ratio between various percentiles for the income distribution, the Theil coefficients and the Atkinson's coefficients.

The final contribution of this Handbook was to provide a basic structure of what each country report ought to include, including a set of minimal indicators that each report will ideally include. Overall, this should ensure that each report is a high quality research output on its own, as well as ensure the comparability of the reports across the different countries.

7. REFERENCES

Alkire, S., & Foster, J. (2011). Counting and multidimensional poverty measurement. Journal of Public Economics, 95(7), 476-487.

Araar, A., & Duclos, J. Y. (2013). User manual DASP version 2.3. *Distributive Analysis Stata Package, PEP, CIRPÉE and World Bank, Université Laval, June.*

Atkinson, A. B. (1970). On the measurement of inequality. *Journal of economic theory*, 2(3), 244-263.

Banerjee, A. K. (2010). A multidimensional Gini index. *Mathematical Social Sciences*, 60(2), 87-93.

Cobham, A., Schlögl, L., & Sumner, A. (2016). Inequality and the tails: the Palma Proposition and Ratio. *Global Policy*, 7(1), 25-36.

Cowell, F. A. (1985). Measures of distributional change: An axiomatic approach. *The Review of Economic Studies*, 52(1), 135-151.

Cowell, F. A., & Flachaire, E. (2007). Income distribution and inequality measurement: The problem of extreme values. *Journal of Econometrics*, 141(2), 1044-1072.

Doyle, M. W., & Stiglitz, J. E. (2014). Eliminating extreme inequality: A sustainable development goal, 2015–2030. *Ethics & International Affairs*, 28(1), 5-13.

Ezcurra, R., & Rodríguez-Pose, A. (2017). Does ethnic segregation matter for spatial inequality?. *Journal of Economic Geography*, 17(6), 1149-1178.

Fields, G. S. (2003). Accounting for income inequality and its change: A new method, with application to the distribution of earnings in the United States. In Worker well-being and public policy (pp. 1-38). Emerald Group Publishing Limited.

Filmer, D., & Pritchett, L. H. (2001). Estimating wealth effects without expenditure data—or tears: an application to educational enrollments in states of India. Demography, 38(1), 115-132.

Foster, J., Seth, S., Lokshin, M., & Sajaia, Z. (2013). *A unified approach to measuring poverty and inequality: Theory and practice*. The World Bank.

Gluschenko, K. (2018). Measuring regional inequality: to weight or not to weight?. *Spatial Economic Analysis*, 13(1), 36-59.

Jenkins, S. P., & Jantti, M. (2005). *Methods for summarizing and comparing wealth distributions* (No. 2005-05). ISER working paper series.

Leibbrandt, M., Woolard, I., & de Villiers, L. (2009). Methodology: Report on NIDS wave 1. *Technical paper*, 1.

Lessmann, C. (2014). Spatial inequality and development—is there an inverted-U relationship?. *Journal of Development Economics*, 106, 35-51.

McKenzie, D. J. (2005). Measuring inequality with asset indicators. J*ournal of Population Economics*, 18(2), 229-260.

Shorrocks, A. F. (1982). Inequality decomposition by factor components. *Econometrica: Journal of the Econometric Society*, 193-211.

Shorrocks, A. F. (2013). Decomposition procedures for distributional analysis: a unified framework based on the Shapley value. The Journal of Economic Inequality, 11(1), 99-126.

Sen, A. (1973). "On Economic Inequality". New York: Norton.

Sen, A. (1992), Inequality Reexamined, Oxford University Press, Oxford.

Stewart, F., Brown, G., & Mancini, L. (2010). *Monitoring and measuring horizontal inequalities. Centre for Research on Inequality, Human Security and Ethnicity, CRISE Working Paper (4)*.

Williamson, J. G. (1965). Regional inequality and the process of national development: a description of the patterns. *Economic development and cultural change*, 13(4, Part 2), 1-84.

Wittenberg, M. (2017). Measurement of earnings: Comparing South African tax and survey data.

Wittenberg, M., & Leibbrandt, M. (2017). Measuring inequality by asset indices: A general approach with application to South Africa. *Review of Income and Wealth.*

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.

8. APPENDICES

A. Instructions for downloading and installing the DASP package.

You can go to the following website to download the DASP package:

http://dasp.ecn.ulaval.ca/downloadhelp.htm

Follow these steps to install the DASP modules on Stata

1. Unzip the file dasp.zip(this is a file name) in the directory c:/
1. Make sure that you have **c:/dasp/dasp.pkg** or **c:/dasp/ stata.toc**
1. In the Stata command window, type the syntax
2. net from c:/dasp
3. Then type the syntax:

    net install dasp_p1.pkg, force replace
    net install dasp_p2.pkg, force replace
    net install dasp_p3.pkg, force replace
    net install dasp_p4.pkg, force replace

4. Create a folder called "personal" in the ado folder (which should be in the c:/ drive), if it doesn't already exist. Copy graph_header.idlg and profile.do into this folder.
5. Close Stata and then reopen it. Check that the DASP menu shows up under the "User" tab on the top menu bar.

B. A do file for estimating asset indices using the UC PCA (from Martin Wittenberg)

Please copy the following ado file into your do file and run it. Then use the *ucpc* command to estimate asset indices using the UC PCA approach. Please do not forget to cite Martin Wittenberg for writing this ado file.

```
*! version 1.0.0  18dec2013
program ucpc, rclass
        syntax  varlist(numeric min=2) [aw fw] [if] [in] [, GENerate(name)]

        if "`weight'" != "" {
                local wght `"[`weight'`exp']"'
        }
        if "`generate'" !="" {
                confirm new var `generate'
        }

// clean up varlist

        marksample touse
        quietly count if `touse'
        if (r(N) == 0) error 2000
        if (r(N) == 1) error 2001

        // local varlist : list uniq varlist
        foreach v of local varlist {
                quietly summ `v' if `touse', meanonly
                if r(mean) !=0 {
                        local vlist `vlist' `v'
                }
                else {
                        dis as txt "(`v' dropped because of zero mean)"
                }
        }

        if "`vlist'" == "" {
                dis as err "all variables dropped because of zero mean"
                exit 498
```

```
        }

        local varlist `vlist'
        local nvar : list sizeof varlist
        if `nvar' < 2 {
                error 102
        }
        foreach X of varlist `varlist'{
                tempvar temp`X'
                qui summ `X', meanonly
                gen `temp`X''=`X'/r(mean)
                local tempvlist `tempvlist' `temp`X''
                }

// create matrix to be analyzed

        tempname C nobs Ev L Score
        quietly matrix accum `C' = `tempvlist' if `touse' `wght' , nocons
        matrix colnames `C' = `varlist'
        matrix rownames `C' = `varlist'
        local nvar = colsof(`C')
        quietly matrix symeigen `L' `Ev' = `C'
        matrix `Score'=`L'[1...,1]
        matrix colnames `Score' = scores
        matlist `Score'
        if "`generate'"!=""{

                tempvar index
                qui gen double `index' = 0
                forvalues i =1/`nvar' {
                        gettoken v tempvlist: tempvlist
                        quietly replace `index' = `index' + `L'[`i',1]*`v' if `touse'
                }
                qui gen `generate' =`index'
        }
end

exit
```

C. Mapping using Stata

In order to do mapping, we have to get shapefiles of geographical units. For example, in our case (Figure 6 above), we have got province shapfiles for 2011. Then, to do mapping using Stata we need to install two add-ons: *spmap* and shp2dta using the *ssc* install Stata commands as follows:

ssc install spmap

ssc install shp2dta

where *spmap* is the graphing command which turns the raw data into a standard Stata .gph output. And shp2dta is the command that converts the spatial data (stored in a .shp file) into .dta (Stata file) format which can then be used by the spmap command.

Using the province shapefiles (labeled "PR_SA_2011.shp" in our shapefiles) use the shp-2dta command as follows:

shp2dta using PR_SA_2011.shp, database(Province) coordinates(PRcoord) genid(PRid) genc(_c)

The above command produces two datasets: database and coordinate

The database file contains a variable with the primary geographical units, which in our case is a province. We named this database "Province" in the shp2dta command above.

The coordinate file contains the coordinates which make up the boundaries of the spatial units (i.e. provinces). We named this file "PRcoord"

The option genid(PRid) generates a variable named PRid in the Province file, which assigns unique numbers for each geographic unit (i.e. Province). And genc(_c) creates x_y coordinates in the Province file. See what each dataset contains. And make sure that the province names in the Province dataset and the data containing our CV estimates (in our case, stored in the CV_province file) are consistent and both are given the same ID number which is PRid. We then merge the CV_province dataset with the Province dataset as follows:

use CV_province, clear

merge 1:1 PRid using Province

drop _merge

gen CV_standP=CV_stand*100

format CV_standP %4.0f

Then use the spmap command to map estimates of the coefficient of variation as follows:

spmap CV_standP using PRcoord , id(PRid) fcolor(BuRd) ocolor(balck ..) osize(thin ..) ///

legend(position(11)) legtitle("CV %") clmethod(eqint) name(CV,replace) ///

label(data(Province) xcoord(x__c) ycoord(y__c) label(PR_NAME) color(white black))

Please use the help file for the spmap command to see what the various options specified in the spmap command are.