The Effect of Remittances on Household Income Inequality in South Africa

A dissertation submitted in partial fulfillment of the requirements for the award of the degree of Masters of Philosophy in Politics, Philosophy and Economics

by
Janina Hundenborn
HNDINE001

University of Cape Town
Faculty of Humanities
School of Economics

Supervisor: Patrizio Piraino

14.11.2014

Declaration

This work has not been previously submitted in whole, or in part, for the award of any degree. It is my own work. Each significant contribution to, and quotation in, this dissertation from the work, or works, of other people has been attributed, and has been cited and referenced.

Signature: ____________________ Date: ____________________

Contents

1 Introduction 1

2 Literature Review 4
- 2.1 Background on Migration 4
- 2.2 Remittances and Inequality 6

3 Methodology 9
- 3.1 Decomposition of Income Inequality 10
  - 3.1.1 The Decomposition of the Gini Coefficient 10
  - 3.1.2 The Derivation of the Decomposed Gini Coefficient 12
- 3.2 Non-Migration Counterfactual 13
  - 3.2.1 Different Approaches to non-migration Counterfactuals 14
  - 3.2.2 Estimation Method 16
- 3.3 Propensity Score Matching and Difference in Difference Matching Estimators 18
- 3.4 Propensity Score Matching 18
  - 3.4.1 Difference in Difference Matching Estimators or Conditional Difference in Difference 21
  - 3.4.2 Different Matching Algorithms 24

4 Descriptive Statistics 25
- 4.1 Data 26
- 4.2 Data Concerns 26
- 4.3 Summary Statistics 28

5 Results 31
- 5.1 Decomposition of the Gini 32
- 5.2 Results of the Counterfactual Analysis 36
  - 5.2.1 Conditional Difference in Difference Matching 36
  - 5.2.2 Estimation of Predicted Income Excluding Remittances 44

6 Conclusion 50

7 Appendix 57

List of Figures

1 Nearest Neighbour Probit Matching - Rural Households 40
2 Nearest Neighbour Probit Matching - Urban Households 40
3 Kernel Matching - Rural Households 57
4 Kernel Matching - Urban Households 57
5 Nearest Neighbour Logit Matching - Rural Households 62
6 Nearest Neighbour Logit Matching - Urban Households 62

List of Tables

1 Summary Statistics of Household Income Components 29
2 Summary Statistics on Household Characteristics 30
3 Gini Decomposition into Income Components 33
4 Probit vs Logit Estimations 37
5 Results of the t-test after ps matching - Rural Households 41
6 Results of the t-test after ps matching - Urban Households 42
7 Regression for Prediction of Income Excluding Remittances - Urban Households 46
8 Regression for Prediction of Income Excluding Remittances - Rural Households 47
9 Gini coefficients in the Counterfactual Scenario - Estimated using different methods 49
10 t-Test after Kernel Matching - Rural Households 58
11 t-Test after Kernel Matching - Urban Households 59
12 t-Test after Logit Matching - Rural Households 60
13 t-Test after Logit Matching - Urban Households 61

Abstract

Household incomes in developing countries often rely on a variety of sources. Analyzing the effects on income inequality of these different sources can help understand developments underlying overall inequality. South Africa's levels of inequality have been characterized as remaining "stubbornly high"(Leibbrandt and Finn, 2012). Studies show that in the past 20 years, Gini coefficients of per capita income have increased from 0.66 in 1993 to 0.70 in 2008. Lerman and Yitzhaki (1985) derived a method to decompose inequality of income by source followed by a derivation of the Gini Coefficient by Stark et al. (1986). It therefore becomes possible to assess the impact of changes in different components on inequality of total household income. This paper utilizes these techniques to focus on the effect of remittances on inequality in South Africa. Applying the decomposition of income sources to the South African National Income and Dynamics Survey (NIDS), the paper will take the analysis one step further by constructing a counterfactual that allows to compare current inequality levels to levels that would have prevailed had migration not taken place. For the construction of this counterfactual, conditional difference in difference matching will be employed and data on matched non-remittance households will be used to predict household incomes excluding remittances for migration households. The findings of this paper show that levels of inequality are still stagnating. While inequality measured by the Gini coefficient is lacking significant improvement, the counterfactual analysis shows that without remittances, inequality would be slightly worse than current levels. The counterfactual estimation thus supports the result of the decomposition of the Gini coefficient that also finds a minor inequality reducing effect of remittances.

1 Introduction

When a household sends out one or more migrants, it is often observed that these migrants send back personal transfers in cash or kind to their regions of origin. These transfers are referred to as remittances. Remittances have long been a topic of academic studies and it is often discussed whether such payments may have an increasing or decreasing effect on levels of inequality. The impact of remittances is unclear a priori to migration due to the unknown placement of migrants in the initial income distribution. On the one hand, remittances have the potential to be equalizing. In this case, remittances would mostly be flowing towards households at the bottom end of the income distribution, thereby lifting these households and making overall income distribution more equal. On the other hand, remittances could increase inequality. There are two cases where this may happen. Firstly, an increase in inequality will occur if migration costs are relatively high. This can prevent poorer households from sending migrants as they face larger constraints. In such a case, remittances flow towards the already better off and levels of inequality worsen. Secondly, it is possible that poorer households face higher opportunity costs of migration and hence abstain from sending out migrants. Opportunity costs can be too high if the labour of a migrant is needed in the immediate household more than financial assistance in the form of remittances are needed for the household income. In this instance, remittances would again increase inequality levels as they would favour households located at the upper end of the income distribution. Additionally, Stark, Taylor, and Yitzhaki (1986) argue that the effects of remittances on income inequality depend on the size of remittance payments relative to income from other sources. This highlights the complex and versatile effects of remittances.

However, remittances can have numerous other effects on sending households. Docquier and Rapoport (2006) summarize some effects found in the remittance literature, which include positive contributions of remittances on liquidity constraints for example on the educational investment of households (Perotti, 1993) or for entrepreneurship (Mesnard, 2004). Furthermore, remittance payments can be used for consumption smoothing in the receiving household as well as for insurance. Rural households are highly vulnerable to fluctuating incomes as they are more dependent on external factors in order to receive income from agricultural activities. Rural households may experience lower agricultural income due to a drought or other unpredictable factors. A member of the household that migrated to urban areas can assist with consumption smoothing as risks associated to urban incomes are generally uncorrelated with those experienced by rural households (Docquier and Rapoport, 2006). Lastly, Docquier and Rapoport (2006) emphasize the positive effects on growth that are associated with the distributive effects of remittances.

In the South African context the source decomposition of income inequality with regards to remittances is especially interesting. In 1994, the system of apartheid officially ended for South Africa with the democratic election of a new government. The onset of a new democracy ended a regime of racial discrimination that had severely obstructed the freedom of the majority of the South African people especially in the later decades of the 20th century. During the apartheid era, migration laws restrained the free movement of labourers and prohibited permanent migration for a majority of the population (Posel, 2010). With the new democracy came new opportunities and for the period of 1993 to 2002, households increasingly contained one or more migrants that were absent from the household to work or in search for work and most of these households also reported the recipience of remittances (Posel, 2010). However, levels of inequality in overall per capita income have increased in the post-apartheid era. According to Leibbrandt et al. (2010), the Gini coefficient was at around 0.66 in 1993 from which there has been an increase in inequality to a Gini of 0.70 in 2008. A decomposition of these inequality levels will allow to disentangle the a priori uncertain effects of remittances as well as the effects of other income sources on inequality.

For this paper it is especially interesting which factors appear to be the driving force behind the high levels of inequality witnessed in South Africa and how remittances may affect present inequality. In order to offer a more thorough analysis, a counterfactual scenario will be constructed that will predict household incomes without migration. Opposed to previous studies, the counterfactual created in this paper uses propensity score matching in order to avoid selection bias in the estimation of remittance-excluding household income. Propensity score matching can be employed given that a certain set of assumptions holds; it will be shown later in this paper that the assumptions for propensity score matching are satisfied by the data. Therefore, propensity score matching is the best solution for the construction of a counterfactual in this paper. For the counterfactual scenario, it is not possible to simply set income from remittances equal to zero and thus exclude them from the analysis. This approach would ignore how households may compensate missing remittance income with income from other sources. Furthermore, one cannot simply take all non-remittance receiving households to predict the counterfactual income for households that receive remittances as it can be assumed that the two types households differ systemically. Therefore, difference in difference matching is employed to find those households that do not receive remittances but are similar to households that do. The data on these matched households of the non-remittance control group will then be used to predict household income for the households in the remittance group in a non-migration counterfactual. The final step of the counterfactual analysis will then see to calculate the Gini coefficients of the predicted income distribution in order to measure the level of inequality that would have prevailed without migration and without remittances.

The remainder of this paper is structured as follows. In the literature review, Section 2.1 will introduce the topic of migration in academic literature and will give a general understanding of incentives for and effects of migration. Section 2.2 will then review existing literature on remittances and income inequality. Literature on the method and application of the decomposition the Gini coefficient according to income components will be discussed and the Gini coefficient as a measurement of inequality will be reviewed. In Section 3 the methodology of the decomposition of the Gini coefficient as well as the methodology for the creation of a counterfactual will be discussed. Furthermore, different possibilities of creating a counterfactual will be discussed in order to establish why propensity score matching is suitable in this context. Section 3.4 therefore reflects on methods of propensity score matching. These methods will then be applied to the data of the South African National Income and Dynamics Survey (NIDS). Therefore, Section 4 will present the NIDS data set in detail. The results of the analysis will then be discussed in Section 5. Section 5.1 thereby presents the results of the decomposition of the Gini coefficient for all waves of the NIDS data separately. In Section 5.2 propensity score matching will be applied in order to estimate the counterfactual levels of inequality. Finally, Section 6 concludes.

2 Literature Review

The literature on remittances is multi-faceted and broad. In order to comprehend those papers focusing on remittances, a general understanding of the theories of migration is instrumental. For this reason, this section offers a short introduction to migration in the literature before moving on to discuss the literature on remittances and inequality specifically. In the migration literature, many theories are developed with regards to international migration, however most of the concepts discussed here are applicable to internal migration and literature on both concepts will be reviewed.

2.1 Background on Migration

Since Roy (1951) the ground rules for the economics of migration were formulated. Roy's model utilizes an income maximization framework and forms the conceptional ground for a major body of literature on migration. The author conceptualizes the idea that migrants move in prospect of higher incomes to be generated at the migration destination. Together with incomes in the source economy and incomes expected at the receiving destination, the cost of this migration forms the third determinant of who migrates and who stays at home. However, further selection effects are found when studying the characteristics of migrants. Roy (1951) defines three possible selection effects that may occur. Firstly, there may be *positive selection*. This means that migrants earn above average wages in the location of origin as well as at the area of destination, this destination can be abroad or a national higher income destination. In this case, migrants are more likely to emanate from households at the upper end of the income distribution in their source areas and may well outperform colleagues at the receiving destination. However, the latter only holds true assuming no discrimination of migrants due to racial or cultural backgrounds. Whenever positive selection occurs, the possibility for a brain drain to exist increases as well (Borjas, 1991). Literature on the brain drain, which is defined by Beine et al. (2008) as "the international transfer of resources in the form of human capital [which] mainly applies to the migration of relatively highly educated individuals from developing to developed countries", forms another branch of migration literature but will not be considered further in this paper.

The second effect Roy (1951) defines is that there may be *negative selection*. This occurs when migrants are under-performing in their home as well as at their destination's labour market. The phenomena of negative selection is observed if the home income distribution is more unequal than at the destination point. This implies that low-income workers would experience better insurance against such low incomes at the destination than at the origin. Therefore, low ability workers can still improve their situation through migration although they would remain at the bottom end of the income distribution. This case is less applicable to internal migration as any form of insurance against low income would be the same across areas within a country. Therefore internal migration would not occur in this instance.

Lastly, there could be *refugee sorting*. This is the case in which migrants are underperforming at home but outperforming natives in the destination country. This is the case for example when migrants were suppressed at their source location. Although less likely, refugee sorting can occur with internal migration if a certain population group is discriminated against in certain parts of a country but not in others.

Further analysis of these selection effects is given for example in Borjas (1991). Borjas (1991) concludes that migrants are a non-random sample of the population. Selection is based on observable characteristics, such as education, and on non-observable variables such as productivity and ability. On top of this first step of selection, migrants also select themselves into different host regions or countries. During the course of his investigation, Borjas (1991) found that laxer restrictions to migration to the United States led to a less competitive migrant labour force as a higher number of less-skilled individuals were attracted compared to labour markets for example in Australia or Canada, where migrants are sorted more stringently according to skills.

Additional analysis by Grogger and Hanson (2011) finds that aside from positive selection, positive sorting is also a common feature in international migration. Positive sorting indicates that migrants with higher levels of education are more likely to sort themselves to migrate to countries that are characterized by higher rewards to skills, according to Grogger and Hanson (2011). Their findings indicate that in terms of selection effects, given any pair of source-destination countries migrants are relatively more educated compared to non-migrants when absolute wage difference due to skill-related variations between the country of origin and the migration destination are larger. Additionally, the study on sorting effects indicates that the stock of more educated migrants at the migration destination is increasing as the absolute difference in wages between low- and highly skilled workers increases. Grogger and Hanson (2011) discovered these findings in a study on international migration, however, it is plausible that similar holds true for wage differences within a country.

Given these studies on migration in general, the next section will focus on the effects of migrants' transfers to their family members remaining in the home location.

2.2 Remittances and Inequality

Stark, Taylor and Yitzhaki (1986) examine the impact of rural-out migration upon household income distribution in Mexico. The authors find that the effects of remittances on income distribution vary depending on the migration history of the respective villages. In a village with a long migration history, members of a village that migrated at an earlier stage contribute to an increased opportunity for members of poorer households to migrate later on. Stark et al. (1986) identify what later became known as network effects. Such effects ease the access to migration for poorer households. In this case, remittances can have a decreasing effect on inequality. However, in villages with a short or recent migration history, remittances will necessarily be distributed rather unequally. For a village at this stage, migration is described as subject to a higher degree of uncertainty as information is scarce, hence, Stark et al. (1986) identify households from the upper end of the income distribution to be more likely to migrate. However, the effect of remittances is also dependent on their relative magnitude in comparison to income from other sources. In this matter, the paper by Stark et al. (1986) holds valuable information due to the fact that their results are found whilst studying two Mexican villages with different characteristics regarding remittances and migration history. Therefore their analysis allows to better compare the effects of internal and international migration with each other.

In their analysis, Stark et al. (1986) implement a new statistical concept to identify the role of remittances as well as other income components in the distribution of village incomes. To complete such an undertaking, they use the decomposed Gini coefficient introduced by Lerman and Yitzhaki (1985) which identifies the role of different sources of income in overall income inequality. Using such a Gini index to measure the effect of remittances on the income distribution, the role of remittances can be decomposed into three parts. Firstly, the share of remittances in overall household income; secondly, the inequality in remittance payments and lastly, the (Gini) correlation between remittances and total household income. The specifics of this method will be applied and discussed later in this paper (see Section 3.1). In their article, Stark et al. (1986) find that especially in more costly international migration, a village with a shorter migration history shows significantly larger inequality in remittances. Furthermore, migrants are predominantly sent from households at the upper end of the income distribution. However, compared to international remittances, internal migrants' remittances report a lower degree of inequality. In a village with a longer migration history, remittances from international migrants are more equal than payments from internal migrants. Therefore, this effect is opposed to the situation of a village with a short migration history. As described above, it is plausible that due to the long migration history, households at the bottom end of the income distribution are able to send out migrants as well. For both villages, the correlation between remittances and total household income is rather small.

This paper follows the approach of Stark et al. (1986) and thus uses the Gini coefficient as a measurement of inequality. There are a number of different measurements for inequality such as the Theil index, the Atkinson index or other indeces; each one has its own advantages and disadvantages. So does the Gini coefficient, although it is by far the most commonly used measurement of inequality according to DeMaio (2007). The Gini coefficient measures the space between the Lorenz curve and the 45° line of equality in a graphic that depicts the cumulative percentage of population and income. The Gini coefficient fails to reflect different shapes of potentially intersecting Lorenz curves that both have the same Gini coefficient but not the same income distribution (Cowell, 2000). Furthermore, the Gini coefficient is strongly sensitive to inequality at the centre of the income distribution. For this paper this is a serious constraint as this may not be where most remittance receiving households are located. Because of its sensitivity to inequality in the middle of the income distribution, the Gini coefficient is not value free. However, the main reason for its popularity is that the Gini coefficient reflects the income distribution as one single summary statistic.

In a more recent study on international migration, McKenzie and Rapoport (2007) analyze the effects that Stark et al. (1986) found. McKenzie and Rapoport (2007) aim to explain the network effects of migrants as well as the diverse consequences of inequality and migration. The authors provide convincing theory that mainly middle-income families are sending out migrants. Families with higher incomes lack the incentives to migrate and low-income families lack the means to do so. Furthermore, they conclude that remittances have a decreasing effect on inequality only in those villages with a long history of migration. In such villages, poorer households are also able to send out migrants as they face lower migration costs due to existing networks in the receiving region or country. Overall, there seems to be Kuznet-shaped relationship[1] among communities with a wider range of migration experiences.

Leibbrandt, Woolard, Finn and Argent (2010) apply exactly those methods introduced by Stark et al. (1986) to South African data. They utilize three consistent sets of national survey data in order to assess developments of inequality over time. The data sets utilized are the survey undertaken in the 1993 Project for Statistics on Living Standards Development (PSLSD), the Income and Expenditure Survey of 2000 and the first wave of the National Income Dynamics Study (NIDS) in 2008. According to their findings remittances have a rather small however possibly negative role on inequality, implying that remittances may serve to lower inequality. They also find strong fluctuations in the share of households receiving remittances. This share was equal to 24% in 1993, rose to 36% in 2000 only to fall back to 14% in 2008. Most importantly, the study finds an increase in overall inequality over the time period under observation.

Although they introduced a novel method of analyzing income inequality according to components, the paper by Stark et al. (1986) has a very strong short-coming, namely, the lack of a counterfactual analysis to determine developments of the village income distributions without remittances. Adams (1989) introduces a non-remittance counterfactual analysis in a paper on remittances of workers and their effects on inequality in rural Egypt.

Adams (1989) characterizes the inability to predict migration and income functions in order to be able to compare pre- and post-migration income data as one of the reasons why the literature has struggled to determine whether remittances increase inequality or not. He therefore develops a counterfactual scenario using household survey data from a rural area in Egypt. Adams (1989) uses three equations for his analysis that determine (1) who is migrating, (2) how remittances affect pre-migration income and (3) remittances' effect on post-migration income. The estimation equations of pre- and post-migration income thereby only differ by a migration dummy being added to the income function excluding remittances. Adams uses data available from the households that are not receiving remittances to estimate a set of parameters predicting income excluding remittances and then applies these parameters to all households in order to generate a non-migration counterfactual. He also predicts household incomes for the migration counterfactual by applying a similar method to households receiving remittances, predicting parameters and then estimating gross household income including remittances for all households (see Adams, 1991). However this estimation procedure is vulnerable to a possible bias between migrant and non-migrant families. Selection bias has long been discussed and established in migration literature, hence the veracity of income predictions not taking this bias into account will be questionable. In Section 3.2.1, a method to account for this bias will be discussed in more detail.

Acosta et al. (2008) use a very similar approach to the one of Barham and Boucher (1998) described in Section 3.2.1. The authors firstly follow the methodology introduced by Stark et al. (1986) to uncover the link existing between remittances and inequality. For this purpose, they use data from household surveys of ten Latin American countries and decompose the Gini coefficient according to income components for all ten countries. As mentioned before, the authors again note that this approach implies remittances being exogenous transfers by migrants. They point out that the implicit counterfactual scenario would mean that there are changes in remittances but no changes in migration. However, following their estimation of household income without remittances Acosta et al. (2008) highlight that one has to control for the possibility of a selection bias. Similar to Barham and Boucher (1998), they add a "propensity to migrate"in the Heckman (1979) two-step estimation framework and simulate an error component in order to account for the possible bias. They do not however include a separate estimation for labour force participation as previously explained by Barham and Boucher (1998). Their results show that unlike in the studies of Adams (1989) or Barham and Boucher (1998), the selection coefficient is significant and positive. Overall, remittances show an inequality reducing effect in the Latin American countries under observation. However, the reducing effects of remittances on inequality appear small, even after applying the procedure of imputing home earnings for migrants. Additionally to the effects of remittances on inequality, the authors also test for the effects of these transfers on poverty. Again, remittances have a decreasing effect on poverty levels, however the effect is rather small. Furthermore, the results differ considerably across countries which makes a common conclusion difficult. All in all, Acosta et al. (2008) offer a substantial analysis of the effects of remittances on different aspects relevant for developing countries. Section 3 will discuss the assumptions relevant for propensity score matching in more detail as well as the overall methodology used in this paper. This includes the decomposition of the Gini coefficient as well as detailed discussion of propensity score matching and matching algorithms. Furthermore, Section 4 will discuss data concerns relevant for the Gini decompositions as well as the construction of a non-remittance counterfactual using propensity score matching.

3 Methodology

Different concepts of analyzing the role of remittances have been discussed in Section 2.2. This section focuses on the methodology implemented in this paper. Firstly, the decomposition of the Gini coefficient will be discussed following the Lerman and Yitzhaki (1985) approach that was further developed by Stark et al. (1986). Furthermore, Adams' approach (1989) of a counterfactual will be explained and further advanced by implementing propensity score matching in the choice of non-remittance households that will be used in the estimation of the counterfactual.

3.1 Decomposition of Income Inequality

In this subsection, the analytical framework introduced by Stark et al. (1986) and Lerman and Yitzhaki (1985) will be presented. The index presented here is a Gini coefficient decomposed into the different sources of income. The decomposition of the Gini coefficient is instrumental to analyze the role of remittances in more depth but also to gain a deeper understanding of the underlying factors of South Africa's persistently high level of inequality. It points out the dominant sources of income in the overall income distribution in South Africa. By decomposing the Gini coefficient it is possible to focus on the following crucial concepts. Firstly, the share of remittances in total household income can be analyzed. Furthermore, decomposing the Gini makes it possible to determine the distribution of remittances (and other income sources), and finally, the correlation of remittances with overall household income can be examined. This correlation reflects how a change in remittances may affect total household income and the distribution of overall income.

3.1.1 The Decomposition of the Gini Coefficient

In this paper, the Gini coefficient of per capita household income in South Africa will be decomposed into K components of household income. Let $y_1, y_2, \ldots, y_K$ denote the different income components. For the remainder of this analysis, the components y_k are assumed to be remittances but also include wage income and income from non-employment sources such as government grants, agricultural income and income of a capital nature. Overall household income, i.e. the sum of the different components y_k, is then given by y_0 and can be formalized by $y_0 = \sum_{k=1}^{K} y_k$. As in Stark et al. (1986), the level of inequality for total household income, the Gini coefficient of y_0, is then assumed to be given by

$$G_0 = \frac{2Cov\left[y_0, F(y_0)\right]}{\mu_0}. \tag{1}$$

In this formulation, G_0 represents the Gini coefficient of total household incomes. In equation (1), μ_0 denotes the mean of household incomes and $F(y_0)$ denotes the cumulative distribution function of total household income y_0. Following Stark et al. (1986), equation (1) can be rewritten as

$$G_0 = \frac{2\sum_{k=1}^{K} Cov\left[y_k, F(y_0)\right]}{\mu_0}, \tag{2}$$

given that $y_0 = \sum_{k=1}^{K} y_k$ and utilizing the properties of covariance. In equation (2), $Cov\left[y_k, F(y_0)\right]$ is the covariance between income source k and the cumulative distribution of income, $F(y_0)$. This redefinition of the Gini coefficient for overall household income (G_0) allows for the final step of the decomposition. Assume that S_k is the share of income from component k in total household income. S_k may then be formalized by $S_k = \frac{y_k}{y_0}$. Furthermore, G_k is the corresponding Gini coefficient that measures the level of inequality in income component k. As such, G_k can be formalized by rewriting equation (1) for y_k. As previously discussed, the decomposition of the Gini coefficient allows to analyze three important concepts. Such concepts include the share S_k of income from source k in total household income as well as k's distribution. The final concept mentioned previously is the correlation of remittances or other incomes y_k with overall household income. This correlation is given by

$$R_k = \frac{Cov\left[y_k, F(y_0)\right]}{Cov\left[y_k, F(y_k)\right]}.$$

In this equation, R_k represents the Gini correlation of income from component k with overall income y_0.

Given the definition of S_k, R_k as well as the Gini G_k of income component k stated above, equation (2) can be rewritten by the following steps. When dividing equation (2) by overall income y_0, one obtains the relative Gini. This is then multiplied and divided by both the covariance of k with the cumulative distribution of k, i.e. $Cov(y_k, F_k)$ as well as y_k, the income from component k. Then the following equation is obtained

$$G_0 = \underbrace{\frac{2\sum_{k=1}^{K} Cov\left[y_k, F(y_0)\right]}{\mu_0 \cdot y_0}}_{\text{Relative Gini}} \cdot \frac{Cov(y_k, F_k)}{Cov(y_k, F_k)} \cdot \frac{y_k}{y_k}.$$

Given this equation, S_k, G_k and R_k can be substituted such that the overall Gini G_0 can be rewritten as

$$\begin{aligned} G_0 &= \sum_{k=1}^{K} \frac{Cov\left[y_k, F(y_0)\right]}{Cov\left[y_k, F(y_k)\right]} \cdot \frac{2Cov\left[y_k, F(y_k)\right]}{\mu_k} \cdot \frac{y_k}{y_0} \\ &= \sum_{k=1}^{K} R_k \cdot G_k \cdot S_k. \end{aligned} \tag{3}$$

Hence, equation (3) yields the source decomposition of the Gini coefficient. Stark et al. (1986) assume that for any given income source k in equation (3), the larger the product of the three components, S_k, G_k and R_k, the larger is the contribution of k to the overall level of income inequality G_0. By definition, the components S_k and G_k are always positive but smaller than 1. The properties of R_k, however, are a mixture of the properties of the Pearson's and Spearman's correlation coefficient. Stark et al. (1986) summarize the properties of the correlation between income source k and total household income y_0 as follows:

1. $-1 \leq R_k \leq 1$. R_k is equal to zero if y_k and y_0 are uncorrelated. R_k is equal to 1 if y_k is an increasing function of total household income and -1 if it is a decreasing function.

2. If y_k and y_0 are normally distributed, then R_k will be equal to the Pearson's correlation coefficient ρ.

These characteristics allow the decomposed Gini to determine the three previously described terms of the role of remittances. Firstly, the share of remittances in overall household income; secondly, the inequality in remittance payments and lastly, the (Gini) correlation between remittances and total household income.

3.1.2 The Derivation of the Decomposed Gini Coefficient

The previous subsection has successfully shown that the decomposition of the Gini coefficient presents a derivation that allows for an analysis of how remittances influence the overall income inequality in different ways. This includes the share of remittances as one income source in the total household income, the inequality in the distribution of remittances as well as the correlation of remittances and overall household income. However, the decomposed Gini as presented in equation (3) offers the opportunity for even more detailed analysis. As such, the formula of the decomposed Gini can be used to examine the effects of a marginal change in one income component of the total distribution of income. This section will carefully explain how the effect of a 1-percentage change in any income component on the overall Gini coefficient G_0 is calculated.

Assume that results of household labour and production decisions are held constant. If then an exogenous change in any income component j by a factor e occurs, income from j is assumed to change according to $y_j(e) = (1+e)y_j$. The first derivative of the decomposed Gini is then stated by Stark et al. (1986) as

$$\frac{\partial G_0}{\partial e} = S_j(R_j \cdot G_j - G_0) \tag{4}$$

where equation (3) is derived with respect to the small percentage change e in income from a particular source j. In equation (4), S_j is the respective jth income share, G_j is the Gini coefficient of income source j and G_0 the overall Gini coefficient. Additionally,

R_j represents the correlation between income from source j with total income before the marginal change. Equation (4) is then divided by G_0 and the following formula is obtained

$$\frac{\partial G_0/\partial e}{G_0} = \frac{S_j \cdot R_j \cdot G_j}{G_0} - S_j. \tag{5}$$

Equation (5) reflects how a small exogenous change in one income component j affects the overall Gini coefficient G_0. As can be seen from equation (5), the relative effect of a marginal change e in income component j on overall inequality equals the relative contribution of that component j to total inequality, i.e. $S_j \cdot R_j \cdot G_j/G_0$, less the share S_j of income from component j.

Given equation (5), the effects of income through remittances can be divided into two categories. Assume income component j represents income from remittances. On the one hand, if R_j states a negative or zero correlation between remittances and total household income, any increase in remittances will necessarily lower inequality due to the fact that any share S_j of income from remittances as well as the Gini indices for remittances and total income, G_j and G_0, are always positive. On the other hand, if R_j is positive, the effect of remittances on inequality depends on the sign of $R_j \cdot G_j - G_0$ in equation (4). As pointed out above, the Gini correlation coefficient R_j is always smaller or equal to 1 (or -1 if not assumed to be positive). Hence, inequality in remittances must be higher than inequality of total income for overall inequality to increase, i.e. $G_j > G_0$ forms a necessary condition given a positive Gini correlation between remittances and total income.

This section has tried to stress how the decomposition of the Gini coefficient and its derivation can offer critical information on the effects of different income sources on overall income inequality. Therefore, section 5.1 will apply this decomposition to South African data. The effects of remittances on income inequality measured by the Gini coefficient will be analyzed. Furthermore, the effect of a marginal change in remittances on total inequality will be discussed. However, it is possible to take the analysis of the Gini coefficient in South Africa even further. For this purpose, this paper creates a counterfactual where income from remittances ceases. The next section will explain how this counterfactual is created in more detail.

3.2 Non-Migration Counterfactual

Borrowing from Adams (1989), the effects of remittances on household incomes and income distribution will be analyzed. The methodology used by Adams (1989) and its shortcomings have been raised in the literature review in Section 2.2. This subsection describes the methodology developed in this paper to reduce bias associated with selection when creating a non-remittance counterfactual. The main goal is to predict household income in case there would be no remittances. For this purpose, Adams (1989) used the non-migration households to estimate predicted household income excluding remittances, denoted as PREX. Subsequently, Adams (1989) applied the regression parameters obtained to all households, thus predicting overall household income without remittances. However, this approach ignores the systematic differences between migration and non-migration households, often referred to as selection bias. Therefore, this section carefully develops a method that will account for the systematic differences between migration households and non-migration households. Nevertheless, it is important to notice that the method described in this section and applied in later parts of this paper is not the only possible way to correct for time invariant unobservables in estimations regarding a non-migration counterfactual.

3.2.1 Different Approaches to non-migration Counterfactuals

Barham and Boucher (1998) analyse the overall effect of migration and remittances on the distribution of incomes in a town in Nicaragua. Barham and Boucher (1998) stress the differences in the role of remittances depending on how these payments are viewed. Income from remittances can be treated as exogenous transfers or they can be treated as a substitute for earnings in the receiving household. The authors determine the latter definition as more interesting since the economic question analyzed becomes one that will have to create a counterfactual that imputes for home earnings of migrants had they not migrated. In order to achieve this, the earnings equation is estimated using a double-selection Heckman model (Heckman, 1979). The first selection rule is thereby given by the choice of migration and the second by the choice of participation in the labour force of non-migrants.

In their paper, Barham and Boucher (1998) decompose the Gini according to remittance and non-remittance sources briefly, treating remittances as an exogenous transfer. Applying the method introduced by Stark et al. (1986) they find that remittances appear to have a lowering impact on inequality in the income distribution observed. More importantly, they develop an econometric model to estimate home earnings of migrants using a double selection model on labour market participation and migration. The authors find that participation rates in the village observed are very low and conclude that migrants may not be randomly drawn from the population. Instead migrants may have self selected due to higher levels of motivation present in individuals who are willing to migrate but not in those that stay in the village under observation. This calls for measures to correct for the bias induced by systematically different subgroups when estimating migrants' home earnings. Covariates used broadly fit into two categories, human capital characteristics, such as level of education and age, as well as variables to characterize the family structure. This stems from the conception that better educated individuals are more likely to migrate and that the head of household for example is less likely to migrate.

In the first step of their estimation, Barham and Boucher (1998) find that the propensity to migrate as well as the probability to participate in the labour force increase with age at first and then decrease. Furthermore, a lower level of education seems to lead to a lower likelihood of that individual to participate in the labour force and a lower propensity to migrate. Finally, men originating from middle income households seem to be most probable to migrate. The results of this bivariate probit estimation using maximum likelihood on the labour force participation and the migration decision are then employed in the estimation of the earnings equations. For this estimation, they include two selection coefficients to correct for a possible bias due to non-random selection and compare it to the estimation without the correction. The selection coefficients are constructed for each individual using the parameter estimates of the bivariate probit estimation and are then included in the OLS estimation of the (selection corrected) earnings equation. The results of this regression show that for the village under observation the selection coefficients appear small and statistically insignificant. This leads to the conclusion that in this case, the subgroup of non-migrant participants of the labour force is selected randomly from the overall population. This is opposed to their former assumption of non-random selection of migrants.

In their further analysis, the authors use their previous estimations to impute individual earnings in order to construct the Gini coefficient of the non-migration counterfactual. Following through with their analysis, the authors construct different non-migration counterfactual Gini coefficients. They find that when following the Lerman and Yitzhaki (1985) approach and treating remittances as an exogenous transfer, these transfers have a decreasing effect on inequality. However, in their counterfactual estimation, treating remittances as substitutes for home earnings, inequality actually rises by the cause of remittances. In summation, potential earnings of migrants had they stayed home would have a more positive effect regarding equality in income distributions than remittances do.

However, it will be shown later in this paper that the double-selection Heckman model is inapplicable in the case under observation as the restrictions of the model cannot be met by the data. Acosta et al. (2008) also implement a selection-corrected Heckman (1979) estimation framework that includes a 'propensity to migrate' but does not perform propensity score matching as suggested in the method developed in the remainder of this section.

3.2.2 Estimation Method

In his paper, Adams (1989) sets out to determine what the income of migration households would look like if migrants had not migrated and would still be part of the receiving households. As this information is unobservable at the time when data on remittances is collected, necessity arises to predict what the non-remittance income of households would look like. For this purpose, Adams (1989) estimated overall household income without remittances from those households that had never received remittances. Through this method, he obtained parameters to predict income excluding remittances (PREX) for all households. In this paper, Adams estimation is slightly modified to suit the South African context. Therefore, the equation for predicted household income excluding remittances can be formalized as:

$$PREX = HH\_ASSET + EDUC + HHSIZE + GEOTYPE + HHRACE \quad (6)$$

Equation (6) represents the identity of the estimation equation. It was estimated using ordinary least squared (OLS) method. Covariates used include HH_ASSET which is the index of all assets owned by the household and $EDUC$ which represents the mean level of education of the household. The variable HH_ASSET is assumed to have a positive effect on household income. It captures household assets such as property in form of a house or dwelling as well as movable assets such as vehicles for private or commercial use and the ownership of any form of livestock including cattle, sheep, chicken etc. The mean level of education in the household is captured in years of schooling whereby post-school education such as academic degrees and certificates are counted as additional years of schooling. It can be assumed that higher educational levels would yield higher incomes, hence $EDUC$ is assumed to have a positive effect in the estimation of $PREX$. $HHSIZE$ is the number of residents in the household and $GEOTYPE$ determines whether the household is located in a rural or urban area. Household size can potentially have a positive effect on household income, however, the positive effect may be dampened by the fact that income has to be spread across more people. Households located in different geographical regions face different labour markets and as such different opportunities for employment as well as training; different skills are needed in rural areas compared to urban regions and schooling opportunities differ across regions. It is therefore necessary to include a variable $GEOTYPE$ that captures the effects of rural versus urban areas. Finally, $HHRACE$ is included to capture possible effects of race on the household income. It is a wild card for a set of racial dummies that will be included in the estimation for $PREX$ in order to capture possible effects of different races on the household income. Categorical variables for each race are included because the association with different ethnic groups has a historically important role in the South African context. Given the suppression of non-whites, especially Africans, under the apartheid regime, the role of race in the income distribution has to be assessed. If there is not equal opportunity for all, race could possibly be correlated with PREX and would be an error term that leads to biased results when not accounted for.

In his estimations, Adams assumes that migrant and non-migrant households only differ in the fact that the one receives remittances and the other one does not (Adams, 1991). In his rather strict assumption, Adams (1991) states that "non-migrant and migrant households are not assumed to differ in any entrepreneurial or any other sense that might affect their income in any manner apart from the relationships captured by the variables used in the predicted income equations". Whilst this assumption simplifies the model in a sense that makes predicted income calculations much easier it runs the risk of simplifying it to a degree where the model is not suitable to make reasonable predictions. This is especially a concern given that time invariant differences between migrants and non-migrants have often been established in the literature. Furthermore, Adams fails to further justify his assumption.

For the reasons outlined above, this paper extends Adams methodology with the purpose of correctly identifying non-remittance households that are similar to remittance receiving households through propensity score matching. Propensity score matching is the best alternative to construct a counterfactual given the structure of the data on remittances. Other counterfactual estimations discussed in the beginning of this Section, such as Barham and Boucher (1998), use a double selection model to account for selection bias. However, the exclusion restriction for double selection cannot be met given the limited information available on remittance payers in the data. This will be discussed in more detail together with other concerns regarding the data used in Section 4.2.

The counterfactual scenario estimated in this paper therefore uses information available on households to match non-remittance households to remittance receiving households based on the propensity score. Then equation (6) will be estimated for the matched households in the control group. In a subsequent step, parameters obtained from this first estimation will be applied to households that receive remittances in order to predict their income excluding remittances. Therefore, equation (6) will be run again on remittance households but with the parameters obtained from the estimation of PREX from matched non-remittances households. Through this process, income excluding remittances is available not only for non-remittance households but also an estimated income without remittances is obtained for remittance households. In a final step of this counterfactual estimation, the Gini coefficient will be calculated for this newly obtained income distribution.

3.3 Propensity Score Matching and Difference in Difference Matching Estimators

Propensity score matching is a widely used method to evaluate the effects of treatments on participants compared to non-participants. The term 'treatment' thereby refers to different economic policies and events. For example, Girma and Görg (2007) test the effect of an acquisition of a company by a foreign multi-national enterprise using matching estimators and Heckman et al. (1998) test their hypothesis on an experiment of a job training program. In this paper, the treatment under observation is the reception of remittances. Hence, households in the treatment group are those receiving remittances and households in the control group are those that do not.

The principal idea of propensity score matching is to identify those individuals within the group of non-participants that are similar to the individuals in the treatment group. The criteria by which this similarity between the individuals in the two groups is determined are a set X of relevant pre-treatment characteristics. This process of establishing an adequate control group is then used to ascertain the effect that can be solely assigned to the treatment by comparing the differences in outcomes between the matched individuals of the control and the treatment group. Rosenbaum and Rubin (1983) define the propensity score as the predicted probability for participation given the set of observable covariates X. Individuals are matched when they have the same propensity score. Detailed discussion of the assumptions and the theoretical background of propensity score matching in general will follow in Section 3.4. Section 3.4.1 will then discuss Difference in Difference matching specifically before Section 3.4.2 reviews different matching algorithms applied in this paper.

3.4 Propensity Score Matching

Much of the literature on propensity matching focuses on the removal of biases and some first literature on the issue of propensity score matching is by Rosenbaum and Rubin (1983). In their paper they discuss the instrumental role of propensity scores for causal effects and focus in their analysis on observational studies. Rosenbaum and Rubin (1983) offer the technical background of an adjustment for the propensity score which is suitable for removing biases in estimation strategies. Rosenbaum and Rubin (1983) also provide the fundamental steps to develop the t-test analytics applied later in this paper.

Heckman, Ichimura, Smith and Todd (1998) focus in their study more precisely on a characterization of selection bias. The authors utilize data that stems from an experiment on a prototypical job training program. The authors can insure an "unusual richness"of the data by combining data from the experiment with data from non-experimental comparison group. Heckman et al. (1998) succeed in proving the effectiveness of propensity score matching in reducing conventional measures of bias. However, they fail to completely remove bias from the estimations. The authors discuss the suitability of difference in difference matching to remove the selection bias. They find that, if all assumptions are satisfied by the data, bias can be reduced but not all bias can be removed.

In their paper on propensity score matching, Caliendo and Kopeining (2008) offer an inclusive summary of the technical concerns regarding propensity score matching. They not only discuss different matching methods but also offer in depth analysis on the model choice and how the variables for matching should be chosen. In their paper they go beyond discussing the requirements for valid propensity score matching but offer theoretical background on alternative matching methods if one assumption is not satisfied by the data.

While these papers offer the necessary theoretical background on propensity score matching, Ham, Li and Reagan (2005) actually implement propensity score matching to estimate the effect of internal migration in the United States on the growth of real wages between first and second jobs of movers. The authors apply propensity score matching in order to deal with selection issues and to further examine how migration affects the growth in wages of young male migrants. In their study, the authors not only instrumentalize propensity scores but also implement a distance-base measure for migration to estimate the effects of migration on wage growth. Ham et al. (2005) conclude that in their case they are able to eliminate most of the permanent component of error in the outcome estimation through the process of matching and differencing. Furthermore, the authors find that migration positively affects wage growth for college graduates, however, migration appears to have a marginally negative yet significant effect for high-school drop-outs.

Finally, Smith and Todd (2005) assess the ability of different evaluation methods of social programmes to overcome the problems of non-experimental estimations. The authors re-estimate the work done by LaLonde (1986) as well as by Dehejia and Whaba (2002) focussing on the implementation of cross-sectional as well as longitudinal variations of propensity score matching estimators. Their findings show that earlier estimations by LaLonde (1986) and Dehejia and Whaba (2002) respectively are strongly dependent on the choice of subsamples in the data. Furthermore the results show strong sensitivity to the variable choice for the estimation of the propensity score.

Additionally to this assessment, Smith and Todd (2005) also offer a good derivation of the methodology of propensity score matching in general and difference in difference matching in particular. In their estimations, Smith and Todd (2005) find that difference in difference matching estimators deliver better results than cross-sectional matching algorithms. The authors also find that no such conclusion can be drawn with regards to different matching algorithms. The results do not differ much between nearest neighbour or local linear matching but are very much dependent on the data available, therefore the data under observation should determine the choice of an evaluation method. Smith and Todd (2005) conclude that propensity score matching is not the best estimator in every case but can be suitable to reduce bias given that the assumptions are met by the available data set.

The purpose of using propensity score matching in this paper is not to determine the treatment effect of remittances but to identify a suitable group that can be used in the prediction of income excluding remittances for remittance households. Therefore, the remainder of this section will focus on the matching process instead of the evaluation of the treatment effect. For this purpose, the principals and assumptions as well as the different methods of propensity score matching that will be used in the later analysis are discussed in the following paragraphs.

Propensity score matching requires two rather strict assumptions:

1. Unconfoundedness

2. Overlap

According to Caliendo and Kopeinig (2008), *Unconfoundedness* can be formalized as $Y(0), Y(1) \coprod D|X$ where $\coprod$ denotes independence and X is a set of observable covariates and unaffected by the treatment. D is the treatment indicator, in this case a dummy variable for the receipt of remittances. D is equal to 1 if a household receives remittances and equal to 0 if not. The unconfoundedness assumptions requires that conditional on X potential outcomes Y are independent of the assignment to treatment group. However, this implies the very strong assumption that all variables that simultaneously influence outcome as well as assignment can be observed by the researcher. Especially with migration, however, this may not be the case. As discussed earlier, Borjas (1991) found that migrants may be characterized by a higher level of motivation or a higher willingness to take on a risk. Such characteristics cannot be observed and are therefore not captured in the data. Conditional difference in difference (DID) matching offers an alternative estimation method for which this assumption does not have to be satisfied. The details about this method will be discussed shortly.

The second assumption that has to hold in order to justify propensity score matching is the *overlap* or *common support* assumption. Formally written as $0 < P(D = 1|X) < 1$, the overlap condition states that individuals with the same values of covariates X are characterized by a positive probability of being in the treatment group or in the control group.

Together, assumptions 1 and 2 are called 'strong ignorability' (Rosenbaum and Rubin, 1983). However, Caliendo and Kopeining (2008) point out that the assumptions for strong ignorability are overly strong and that what is commonly known as 'CIA', conditional mean independence assumption, is sufficient to estimate average treatment effect as well as average effect of the treatment on the treated. Conditional mean independence requires that the outcome is independent of the treatment conditional on the propensity score. Although this is a less strict requirement than strong ignorability, in the presence of a selection bias this assumption does not hold either. Hence, the following section will discuss conditional difference in difference estimations as an alternative that can be used when unconfoundedness is not satisfied.

3.4.1 Difference in Difference Matching Estimators or Conditional Difference in Difference

Heckman et al. (1998) are the first to suggest difference in difference matching. It has the great advantage that the assumptions required by this conditional version are weaker than those required by propensity score matching. Girma and Görg (2007) point out that common matching estimators are often dissatisfactory. This is for example caused by strong limitations due to assuming that all characteristics that need to be controlled for are observable. Therefore, this paper utilizes conditional difference in difference matching in order to correct for a selection bias in the estimation of a non-remittance counterfactual. Heckman et al. (1998) discuss the effectiveness of difference in difference matching and find it useful in the case where a time invariant bias is present that can be differenced out. It can be assumed that the differences in characteristics of migration households are constant over time. Therefore it is a suitable approach to take in the estimation of a non-remittance counterfactual. Smith and Todd (2005) find that difference in difference matching estimators are more robust than traditional cross-section matching estimators. Difference in difference (DID) matching estimators or conditional DID matching can be implemented easily in a panel or repeated cross-sectional data set. Therefore, the conditional DID approach will be used to determine which non-remittance households are similar to the ones receiving remittances. These matched households are the ones that will be applied to Adams' approach of predicting household income excluding remittances in the hope that selection bias in the estimation will be reduced in the results.

A DID matching estimator requires weaker assumptions than the matching estimators explained above. The assumption of conditional mean independence is relaxed such that it allows for unobservable and time invariant differences across groups of participants and non-participants. However, the assumption of common support still has to hold. Conditional DID matching was introduced by Heckman et al. (1998). Traditional difference in difference estimations and the related fixed effects estimations are commonly used measures which this approach extends by conditioning outcomes on the propensity score. Furthermore, differences are constructed using semi-parametric methods.

The matching process of conditional DID matching estimators can be explained as follows. If $REM_{it} \in 0, 1$ is an indicator of whether or not household i is receiving remittances in time period t, then y^1_{it+s} denotes household income in period $t + s, s \geq 0$ including remittances and y^0_{it+s} denotes the household income in households that do not receive remittances. The causal effect of remittances can then be derived by $y^1_{it+s} - y^0_{it+s}$. The aim of the matching process is to estimate the counterfactual household income y^0_{it+s} for households that received remittances in $t + s$, however for these households we can only observe y^1_{it+s}. According to Girma and Görg (2007) this situation may be described as a missing data problem. Alternatively to propensity score matching, the income of households had they not received remittances could be estimated by the average of the income of non-remittance households. However, this would only be a valid estimation if there were no other simultaneous effects that are associated with the receipt of remittance income that are not controlled for. As explained above, it can be assumed that remittance receiving households differ systemically. Therefore, it is important to select a valid control group in order to reduce bias in the construction of the counterfactual. This paper aims to achieve this by employing matching techniques.

The matching process involves pairing remittance receiving households with non-remittance households based on an index of observables X. In this case, X is a vector of observable covariates that include mean household age, number of household residents, number of adults in the household, an indicator for the years of schooling in the household and a number of indicators of the race of the household. The choice of variables is influenced by what the literature on migration and remittances describes as determinants of migration (see for example Barham and Boucher, 1998 or Adams, 1989). Propensity score matching as defined inter alia by Rosenbaum and Rubin (1983) estimates the probability of receiving a treatment, or in this case remittances, conditional on this set of observable covariates X. Hence, the probability of receiving remittances, or the propensity score, can be defined as

$$P(REM_{it} = 1) = F(X_{it-1}). \tag{7}$$

The propensity score as defined in equation (7) is obtained using a non-linear model such as probit or logit model, depending on the distributional assumptions on the error terms. Let p_i denote the thus defined probability for a household to receive remittances in the group of remittance households R and p_j the probability of receiving remittances for a household in the control group C. As stated by Girma and Görg (2007), the standard matching estimator can be written as

$$\mu = \sum_{i \in R} \left(y_i - \sum_{j \in C} g(p_i, p_j) y_j \right). \tag{8}$$

In equation (8), $g(\cdot)$ is a function to assign weights on the comparison household j while constructing the counterfactual for household i. The different matching estimators explained in Section 3.4.2 differ in the weights assigned in this function $g(\cdot)$.

Conditional DID matching is a variation of this standard matching method in which Δy denotes the difference in household income at the end of the panel and income in the first wave. Following Heckman et al. (1998) the DID estimator is then defined as

$$\delta = \sum_{i \in R} \left(\Delta y_i - \sum_{j \in C} g(p_i, p_j) \Delta y_j \right). \tag{9}$$

This estimator is conditional on the probit estimations for the propensity score and thus on the set of observable characteristics X, hence "conditional difference in difference matching". These weights of the propensity score are dependent on the respective cross-sectional matching algorithm used. Similar to equation (8), DID matching re-weights observations according to the matching estimator's weighting function $g(\cdot)$. Smith and Todd (2005) argue the superiority of the conditional DID approach over its unconditional counterparts as the estimation of the conditional outcome variable in this approach is free of restrictions on linear functional form.

This paper will make use of DID matching to determine those non-remittance households that are most similar to households receiving remittances. Subsequently, these matched non-remittance households will be used in the estimation of predicted income excluding remittances (PREX) in the non-migration counterfactual. Implementing the DID matching method to determine which households to use for the prediction of non-remittance income is thought to reduce biases that arise in the estimations due to the fact that time invariant differences like the selection bias between migrants and non-migrants are accounted for. The following paragraph will explain the different algorithms that this paper utilizes for the matching process.

3.4.2 Different Matching Algorithms

Logit vs. Probit Matching

The most commonly used model in the estimation of propensity scores is a binary probit model. However, any discrete choice model can be used for the estimation according to Caliendo and Kopeinig (2008). Principally, logit or probit models are preferred over linear probability models as the functional form assumed in probit and logit models is more eligible. The choice between logit and probit is less important as the two yield very similar results for cases of a binary treatment although the distribution in the logit estimation yields higher density masses at the margins. In the later estimation, both logit and probit models will be implemented to test for robustness of the results against different estimation methods.

Nearest Neighbour Matching

The method of nearest neighbour matching is the most straightforward of the different matching methods according to Caliendo and Kopeinig (2008). An individual from the control group is matched to an individual in the treatment group when it is closest in terms of the propensity score. There are different methods of nearest neighbour matching. Firstly, matching can be done with replacement; i.e. an individual from the comparison group can be matched to more than one individual in the treatment group. When choosing nearest neighbour matching with replacement, one has to decide on a trade off between bias and variance. Matching with replacement will increase the matching quality and decrease the bias. However, variance will increase due to the fact that the counterfactual outcome is calculated with a reduced number of non-participants. Matching without replacement on the other hand does not face the problem of increased variance but requires special caution with regards to the ordering of the matched observation. It is important to make sure that this order is random in the matching process as nearest neighbour matching without replacement is sensitive to this order. Smith and Todd (2005) discuss this trade-off between bias and variance in more detail. Lastly, one disadvantage of nearest neighbour matching is that it faces bad matching in cases where the nearest neighbour is relatively far away.

Kernel Matching

As opposed to nearest neighbour, kernel matching is a non-parametric matching estimator. Instead of matching one non-participant to many participants, Kernel matching employs weighted averages of potentially all members of the comparison group in order to establish the counterfactual outcome. This results in lower variance compared to nearest neighbour matching as information of more individuals is used in the construction of the counterfactual. However, as Kernel matching uses weighted averages of a large fraction of the control group it is possible that these averages are calculated using observations that are bad matches. Heckman et al. (1998) offer an extensive discussion of the theoretical background of kernel matching. The fact that averages are calculated over large parts of the groups of non-participants makes it imperative to carefully establish the condition of common support explained above. Caliendo and Kopeinig (2008) explain that weights used in the kernel method are "dependent on the distance between each individual from the control group and the participant observation for which the counterfactual is estimated". The authors also stress that in the case where weights are implemented that stem from a kernel that is uni-modal, symmetric and non-negative the derived averages impose lower weights on matches that are more distant with regards to the propensity score than weights imposed on those that show propensity scores closer to one another.

All matching procedures in this paper are done using nearest neighbour matching. However, kernel matching will be implemented to test for robustness. If matches found in the data are of high quality, the different matching algorithms should yield similar results. Further matching methods that will not be implemented in this paper include caliper or radius matching as well as stratification or local linear matching. Section 5.2 will implement the matching algorithms discussed above in order to determine the appropriate control group of non-remittance households that can be used in the estimation of predicted income excluding remittances (PREX) discussed in this section. Section 4 will now review the data used in the decomposition of the Gini coefficient and the creation of a non-remittances counterfactual performed in Section 5.

4 Descriptive Statistics

This section discusses the data underlying the analysis in this paper and reviews the descriptive statistics of the relevant variables used in later assessment regarding the Gini coefficient in South Africa.

4.1 Data

The data used in this paper stems from the panel data set of the National Income Dynamics Study (NIDS). The NIDS data set offers great detail in the information collected on income from different sources and therefore is the suitable choice in the context of this paper. The data has been collected from 2008 to 2012 in two-year intervals. NIDS is a nationally representative study of South African individuals, this was achieved through assigning probabilities to the different provinces according to size in the Master Sample of primary sampling units (PSU's). This process was necessary in order to avoid a concentration of the sample in a few provinces. The sampling was assigned to Statistics South Africa and to draw the sample, a two-stage cluster sample design was implemented.

NIDS is the first panel data set for South Africa that collects information on individuals of all ages. Individual's information was collected mostly in private households but the study also comprises individuals residing in convents or worker's hostels. In the three waves, information about labour market participation, individual and household income from employment and non-employment sources as well as data on individual health and education has been conducted. Attrition rates have improved across waves due to better tracking in Wave 3 (NIDS, W3 User Manual). As such household attrition is at about 13% in Wave 2 and 10% in Wave 3. Individual level attrition is slightly higher due to the fact that within a household some individuals may agree to an interview while others refuse. Reasons for attrition may be interview refusal, on a household level or individual level, not tracked, moved outside SA or deceased. Individual attrition levels are 19% in Wave 2 and 16% in Wave 3.

4.2 Data Concerns

The main objective in this paper are different sources of income. For this purpose, the derived income variables provided by NIDS[2] will be used throughout the paper. These derived income variables contain information on labour market income (wage), income of a capital nature and investment sources, agricultural activities and implied rent. Furthermore, income from remittances and government grants are included. All income is derived as per month income. The variable on labour market income contains information not only on the first paid job but also on any other paid jobs as well as self-employment activities. In the case an individual earned an income from casual work, helping out a friend or a 13th paycheck or any other bonuses, the wage variable comprises that information as well. Investment income contains information on incomes from stocks, loans, rentals, as well as private pensions and retirement annuities. The variable on income of a capital nature contains a broad band of (mostly once-off) payment incomes such as labola (bridal payments), a monetary gift or inheritance, retrenchment payments as well as the pay-back of a loan to the income receiving person. Income from agricultural activities includes everything from the income from selling products of the land such as crops or vegetables as well as animal products such as milk, eggs or wool. All agricultural income is net of input costs which include for example the cost for labour and fertilizers or the cost of feeding the animals. The implied rent variable is included to account for income derived from owning a house or dwelling. It is constructed from the market value of the property and a variable that captures the amount a family would be willing to pay in order to occupy the dwelling that they currently own. However, when constructing this variable, the different kind of home-owners and non-rent payers have to be taken into account. Firstly, there are families living in the dwelling of someone else that are not paying rent. In the NIDS Technical Paper on household income in Wave 1 (Argent, 2009), these people are described to be constituting an income from the use of a dwelling by not paying for it. Hence, the implied rent variable takes the market value for those occupied dwellings as the rent implied although no actual rent is paid. Furthermore, the people who actually own the home their families occupy are separated into two categories, the ones with a mortgage on their dwelling and the ones without. While this has different effects on their expenditure, home-owners with and without are treated the same in terms of implied rent which equals the market rental price. This value is thought to be captured by asking[3] for the amount of rent the individual thinks could be collected were the place to be rented out. However, any such question may lead to biased results. Implied rent is therefore a slightly problematic variable but it will nevertheless be included in the later analysis of income sources due to its importance in the income decomposition. The variable on government grants captures any incomes received through state pension funds, grants for disability, children, foster children or care dependency. Furthermore, "other government grants"include income from the unemployment insurance fund and workman's compensation. Lastly, the derived variable on income from remittances not only contains the monetary contributions a household receives from a distant remitter but also the values stated of contributions in kind. All income variables applied in the following analysis are used in per capita terms. Per capita values are calculated using the number of household residents. A resident is defined as someone who spends at least 4 nights a week in the household and is not away from the household for more than three months a year. Opposed to residents, there are also household members. These are individuals associated with the household but spending more time away than the given thresholds. Household members could possibly be remittance payers, however, this study does not analyse the role of household members any further.

A migrant household is strictly one that receives remittances. Although possibly not the best definition of a migrant household, the quality of the data given demands such definition seeing that the purpose of this paper is the analysis of the role of remittances. Given the definition of household members and residents according to the criteria outlined above, households that report a member being away, for work or in the search of work often do not report an income from remittances and vice versa, not every household that receives remittances reports a household member away. In many cases, households do not report detailed information on the person sending remittances beyond the income received. Therefore it is not possible to gain further information on the remitters in the majority of cases. For the further analysis in this paper, this definition of a migrant household also implies the rather strict assumption that any remitter has been a member of the remittance receiving household and this household sent the migrant away.

Lastly, the analysis in this paper categorizes households into urban and rural households according to the geographic variables in the 2011 Census that are included in NIDS. The Census data reports households in three different geographical regions. These different geotypes include urban areas, farms and regions labelled "traditional". This is a new definition compared to the 2001 Census geotypes that saw urban formal and urban informal areas as well as rural formal and tribal authority geotypes. For the rest of the paper, rural areas comprise all households situated in traditional or farm areas on the one hand and urban areas on the other.

4.3 Summary Statistics

Table 1 reports the summary statistics in per capita values for the relevant variables that are used in the later analysis of the Gini coefficient. Although NIDS is a panel data set, this paper will exploit the repeated cross-sectional characteristics of the data. Therefore, all variables are reviewed separately for the different waves. Number of observations (N), mean and standard deviation for components of household income are reported in Table 1. The mean of total household income per capita increases from ZAR 1198.7 in Wave 1 to ZAR 2028.95 in Wave 3. Seeing that overall household income increases in the observed time span, most income sources increase as well. For example, income from labour market activities increases from ZAR 635 in Wave 1 to ZAR 914 in Wave 2 and ZAR 1186 in Wave 3. The large share of labour market income in overall income already becomes apparent. Income from remittances in per capita terms first decreases from ZAR 48 in Wave 1 to ZAR 39.5 in Wave to before it increases to ZAR 77.3 in Wave 3. The different numbers of observation reflect the handling of non-responses in the data. All income sources are reported separately in Table 1 to allow the reader a deeper understanding of the underlying income components. However, income from different sources will be grouped together for the later decomposition exercise. As such, all income from government grants will be

Table 1: Summary Statistics of Household Income Components

Variable	N	Mean	Std. Dev.
WAVE 1			
HH Income	7212	1198.77	7769.189
Labour Income	7212	634.7445	1893.82
Government Grants	7212	146.8864	203.0866
Other Government Grants	7212	3.333955	59.96827
Investment	7212	51.45503	481.6339
Capital	7212	138.8342	6878.765
Remittances	7212	47.93394	236.2791
Implied Rent	6538	190.9847	491.0229
Agriculture	1655	10.78338	57.58349
WAVE 2			
HH Income	6495	1561.217	7267.143
Labour Income	6495	914.4204	3080.15
Government Grants	6495	180.9422	252.1437
Other Government Grants	6495	2.404023	49.3447
Investment	6495	75.56537	686.6517
Capital	6495	118.1125	6341.453
Remittances	6495	39.53614	438.4983
Implied Rent	6409	242.0276	743.0446
Agriculture	6495	13.19709	639.8328
WAVE 3			
HH Income	6865	2028.949	12208.21
Labour Income	6865	1185.539	2680.929
Government Grants	6865	210.784	301.3488
Other Government Grants	6865	3.879027	84.4851
Investment	6865	100.9149	1061.168
Capital	6865	193.8813	11587.67
Remittances	6865	77.29944	783.0679
Implied Rent	6306	277.7176	929.5288
Agriculture	6865	15.00476	477.1017

reported as one category. Furthermore, income from investment and capital sources will be reported together and income from labour market as well as agricultural activities will be analyzed as one variable. This allows to focus on the few income categories of interest in this paper as the decomposition exercise serves to examine the role of remittances in the income distribution in South Africa.

Additionally, Table 2 gives information about characteristics of the households such as

Table 2: Summary Statistics on Household Characteristics

Variable	N	Mean	Std. Dev.
WAVE 1			
HH Size	7297	5.925449	150.1603
HH Assets	5711	7.331933	3.80167
Mean Age	7297	29.52241	16.35262
No. of Adults	7297	1.734411	1.51732
Mean Education	7296	9.773163	3.485987
African	7297	.6639715	.4723811
Coloured	7297	.11909	.3239166
White	7297	.0501576	.218285
Asian/Indian	7297	.0106893	.1028422
Number of HH's receiving remittances		2658	
WAVE 2			
HH Size	7117	4.507517	2.91882
HH Assets	4784	8.084681	3.775879
Mean Age	7117	23.75067	17.29908
No. of Adults	7117	2.032598	1.346849
Mean Education	6062	10.23111	3.039128
African	7117	.7671772	.42266
Coloured	7117	.1265983	.3325458
White	7117	.0303499	.1715602
Asian_Indian	7117	.0091331	.0951363
Number of HH's receiving remittances		2831	
WAVE 3			
HH Size	8156	4.260544	2.836033
HH Assets	8156	6.242819	5.658522
Mean Age	8156	24.83785	16.80638
No. of Adults	8156	1.883276	7.715582
Mean Education	7250	10.46041	2.977154
African	8156	.8064002	.3951431
Coloured	8156	.1227317	.3281491
White	8156	.0290584	.1679805
Asian_Indian	8156	.0093183	.0960864
Number of HH's receiving remittances		2867	

mean age and mean level of education. Level of education is measured by years of schooling that range from 0 to 12. However, post-schooling information is added to include post-school diplomas, NTC's as well as any tertiary education. For these certificates, the measurement ranges from 13 to 17, whereas 17 is a PhD degree. For the remainder of this paper, whenever the estimation of predicted income excluding remittances refers to the education variable, it speaks about this mean level of household education. Mean household education is calculated as the mean of the years of schooling of all household members. Furthermore, Table 2 reports the number of household residents, i.e. the size of the household, as well as the number of adults residing in the home. Household size as well as number of adult are theorized as having a positive influence on sending out migrants. Firstly because as the number of residents increases so does the likelihood of one going away. Additionally, if there is a larger number of adults left in the households, they are able to compensate the loss of the migrant's assistance in the household. Hence, the household faces lower opportunity costs. The number of adults is added as the composition of a household is theorized to have an influence on migration status as well as mean household age. The age average of all household members is included in the calculations as younger household may be more willing to take the risk of sending out a migrant than households with a higher number of older members. Table 2 also reports a set of race indicators included in the estimation. The race for a household is determined by the mode of the race of the individuals; the mode is the value occurring most frequently. The household race variable is also provided in the derived variables of NIDS. For all three waves, African households dominate all other household race dummies. Interestingly, their share seems to be increasing. The second most common are coloured whose share is stagnant over time and the least common are Asian and Indian households. Furthermore, a variable was created that reports the log values of the assets that the household owns. Included in the variable for household assets are vehicle assets, such as the value of a car or truck owned for private or economic use (work). Furthermore, livestock assets such as cattle, chicken or sheep are included. Lastly and most importantly, the asset value of property or a dwelling owned are included. This is measured by the log of the market value stated by the individuals. In most cases this variable comprises the largest share in the household asset variable. However, as can be seen from the fluctuating number of observation, information on household assets is not easily obtained. The problematic of this issue has been discussed in the previous subsection. The variables presented in Table 2 will be relevant in the construction of the counterfactual.

5 Results

After the descriptive statistics have been discussed in the previous section, this section uses the NIDS data set to analyze levels of inequality in South Africa by decomposing the Gini coefficient as a measure of inequality according to income sources. Furthermore, the data will be used to create a counterfactual that allows the assessment of levels of inequality that would have prevailed had migration not taken place.

5.1 Decomposition of the Gini

This section presents the decomposed Gini coefficients for different sources of income for South African households in 2008, 2010 and 2012 (NIDS Wave 1, Wave 2 and Wave 3). Gini coefficients are calculated using per capita income of households.

In Table 3, household income is decomposed into five different sources. The first income variable noted is labour incomes. It is a sum of all wage income in per capita terms. That indicates, incomes from the main job as well as any second jobs, bonuses as well as self-employment and other wage income are included. Furthermore, this variable comprises any income from subsistence agricultural activities as well as from consuming what is produced by the household. Government grants include state old age pension funds, incomes from disability grants and child support grants as well as foster care and care dependency grants. A small number of households reported income from unemployment insurance funds (UIF) and workman's compensation which are also included in this income variable. Furthermore, income from remittances, the variable of interest in this paper, is included and lastly, income from implied rent is covered. NIDS constructed a derived variable for implied rent in order to take into account the income a family derives from home ownership in the form of not having to pay rent. The capital variable covers incomes from investment, i.e. interest or dividend generating incomes, such as private pension funds and retirement annuities as well as stocks, loans and rentals but also incomes of a capital nature. This includes inheritance income, bridal payments such as the traditional "labola"and other monetarist gifts. However, also the kind of income that is generated when receiving the repayment of a loan to the person and received income from sale of household goods are interpreted as capital income. Finally, the variable for capital includes an assembly of different kinds of incomes such as the war veterans pension that does neither fit into another category nor do many people receive income from such a source. The fact that the income variable 'Capital' comprises income from this many sources explains the relatively high share of income from this source in overall income.

The decomposition exercise[4] shows that income inequality in South Africa is still stagnating. The overall Gini coefficients across all three waves are at about 0.64 with a mild increase in Wave 2 to 0.648. Table 3 firstly reports the decomposition into different income sources for Wave 1. The share s reports the share of the respective source of income in the overall household income. Labour market income is the biggest source of income for all three waves and fluctuates around 55 percent. Remittance income remains almost the same in its share in overall income at about 3.5 percent in Wave 1, 2.4 percent in Wave 2 and about 3.2 percent in Wave 3. Given that the income variables on government grants and capital comprise income from several sources it is not surprising that Table 3 reports a relatively large share in total income for government grants and capital income respectively. Lastly, the variable on implied rent income reports a relatively large share as well, with up to 17% in Wave 2 and around 14% in Wave 1 and Wave 3.

In the second column, g reports the Gini coefficient of the respective income source.

Table 3: Gini Decomposition into Income Components

Variable	Share s	Coeff. g	Corr. r	Contri. s*g*r	%Contri. s*g*r/G	Elasticity s*g*r/G-s
Wave 1						
Labour Income	0.5522	0.7713	0.9307	0.3964	0.6155	0.0633
Government Grants	0.1157	0.6194	0.1058	0.0076	0.0118	-0.1039
Remittances	0.0348	0.9360	0.4517	0.0147	0.0228	-0.0120
Implied Rent	0.1407	0.6796	0.7930	0.0758	0.1177	-0.0230
Capital	0.1566	0.9924	0.9623	0.1496	0.2322	0.0756
Total HH Income	1.0000	0.6441	-	0.6441	1.0000	-
Wave 2						
Labour Income	0.5688	0.8199	0.9255	0.4316	0.6658	0.0970
Government Grants	0.1394	0.6250	0.1956	0.0170	0.0263	-0.1131
Remittances	0.0241	0.9648	0.5051	0.0118	0.0181	-0.0060
Implied Rent	0.1772	0.7426	0.8468	0.1114	0.1719	-0.0053
Capital	0.0906	0.9793	0.8620	0.0764	0.1179	0.0274
Total HH Income	1.0000	0.6482	-	0.6482	1.0000	-
Wave 3						
Labour Income	0.5401	0.7769	0.9188	0.3855	0.6016	0.0615
Government Grants	0.1222	0.6218	0.1000	0.0076	0.0119	-0.1103
Remittances	0.0322	0.9434	0.4390	0.0133	0.0208	-0.0114
Implied Rent	0.1481	0.7106	0.8263	0.0870	0.1357	-0.0124
Capital	0.1575	0.9903	0.9455	0.1475	0.2301	0.0726
Total HH Income	1.0000	0.6408	-	0.6408	1.0000	-

Across all three waves, these Gini coefficients are rather high. The Gini coefficient of remittance income is between 0.94 and up to 0.96 in the three waves. As a majority of households report zero income from remittances, a high level of inequality within this income source would be expected. Similar holds true for income from capital and investment sources. These mechanisms are not available to everyone and hence, high levels of inequality within these sources and a Gini above 0.97 in all waves are to be expected. The lowest Gini is reported for income from government grants, although government grants still report Gini coefficients of about 0.62. Therefore, income from this source is rather unequal as well. However, the biggest concerns should stem from the steep level of inequality within the largest income source, labour income. Since labour income makes up such a large share in overall household incomes, high levels of inequality in this variable have more severe effects on overall inequality. This can also be seen in the following columns of Table 3.

The third column reports the Gini correlation r of income from the respective source with the distribution of total household income. This correlation coefficient will be zero if the respective source of income k and total income are independent. In the same way, r will be equal to 1 or -1 if income source k is an increasing or decreasing function of overall income. Most of the income sources under observation report a rather strong positive correlation with overall income. For example income from labour sources reports a correlation with overall inequality of 0.93 for Wave 1 and Wave 2 and 0.92 for Wave 3. This supports the concern about the steep level of inequality within this income component as it appears to be a driving force of overall inequality in South Africa. The income source with the lowest correlation is income from government grants which reports a correlation of only about 0.1 for Wave 1 and Wave 3, however in Wave 2 it reports a correlation almost twice as high with 0.19. This can be explained by the fact that grants should mainly benefit households at the bottom end of the income distribution and should therefore not be strongly associated with overall inequality. Remittances report an r of 0.45 in Wave 1, 0.5 in Wave 2 and 0.44 in Wave 3.

In the next column, the contribution of the respective source to the Gini coefficient of overall household income is reported in absolute values and the second last column reports the contribution in percentage terms. The second last column emphasizes that the distribution of wages play an important role in overall inequality. This was to be expected given that they make up such a big share in overall income. The percentage contribution of inequality in the labour market income to overall income inequality is about 61 percent for Wave 1 but increases to about 67 percent in Wave 2 only to fall back to about 60 percent in Wave 3. Remittances on the other hand make up a rather small part of the overall income inequality; their contribution is around 2 percent for all three waves. Government grants also report only a small contribution to the overall Gini coefficient with 1.2 percent in Wave 1, 2.6 percent in Wave2 and 1.2 percent in Wave 3.

Finally, the last column in Table 3 holds important information for the analysis of the different income components. There the elasticity $S_k G_k R_k / G - S_k$ is reported or in other words the effect of a 1% change in the respective income source on the overall Gini coefficient holding income from all other sources constant. Changes in wage income as well as income from capital sources seem to have the largest positive effects on the overall income inequality. A minor increase in both these income sources would lead to an increase in overall income inequality. The effect of a marginal increase in labour market income ranges from an increase in overall inequality of about 0.06 percent in Wave 1 and 3 up to an inequality increasing effect of 0.09 percent in Wave 2. Similarly, a minor increase in capital income would lead to an increase in the overall Gini coefficient of about 0.07 percent for Wave 1 and Wave 3. In Wave 2 however, a 1% increase in capital income would only lead to an increase in overall inequality of about 0.026 percent. Income from remittances has a negative effect on overall inequality. If remittances increase by 1%, overall inequality will decrease by 0.012 percent in Wave 1 and 0.011 in Wave 3. For Wave 2, the Gini coefficient would only decrease by 0.006 percent if income from remittances would increase by 1%. Hence, it can be assumed that remittances primarily flow towards households near the lower end of the income distribution, lifting them up and improving general inequality. This has implications for economic policy as well. Policies that make migration easier may lead to an increase in remittance payments. The largest negative effect on overall inequality stems from income from government grants. When income from this source increases by 1%, the overall Gini coefficient would decrease by 0.10 percent in Wave 1 and 0.11 percent in Wave 2 and Wave 3. This result stresses the role of government and economic policy in reducing inequality in South Africa. It supports that grants are favouring those households at the bottom end of the income distribution and should therefore be continued. Lastly, implied rent has a negative elasticity as well. However, like income from remittances the effect is rather small. A marginal increase in implied rent would lead to a decrease of about 0.02 percent in Wave 1, 0.007 in Wave 2 and 0.012 in Wave 3. Despite these small effects, increased home ownership could also lead to a reduction in present levels of inequality. This could be achieved for example by following the government's plan to increase housing available in the townships but also in other places and by improving channels that ease access to home ownership in general.

The decomposition exercise in this subsection has highlighted the inequality reducing effects of remittances. However, the decomposition of the Gini coefficient according to income sources has also shown the strong influence of labour income and government grants on overall inequality as well as the effect of income from a number of capital and investment sources. The counterfactual scenario that will be created in the subsequent paragraphs aims at testing the effect of remittances by assessing the levels of inequality that would have occurred if there were no remittance payments. It can be assumed that households without remittances compensate this lack in income from one source by increasing income from other sources. Therefore, the counterfactual scenario that will be created in this paper matches non-remittance households to remittance receiving households using propensity score matching. Using information on matched households, a new income distribution will be estimated. This exercise is necessary as simply excluding remittances as an income source would not account for the compensation from other income sources mentioned. Hence, such an estimation would lead to a bias in the results. The counterfactual estimated in this paper tries to reduce such bias. The following subsection will explain the process and the results of the counterfactual in more detail.

5.2 Results of the Counterfactual Analysis

The counterfactual that will be presented in this subsection was created by matching remittance and non-remittance households in a conditional difference in difference (DID) matching estimation. The non-remittance households that were thus identified to being similar to remittance receiving households are used to estimate predicted household income excluding remittances (abbreviated with PREX). Subsequently, the parameters obtained in this estimation will be applied to remittance receiving households to determine how the income distribution would have looked like had these households not received remittances. The detailed results of this procedure will be discussed in the remainder of this section.

5.2.1 Conditional Difference in Difference Matching

All matching procedures are done separately for rural and urban households. The necessity for this separate process arises due to the fact that matching requires to compare individuals in the same economic environments (Caliendo and Kopeining, 2008). In other words, for the matching to be successful, it is important to compare similar groups in treatment and control. As such, not differentiating between urban and rural households would lead to less robust matching results. Heckman et al. (1998) stress the point that comparing incomparable groups may contribute considerably to selection bias measured. In a first step, difference in difference matching estimators are created by using the nearest neighbour method. For this, a probit estimation was run to determine the probability of a household to receive remittances. Based on this, the propensity scores were estimated which were then utilized in the conditional DID matching estimator. The entire process of conditional DID matching will be done using logit as well as probit algorithms to ensure the results obtained are robust against the use of different matching methods. Table 4 presents the results of the probit and logit estimations of equation (7). To recall, equation

$$P(REM_{it} = 1) = F(X_{it-1})$$

estimates the probability of receiving remittances based on a set of observable covariates X. Hence, the probit as well as the logit estimation was run on a binary variable that observes the reception of remittances. Tested was the effect of different observable covariates that are assumed to affect whether a family will send a migrant or not. As

Table 4: Probit vs Logit Estimations

	(1) Probit Rural	(2) Probit Urban	(3) Logit Rural	(4) Logit Urban
1 = Receiption of Remittances				
Household Income (ln)	0.0764***	-0.0318*	0.125***	-0.0510*
	(5.25)	(-1.98)	(5.27)	(-1.97)
Number of Household Residents	0.0543***	0.0600***	0.0865***	0.0951***
	(8.15)	(6.36)	(8.01)	(6.26)
Mean Household Age	-0.00977***	-0.00167	-0.0163***	-0.00283
	(-6.25)	(-1.00)	(-6.31)	(-1.04)
No. of Adults	-0.0653***	0.00234	-0.105***	0.00641
	(-4.89)	(0.14)	(-4.86)	(0.24)
Mean Household Education	0.0170***	0.0302***	0.0280***	0.0490***
	(3.59)	(4.57)	(3.64)	(4.55)
African	-0.152	0.0384	-0.240	0.0614
	(-1.61)	(0.45)	(-1.58)	(0.45)
Coloured	-0.792***	0.0681	-1.307***	0.106
	(-7.32)	(0.78)	(-7.38)	(0.76)
Asian_Indian	0.198	-0.239	0.327	-0.403
	(1.30)	(-1.63)	(1.33)	(-1.68)
White	-0.354	-0.279*	-0.566	-0.462**
	(-1.95)	(-2.55)	(-1.93)	(-2.59)
Constant	-0.579***	-0.535***	-0.935***	-0.861***
	(-4.02)	(-3.59)	(-4.02)	(-3.59)
Observations	12833	9557	12833	9557

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

was to be expected, the two different algorithms do not differ much in the results and for all variables the estimations show the same signs. The set of covariates X includes a number of observed variables, such as the number of adults in a household and mean age of the household. In urban areas, the number of adults in a household as well as the mean age are not significant. The variable on household age captures the effect of mean age in the household, it is significant on a 1%-level for rural families. Mean household age thereby has a negative effect, i.e. the older the mean age of a household, the less like this household is to send out a migrant and receive remittances. Although statistically insignificant, mean age also has a negative sign for urban households.

The variable that captures the effect of the number of adults, i.e. the number of household members that are 18 years or older, is significant in rural households and also has a negative effect. This implies that the probability of receiving remittances as an additional income is less when households report a higher number of possibly income earning adults. For all households the variable that captures income is significant. Interestingly, the variable has a positive sign for rural households but a negative sign for urban households. This implies that while rural household are more likely to receive remittances with higher incomes, urban households are less likely to receive remittances when they receive high overall household incomes. Logically, rural households are only able to send out migrants if they have a high enough income to accommodate possible migration costs. Households with lower incomes in rural areas may lack the means to migrate. Urban households, however, have little incentive to send out migrants when they already have a high household income. Therefore urban high-income households are less likely to receive remittances relative to urban low-income households that might expect higher income through remittances. To check for robustness of this linear relationship of income, a second probit and logit estimation was run including a squared income term. However, the results on this variable are insignificant. Therefore, the squared income term was excluded in the estimations reported in Table 4.

Furthermore, the mean level of education is significant on a 1%-level for all households and has a positive sign. Therefore, the higher the mean level of education, the more likely it is for a household to receive remittances. It is important to notice that the categorical variables that were included in order to control for race show very mixed results. Overall, race does not seem to be very significant in this estimation. For rural households, only the bivariate variable for coloured households is significant, for urban households the variable indicating a white household is. In both cases the significant variables report negative effects on the reception of remittances. All in all, differences between logit and probit models can be neglected, as for every variable tested the two estimation methods report variables with the same sign and comparable values.

The propensity scores for the matching process were calculated based on the probit model reported in Table 4 for nearest neighbour and kernel matching.[5] Additionally, the results of the logit estimation model will be applied to nearest neighbour matching to test whether the different estimation models affect the matching. The propensity score thereby reports the probability to receive remittances conditional on the set of covariates X discussed previously. The methodology of the matching process has been discussed in detail in Section 3.2. The conditional difference in difference matching process was employed on the differences in household income between the first and the last observation period in order to calculate equation (9) using different matching algorithms

$$\delta = \sum_{i \in R} \left(\Delta y_i - \sum_{j \in C} g(p_i, p_j) \Delta y_i \right).$$

Hence, the probit as well as the logit estimations are used to calculate the propensity scores on the difference in household income in 2012 (NIDS Wave 3) and 2008 (NIDS Wave 1) in the equation above. These propensity scores are then used to match households according to equation (9). The results of these matching processes for the probit estimations are represented in Figures 1 and 2 as well as Tables 5 and 6.

Figures 1 and 2 serve to report on the quality of the processes for nearest neighbour matching. As mentioned previously, propensity score matching is only possible if certain assumptions hold. One of these assumptions is common support. If this assumption was violated in the data, Figures 1 and 2 would report those propensity scores of households that are not supported. However, in this estimation the matching shows good balancing in the results and all households satisfy the assumption of common support. Propensity scores for the treated households are reported in white in the top half of the figure, the bottom half reports the propensity scores for untreated households in grey. Then, households from the treatment and control group were matched based on this propensity score using nearest neighbour matching. As can be seen from the figures, the matching process is successful for the majority of the treated households. Only in one case on the extreme ends of the propensity scores no match can be found in the control group for rural households. The fact that the two halves of the figures are balanced relatively well confirms the successful matching process. Similar holds true for the matching using the kernel algorithm and when the propensity score is estimated using a logit model. Figures 3 to 6 display these matching outputs in the appendix.

Following Rosenbaum and Rubin (1983), the level of balancing achieved can also be

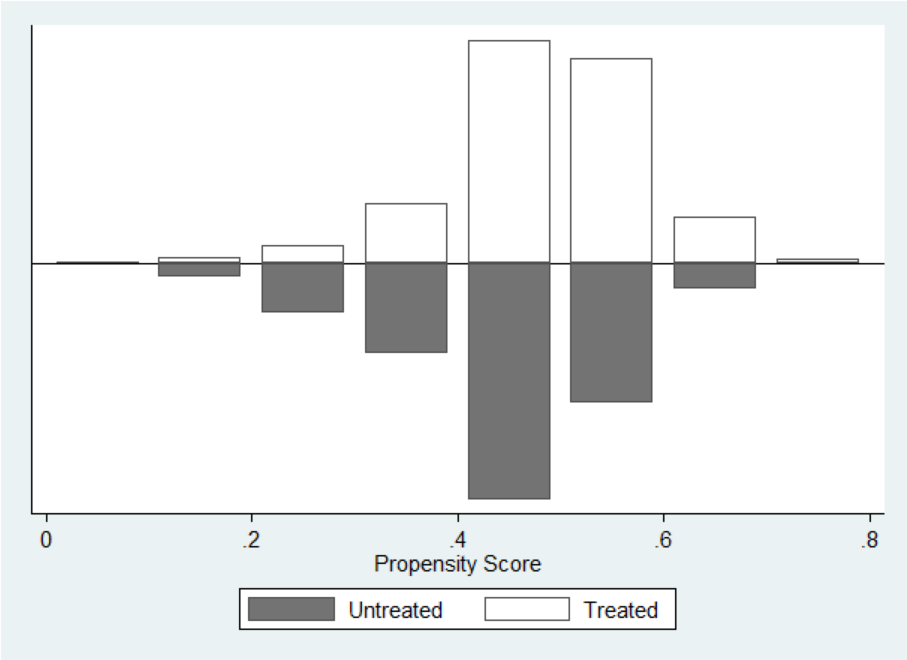

Figure 1: Nearest Neighbour Probit Matching - Rural Households

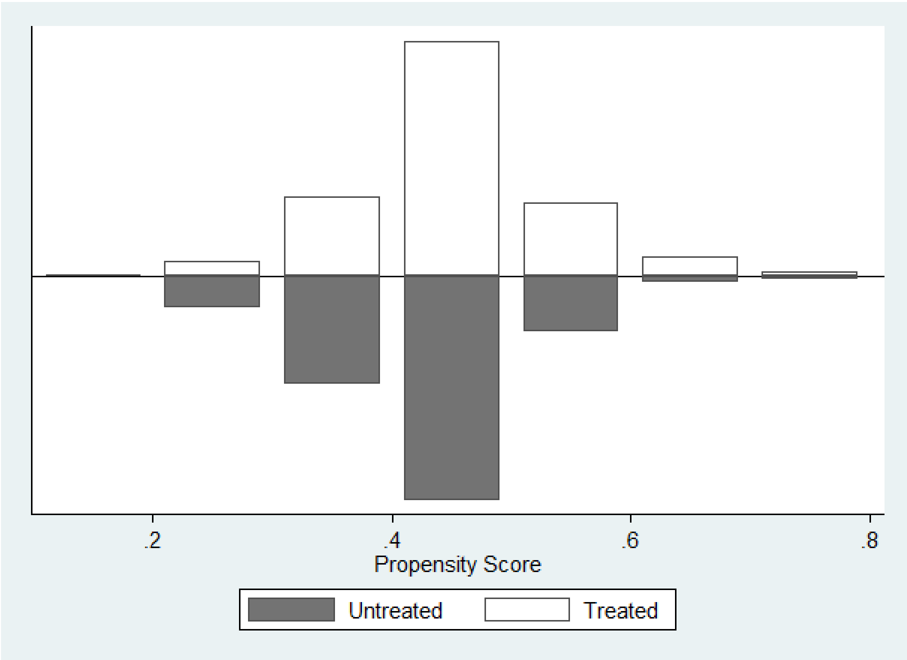

Figure 2: Nearest Neighbour Probit Matching - Urban Households

tested by running a t-test after the matching was performed. This test is suitable to check whether there are significant differences between both groups with regards to the means of the covariates. Before matching, differences would be expected but after the matching, there should not be any significant differences. Tables 5 and 6 report to which

Table 5: Results of the t-test after ps matching - Rural Households

		Mean			%reduct
Variable	Sample	Treated	Control	%bias	bias
HH Income (ln)	Unmatched	7.6174	7.513	11.7	
	Matched	7.6174	7.6239	-0.7	93.8
HH Size	Unmatched	6.8904	5.918	29.1	
	Matched	6.8904	6.4299	13.8	52.6
Mean Age	Unmatched	23.48	26.268	-28.0	
	Matched	23.48	23.838	-3.6	87.2
No. of Adults	Unmatched	3.0541	2.8515	12.5	
	Matched	3.0541	2.8736	11.1	10.9
HH Education	Unmatched	10.22	9.7146	18.4	
	Matched	10.22	10.248	-1.0	94.4
African	Unmatched	.94705	.90515	16.1	
	Matched	.94705	.90716	15.3	4.8
Coloured	Unmatched	.0224	.07026	-22.9	
	Matched	.0224	.01222	4.9	78.7
Asian_Indian	Unmatched	.01103	.00698	4.3	
	Matched	.01103	.01544	-4.7	-9.0
White	Unmatched	.00458	.00727	-3.5	
	Matched	.00458	.00458	0.0	100.0
Pseudo R2	Matched	0.017			
N	Treated	5,892			
	Untreated	6,874			

extent balancing has been achieved on the matched households for the nearest neighbour method. When analyzing Tables 5 and 6, the mean reported for the respective variables of the matched subsamples is the most important feature to interpret. Matching can be seen as successful where the characteristics of the matched households of treatment and control group are similar. For all variables these means differ only minimally leading to the conclusion that matching was successful. For example, in Table 5 rural households report

Table 6: Results of the t-test after ps matching - Urban Households

Variable	Sample	Mean Treated	Mean Control	%bias	%reduct bias
HH Income (ln)	Unmatched	8.0913	8.0701	2.0	
	Matched	8.0913	8.0471	4.2	-108.8
HH Size	Unmatched	5.7691	4.8921	30.8	
	Matched	5.7691	5.6339	4.7	84.6
Mean Age	Unmatched	26.039	27.988	-18.5	
	Matched	26.039	26.501	-4.4	76.3
No. of Adtults	Unmatched	3.0967	2.7545	22.3	
	Matched	3.0967	3.0915	0.3	98.5
HH Education	Unmatched	11.08	10.803	11.6	
	Matched	11.08	11.228	-6.2	46.3
African	Unmatched	.67755	.65966	3.8	
	Matched	.67755	.70135	-5.1	-33.1
Coloured	Unmatched	.26847	.24986	4.2	
	Matched	.26847	.24123	6.2	-46.4
Asian_Indian	Unmatched	.00883	.01521	-5.8	
	Matched	.00883	.00589	2.7	53.8
White	Unmatched	.0292	.05894	-14.5	
	Matched	.0292	.01791	5.5	62.0
Pseudo R2	Matched	0.007			
N	Treated	4,075			
	Untreated	5,327			

a mean of the treatment group for the variable that captures household age of 23.5 years compared to 23.8 years in the control group. Therefore good balancing was achieved for this variable. A little less well balanced but still acceptable is the variable that captures the number of residents in a household. For rural households that receive remittances, the average number of household members is 6.89 and households in the control group show on average 6.4 household members. Nevertheless, this is a minor difference that does not throw off the balancing of the matching. Similar holds true for the balancing of the remaining set of covariates for rural households. Results for urban households as displayed in Table 6 are equally well balanced. For example the mean household size for matched urban households is 5.7 for remittance households and 5.6 for non-remittance households. These levels are slightly below the mean size of rural households. Urban households also report on average older households. Households receiving remittances thereby report a mean age of 26 years. Matched households that do not receive remittances report a mean age of 26.5 years. All in all, all covariates are well balanced for urban households as well.

Additionally to the mean values of the different covariates, Tables 5 and 6 further report the relative differences in the covariates between matched and unmatched households. The column (%bias) shows that an excessive reduction in the differences between the two groups was achieved for all but the African race indicator for rural households. This indicator shows a small reduction in bias, however when looking at the mean values reported for treatment and control group of matched households, the two values are fairly the same. Treated households report a mean of 0.947 and control households a mean of 0.907. Thus the small bias reduction should be of little concern. For example, the differences in household education levels dropped from 18.4% to 1% for rural households. For urban households the relative reduction in differences was also quite successful. The African race indicator for example changed from 11.6% to -6.2%. In the case of urban households this kind of reduction in relative bias was achieved for most covariates. The relative reduction in this bias is reported in the next column (%reduct. bias). The higher the relative reduction, the more bias was reduced through matching. For example, the variable for household income in Table 5 reports a relative reduction in bias of 93.8%. This means that almost all bias was eliminated through matching.

Most variables report rather large values in this column, confirming the good balancing of the matching. Finally, Tables 5 and 6 state the number of treated and untreated households that were matched and the pseudo-R^2. Caliendo and Kopeinig (2008) suggest a re-estimation of the propensity score only on the matched households in treatment and control group. Then the pseudo-R^2 serves as an indicator for how well the set of covariates X explains the probability of receiving remittances. As the distribution of the covariates should report no systematic differences between the treatment and control group, the pseudo-R^2 should be rather low. Therefore, a pseudo-R^2 of 0.017 for urban households and 0.007 for rural households respectively indicates that matching has successfully taken care of any systematic differences across the two groups. Similar holds true for kernel and logit estimations. Results of the t-test for kernel matching and the logit estimation are reported in Tables 10 to 13 in the appendix.

This section has presented the results of propensity score matching when it is applied to rural and urban households in the NIDS data set. For this reason, probit and logit model of the probability to receive remittances were estimated. The propensity scores thus estimated were then employed in nearest neighbour and kernel matching on a conditional difference in difference estimator. The good balancing achieved by these matching algorithms was supported by the results of the t-test as well as the pseudo-R^2. In the following subsection, non-remittance households thus matched will be used to estimate the household income of remittance households had they not received those payments.

5.2.2 Estimation of Predicted Income Excluding Remittances

The results discussed in the previous section support the quality of the matching process. In this section the households thus matched will now be used to predict the counterfactual income excluding remittances for otherwise remittance receiving households. Equation (6) displayed below was discussed in section 3.2.

$$PREX = HH\_ASSET + EDUC + HHSIZE + GEOTYPE + HHRACE$$

This subsection will first estimate equation (6) for the matched households of the non-remittance control group. The coefficients of this estimation will then be applied to the remittance receiving households to predict their income in the counterfactual scenario of no remittances. Utilizing this newly obtained information on household income, the counterfactual Gini coefficient will be calculated to be able to compare the results of the counterfactual with the actual results.

Tables 7 and 8 report the regression of equation (6) on matched households that do not receive remittances for Wave 1, Wave 2 and Wave 3. Variables used in the regression are the log of household assets, the number of years of schooling, number of household residents as well a variable that indicates race. Although equation (6) states $GEOTYPE$ as a covariate, this variable was left out in the estimation. Instead equation (6) was estimated separately for rural and urban households as a necessary consequence from the separate propensity score matching. As explained earlier, the propensity score matching had to be run separately for rural and urban households in order to avoid comparing the incomparable (Heckman et al., 1998).

For each of these separate estimations of equation (6), standard errors are reported in parenthesis. For all three waves, most of these variables are significant for the regression. The only exception is the number of household residents for urban households in Wave 1. The number of observations reported is the number of matched households that do not receive remittances. Income excluding remittances is estimated in Rand values and not in log values as the Rand value is needed to calculate the Gini in the next step of the creation of the counterfactual. Table 7 shows that for urban households, the sum of household assets is significant on a 1%-level for all three waves and has a positive effect on income excluding remittances. Hence, income is increasing with household assets. In Wave 1, this effect is the strongest. The mean level of education is also significant on a 1%-level and has positive values for all three waves. The mean level of education is thereby more significant in Wave 3. The positive sign of the education variable implicates that higher educated individuals in a household are more likely to contribute more positively to overall household income, this can be associated to the fact that higher levels of education are generally rewarded with higher returns to skill. The number of household residents has ambiguous effects on household income excluding remittances. In Wave 1, the variable is insignificant for urban households. In Wave 2 and Wave 3, the number of residents is significant on a 5%-level, however, it has a positive effect in Wave 2 and a negative effect in Wave 3. Finally, a variable capturing race is significant on a 1%-level and positive for all three waves. To recall, the race indicator is 1 for African households and 4 for white households. The indicator has a strong positive effect for urban households.

Table 8 reports the result of the estimation of income excluding remittances for matched rural households. Again, the race indicator has a strong positive effect on income. In rural households, the number of household residents has a positive effect on income and is significant in all three waves. Hence, the more residents, the higher the income of the household. Both household assets as well as level of education have a positive effect on household income in rural areas. With the exception of the number of household residents in Wave 2, which is significant at a 5%-level, all variables are significant on a 1%-level.

The coefficients obtained from the estimations displayed in Tables 7 and 8 are subsequently used to predict income excluding remittances (PREX) for households that initially did receive income from remittances. Although not all households were matched, the coefficients gained from the matched control households will be applied to all households that received remittances. It is important to note that this might allow for a new bias to arise in the estimation. This is a notable shortcoming in the estimation of this paper,

Table 7: Regression for Prediction of Income Excluding Remittances - Urban Households

	(1) PREX1	(2) PREX2	(3) PREX3
Sum of all HH assets (ln)	422.9***		
	(10.16)		
Mean Education	1082.2***		
	(16.30)		
Number of household residents	-80.97		
	(-1.33)		
HH Race	4585.6***		
	(22.30)		
Sum of all HH assets (ln)		310.9***	
		(6.33)	
Mean Education		1479.3***	
		(18.08)	
Number of household residents		153.8*	
		(2.22)	
HH Race		5324.6***	
		(21.28)	
Sum of all HH assets (ln)			233.6***
			(7.77)
Mean Education			1632.7***
			(24.53)
Number of household residents			-143.7*
			(-2.32)
HH Race			4994.2***
			(25.27)
Constant	-15152.5***	-19505.5***	-18450.1***
	(-19.74)	(-19.82)	(-23.62)
Observations	4472	4161	4868

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 8: Regression for Prediction of Income Excluding Remittances - Rural Households

	(1) PREX1	(2) PREX2	(3) PREX3
Sum of all HH assets (ln)	293.8***		
	(6.56)		
Mean Education	214.3***		
	(4.26)		
Number of household residents	190.9***		
	(4.15)		
HH Race	3220.7***		
	(8.45)		
Sum of all HH assets (ln)		285.6***	
		(11.06)	
Mean Education		511.9***	
		(14.61)	
Number of household residents		63.40*	
		(2.17)	
HH Race		4997.7***	
		(20.37)	
Sum of all HH assets (ln)			130.3***
			(11.62)
Mean Education			426.6***
			(19.77)
Number of household residents			160.3***
			(8.73)
HH Race			4374.1***
			(30.53)
Constant	-5676.3***	-9405.6***	-6959.7***
	(-8.33)	(-20.24)	(-24.87)
Observations	6052	5473	6029

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

however this step is necessary for the results of the counterfactual to be comparable to the results of the decomposition exercise. Estimating income excluding remittances for the matched households only would not allow for such comparability. For these matched households, however, bias should be reduced in the estimation of income excluding remittances. Furthermore, the estimation still takes into account the behaviour of households that may substitute the loss of income from remittances through compensating with income from other sources. The estimation of the counterfactual is therefore still deemed significant and superior to a counterfactual in which remittances are simply set to zero. The counterfactual estimation will proceed as previously discussed.

The new income distribution is obtained by imputing the values of $PREX$ for remittance receiving households into the initial income distribution. Table 9 reports the Gini coefficients that are obtained by using the different matching algorithms and when the different regression methods are used. All Gini's presented in Table 9 are calculated using analytical weights provided by NIDS. This summary of the counterfactual Gini coefficients shows the robustness of the results obtained by using different estimation methods. The Gini coefficients do not differ across kernel or nearest neighbour matching methods nor across probit and logit estimations. As discussed previously, Smith and Todd (2005) have found that the results of propensity score matching do not differ much across different matching methods. Hence, the robustness of the Gini across different estimation methods is owed to the fact that although different matching algorithms were employed, the same or similar households were assigned to treatment and control group after the matching. As the prediction of income excluding remittances takes the same values for the same matched households, overall results do not seem to differ. The fact that similar or the same households were matched in either estimation method also supports the good quality of the matching estimations applied.

The Gini coefficients of the counterfactual scenario converge towards the decomposed Gini coefficients. Furthermore, the counterfactual reports generally higher Gini coefficients than the decomposition exercise. The biggest difference between the counterfactual Gini and the decomposed Gini is observed in Wave 1. The counterfactual Gini reported is 0.67 whereas the decomposed Gini reports a value of 0.644. The differences in the consecutive Waves are more narrow and the smallest difference is reported in Wave 3 where the counterfactual Gini reports a value of 0.6456 and the decomposed Gini states a level of inequality of 0.6408. Overall, the differences between the counterfactual and the observed Gini coefficients are small. However, the fact that the levels of the counterfactual Gini are generally higher than present values means that inequality levels would be higher, had there been no remittance payments. This result supports the finding from Table 3 that a marginal change in income from remittances would lead to a (slight) decrease in the overall Gini. The results of the decomposition may also explain why the difference between the counterfactual and the decomposed Gini is more substantial for Wave 1. The decomposition exercise revealed that in this wave, remittances have a relatively larger share in overall household income compared to the other waves. Furthermore, the elasticity reported is also slightly higher in Wave 1.

Finally, Table 9 reports the result of the level of inequality present if remittances were simply set to zero. This paper has argued that there is possibility for bias if a non-remittance counterfactual would not take into account the different characteristics of migrants. Furthermore, the counterfactual created in this paper has tried to account for the fact that households that do not receive remittances will compensate the lack of income from this source by increasing income from other sources. The levels of inequality reported for the scenario where remittances are set to zero stress the necessity to account for these biases. The Gini coefficients increase from Wave 1 where the Gini was 0.656 to 0.689 in Wave 2 only to decrease in Wave 3 to 0.648. This shows, that the Gini coefficient in this scenario converges towards but is still slightly larger than the decomposed Gini in Wave 3. The trend that the Gini increases in Wave 2 is surprising although the decomposed Gini also increases mildly in Wave 2. Nevertheless, the differences between the case of remittances being set to zero and the counterfactual created in this paper stress that the households' behaviour of diversifying income towards other income sources should not be underestimated.

Table 9: Gini coefficients in the Counterfactual Scenario - Estimated using different methods

Estimation Method	WAVE 1	WAVE 2	WAVE 3
Probit MATCHING	0.6787	0.6651	0.6456
LOGIT MATCHING	0.6787	0.6651	0.6456
KERNEL MATCHING	0.6787	0.6651	0.6456
Decomposition	0.6441	0.6482	0.6408
Remittances set to zero	0.6564	0.6896	0.6482

In summation, this section has shown the results of the decomposition of the Gini coefficient according to income sources. This has revealed the strong relative contribution of income from labour market activities to the overall Gini as well as the strongly decreasing effect of government grants on total inequality. However, the decomposition exercise also showed the decreasing effects of remittances on overall inequality. This was supported by the results of the estimation of a counterfactual scenario. This section has employed propensity score matching using logit and probit models on a conditional difference in difference matching estimator in order to determine non-remittance households that are similar to remittance receiving households. Variables used in the creation of this counterfactual are significant and robust across different matching methods applied. Instrumentalizing propensity score matching enabled the estimation of household income excluding remittances in a counterfactual scenario without remittance payments. These estimations were then imputed in the distribution of household incomes and the counterfactual level of inequality was calculated. This has revealed that without remittances, there would be higher levels of inequality prevalent than the current.

6 Conclusion

This paper aimed at analyzing the effects that remittances from internal migration in South Africa have on inequality. The NIDS panel data set offers comprehensive information on household incomes from different sources and thus the ground for the analysis completed in this paper. The information on different components of household income was used to decompose the level of inequality prevalent in South Africa with respect to different income sources. This decomposition featured not only the examination of the Gini coefficients of the different income components but also the analysis of the role of income from different sources in overall inequality. As such, the decomposition according to the Lerman and Yitzhaki approach (1985) enabled the assessment of the effect of remittances according to three important factors. Firstly, the share of remittances in overall household income could be analyzed. Secondly, the level of inequality prevalent within remittances was discussed and lastly, the correlation of remittances with the overall Gini coefficient was estimated. However, not only remittances but income from different components could be analyzed according to these three criteria. Furthermore, the decomposition of the Gini coefficient allowed to assess the effect of a marginal change in one income component on the overall level of inequality. To analyze the role of remittances in this way is necessary as the role of remittances is not known a priori. Remittances may have inequality reducing or increasing effects depending on where in the income distribution the remittance receiving household is located.

The decomposition of household income performed in this paper showed that remittances have a potentially decreasing effect on inequality for South Africa, however, this effect appears to be rather small and is dominated by largely negative effects that stem from the high levels of inequality prevalent in income from the labour market. The paper finds that overall levels of inequality are stagnating over the period of time that is covered by the NIDS data set. Whereas in 2008 a Gini coefficient of 0.6441 was reported, South African inequality shows a Gini of 0.6408 in 2012. This confirms findings of Leibbrandt and Finn (2012) who characterized levels of inequality in South Africa as remaining "stubbornly high". However, for the first time in the post-apartheid era, levels of inequality are slightly decreasing. Leibbrandt et al. (2010) show in their decomposition of income that inequality had increased since 1993. From the decompositions results presented in this paper it seems as though the increased increase in the share that government grants make up in overall household income is one of the driving forces behind the decrease in overall inequality. The decomposition in this paper shows that a marginal increase in state transfers of 1 percent can lead to a decrease in the Gini coefficient of about 0.11 percent. This finding, together with the increasing role of state transfers in the composition of household income has certainly influenced the development of inequality levels in a positive manner. However, the decomposition exercise also shows the inequality decreasing effect of remittances. While much smaller, a marginal increase in remittances may also lead to a decrease of 0.01 percent in income inequality, the positive effect of remittances is supported by the counterfactual constructed in this paper.

The counterfactual was meant to assess the income distribution without remittances. Simply analyzing household incomes without remittances meanwhile holding income from other sources constant, however, would lead to biased results. Therefore, this paper implemented propensity score matching in the form of difference in difference matching and several different matching algorithms to determine those non-remittance households that are comparable to remittance receiving households. Smith and Todd (2005) have shown that difference in difference matching estimators are suitable to reduce bias in estimations given that the assumptions for this matching estimator are satisfied by the data. This paper has shown that conditional difference in difference matching is the best alternative to reduce bias in the estimation of a non-remittance counterfactual despite certain shortcomings. While the NIDS data satisfies the assumptions for difference in difference matching estimations, the data available on remittance payers does not satisfy the exclusion restriction for alternative estimations such as a double-selection Heckman (1979) model. The coefficients obtained from those households that were matched using conditional difference in difference estimators were then used to predict income excluding remittances in the counterfactual scenario of no remittances. As a result of the method applied in this paper, the counterfactual income distribution is more unequal for all three waves. Furthermore, the results of the counterfactual scenario are strongly robust against different matching methods, which supports the good quality of the matching achieved. Higher levels of inequality present in the counterfactual income distribution are supporting the previous finding that remittances indeed act as a reduction on inequality. However, there is an overall trend of decreasing inequality apparent in the Gini coefficients of the counterfactual. Finally, the comparison of the counterfactual results with the levels of inequality present if remittances were simply set to zero shows that with the exception of Wave 1, the latter methods leads to (much) higher results for the Gini coefficient than the counterfactual. These results support the assumption that households divert to other sources of income in case of a fall-out from one income component such as remittances. Through the method applied to estimate the counterfactual scenario in this paper it was possible to account for this behaviour while the alternative of setting remittances to zero was not able to do so.

All in all the decomposition of household income into different sources of income has shown the strong positive effects of government grants. However, given this strong effect of state transfers, the government has to be cautious to avoid overly strong dependency of households on government grants as an income source and should strengthen alternative ways to create income. Furthermore, the positive effects of government grants is dominated by the strong negative effect of the largest income component, income from labour market activities. The incredibly high levels of inequality within this income component and its large contribution to the overall Gini stress how urgently the high levels of inequality in labour market income in South Africa need to be addressed. This demand is supported by the results of the counterfactual analysis that have confirmed the decreasing effect of remittances from internal (work) migrants on inequality. Thus, a possible way to reduce inequality could be to ease ways for worker to move as well as improve channels through which remittances can be sent home. Given the inequality reducing effect of remittances, such policies may help to decrease the persistently high levels of inequality. However, such policies may not have the desired outcome of an inequality reduction given that remittances only have a small positive effect on inequality. More urgently, the high level of inequality within labour income needs to be addressed. That calls for policies that lead to job creation to tackle high levels of unemployment within the country or to make room for entrepreneurship by improving channels to create (formal) self-employment. As pointed out in Section 2.1, literature has discussed entrepreneurial potential of former migrants as well as remittances being used to start own businesses. In a South African context, these possibilities for the use of remittances need to be analyzed in order to address inequality well-considered. Hence, future research may look into the analysis of the use of remittances in South Africa. Other topics brought up in this paper and necessary to investigate further may include a decomposition of the no-remittance counterfactual. Such a decomposition would allow for a deeper understanding of how the loss of income from remittances may be compensated for. Furthermore, the role of other sources of income that have been discussed briefly in this paper, such as income from government grants or labour market income, could be analyzed more closely to dissect the reasons for persistently high inequality more closely. The high levels of income inequality in South

Africa have been discussed thoroughly in this paper with regards to one income source, remittances. In the future, there is room however, to analyze further income sources in order to tackle South African inequality conclusively.

Notes

[1]The Kuznet curve generally describes the initial worsening of the degree of inequality during the developing stage of an economy which then relaxes into lower levels of inequality, illustrated as an inverted u-shape.

[2]NIDS provides assistance with derived variables to allow the researcher better insight. A set of Stata Do-Files can be downloaded from [http://www.nids.uct.ac.za/nids-data/program-library/derived-files]. Extreme outliers are controlled for.

[3]According to Argent (2009), problems of asking for the market value of a dwelling may be caused by people confusing the question "How much rent would you pay?"that is supposed to be capturing the market value of that dwelling with how much they are willing to pay for it. This may cause a bias in the values stated, however, there is hardly a way to correct for this.

[4]Decomposition computed making use of Van Kerm, P. (2009), "sgini Generalized Gini and Concentration coefficients (with factor decomposition) in Stata", v1.1, CEPS/INSTEAD, Dierdange, Luxembourg.

[5]Propensity score matching was completed using Leuven, E. and Sianesi, B. (2014)"PS-MATCH2: Stata module to perform full Mahalanobis and propensity score matching, common support graphing, and covariate imbalance testing", Statistical Software Components.

References

[1] Acosta, P., Calderon, C., Fajnzylber, P., and Lopez, H. What is the impact of international remittances on poverty and inequality in Latin America? *World Development 36*, 1 (2008), 89 – 114.

[2] Adams, R. Worker remittances and inequality in rural Egypt. *Economic Development and Cultural Change* (1989), 45 – 71.

[3] Adams, R. *The effects of international remittances on poverty, inequality, and development in rural Egypt*, vol. 86. Intl Food Policy Res Inst, 1991.

[4] Argent, J. Household Income: Report on NIDS Wave 1. Tech. Rep. No. 3, N.I.D.S., July 2009.

[5] Barham, B., and Boucher, S. Migration, remittances, and inequality: estimating the net effects of migration on income distribution. *Journal of Development Economics 55*, 2 (1998), 307 – 331.

[6] Beine, M., Docquier, F., and Rapoport, H. Brain Drain and Human Capital Formation in Developing Countries: Winners and Losers. *The Economic Journal 118*, 528 (2008), 631–652.

[7] Borjas, G. Immigration and self-selection. *NBER Working Paper Series*, 2566 (1991), 29 – 76.

[8] Brown, M., DeVilliers, L., Woolard, I., Daniels, R., and Leibbrandt, M. Wave 3 user manual. *National Income Dynamics Study 1.1* (2013), 1–43.

[9] Caliendo, M., and Kopeinig, S. Some practical guidance for the implementation of propensity score matching. *Journal of economic surveys 22*, 1 (2008), 31 – 72.

[10] Cowell, F. Chapter 2: Measurement of inequality. *Handbook of Income Distribution 1* (2000), 87–166.

[11] Dehejia, R., and Wahba, S. Propensity score-matching methods for nonexperimental causal studies. *Review of Economics and statistics 84*, 1 (2002), 151 – 161.

[12] DeMaio, F. Income inequality measures. *Epidemiol Community Health 61* (2007), 849–852.

[13] Ellison, G. Letting the gini out of the bottle? challenges facing the relative income hypothesis. *Social Science & Medicine 54* (2002), 561–576.

[14] Girma, S., and Görg, H. Evaluating the foreign ownership wage premium using a difference-in-differences matching approach. *Journal of International Economics 72*, 1 (2007), 97 – 112.

[15] Grogger, J., and Hanson, G. Income maximization and the selection and sorting of international migrants. *Journal of Development Economics 95*, 1 (2011), 42 – 57.

[16] Ham, J. C., Reagan, P. B., and Li, X. Propensity score matching, a distance-based measure of migration, and the wage growth of young men. *Federal Reserve Bank of New York Staff Reports*, 212 (2005).

[17] Heckman, J., Ichimura, H., Smith, J., and Todd, P. Characterizing selection bias using experimental data. *Econometrica 66*, 5 (1998), 1017 – 1098.

[18] LaLonde, R. Evaluating the econometric evaluations of training programs with experimental data. *The American Economic Review* (1986), 604 – 620.

[19] Leibbrandt, M., Woolard, I., Finn, A., and Argent, J. Trends in South African income distribution and poverty since the fall of apartheid. *OECD Social, Employment and Migration Working Papers No. 101* (2010).

[20] Lerman, R., and Yitzhaki, S. Income inequality effects by income source: a new approach and applications to the United States. *The Review of Economics and Statistics* (1985), 151 – 156.

[21] McKenzie, D., and Rapoport, H. Network effects and the dynamics of migration and inequality: Theory and evidence from Mexico. *Journal of Development Economics 84* (2007), 1 – 24.

[22] Mesnard, A. Temporary migration and capital market imperfections. *Oxford Economic Papers 56*, 2 (2004), 242 – 262.

[23] Perotti, R. Political equilibrium, income distribution, and growth. *The Review of Economic Studies 60*, 4 (1993), 755 – 776.

[24] Posel, D. Households and labour migration in post-apartheid South Africa. *Studies in Economics and Econometrics 34*, 3 (2010), 129 – 141.

[25] Rapoport, H., and Docquier, F. The economics of migrants' remittances. *Handbook of the economics of giving, altruism and reciprocity 2* (2006), 1135 – 1198.

[26] Rosenbaum, P., and Rubin, D. The central role of the propensity score in observational studies for causal effects. *Biometrika 70*, 1 (1983), 41 – 55.

[27] ROY, A. D. Some thoughts on the distribution of earnings1. *Oxford Economic Papers 3*, 2 (1951), 135 – 146.

[28] SALDRU. National Income Dynamics Study 2008, Wave 1 [dataset]. *Southern Africa Labour and Development Research Unit [producer]* (2009). Version 5.2.

[29] SALDRU. National Income Dynamics Study 2010 - 2011, Wave 2 [dataset]. *Southern Africa Labour and Development Research Unit [producer]* (2014). Version 2.2.

[30] SALDRU. National Income Dynamics Study 2012, Wave 3 [dataset]. *Southern Africa Labour and Development Research Unit [producer]* (2014). Version 1.2.

[31] SMITH, J., AND TODD, P. Does matching overcome LaLonde's critique of nonexperimental estimators? *Journal of Econometrics 125*, 1 (2005), 305 – 353.

[32] STARK, O., TAYLOR, E., AND S.YITZHAKI. Remittances and Inequality. *The Economic Journal, 96* (1986), 722 – 740.

[33] VAN KERM, P. sgini - generalized gini and concentration coefficients (with factor decomposition) in stata. *CEPS/INSTEAD, Luxembourg* (2009). Version 1.1.

Word Count: ca. 23000

7 Appendix

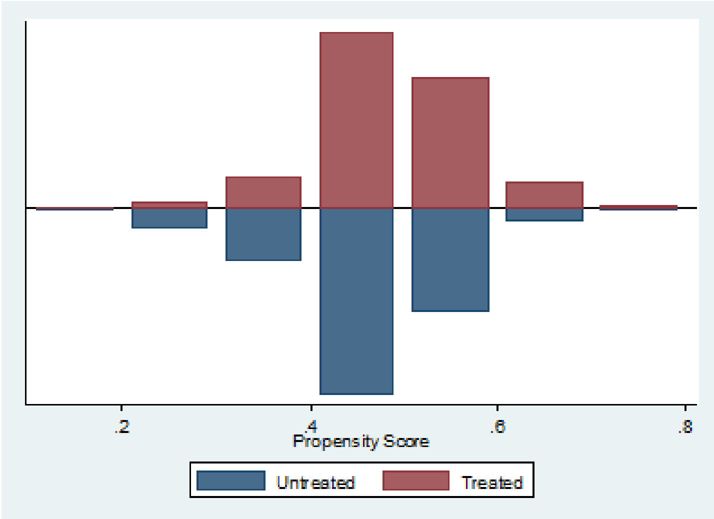

Figure 3: Kernel Matching - Rural Households

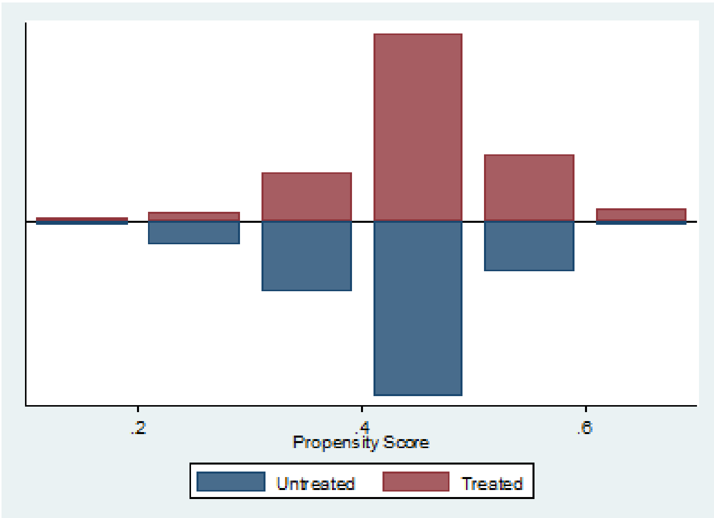

Figure 4: Kernel Matching - Urban Households

Table 10: t-Test after Kernel Matching - Rural Households

Variable	Sample	Mean Treated	Mean Control	%bias	%reduct. bias
ln_PRIM1	Unmatched	7.6174	7.513	11.7	
	Matched	7.6174	7.6119	0.6	94.7
hhsizer1	Unmatched	6.8904	5.918	29.1	
	Matched	6.8904	6.7866	3.1	89.3
meanAGE1	Unmatched	23.48	26.268	-28.0	
	Matched	23.48	23.551	-0.7	97.5
adultage1	Unmatched	3.0541	2.8515	12.5	
	Matched	3.0541	3.0332	1.3	89.7
hhedu1	Unmatched	10.22	9.7146	18.4	
	Matched	10.22	10.212	0.3	98.4
African	Unmatched	.94705	.90515	16.1	
	Matched	.94705	.94547	0.6	96.2
Coloured	Unmatched	.0224	.07026	-22.9	
	Matched	.0224	.02527	-1.4	94.0
Asian_Indian	Unmatched	.01103	.00698	4.3	
	Matched	.01103	.01017	0.9	78.7
White	Unmatched	.00458	.00727	-3.5	
	Matched	.00458	.0041	0.6	82.1
Pseudo R2	Unmatched	0.033	Matched	0.000	

Table 11: t-Test after Kernel Matching - Urban Households

		Mean			%reduct
Variable	Sample	Treated	Control	%bias	bias
ln_PRIM1	Unmatched	8.0913	8.0701	2.0	
	Matched	8.0913	8.0769	1.4	32.0
hhsizer1	Unmatched	5.7691	4.8921	30.8	
	Matched	5.7691	5.6497	4.2	86.4
meanAGE1	Unmatched	26.039	27.988	-18.5	
	Matched	26.039	25.977	0.6	96.8
adultage1	Unmatched	3.0967	2.7545	22.3	
	Matched	3.0967	3.0766	1.3	94.1
hhedu1	Unmatched	11.08	10.803	11.6	
	Matched	11.08	11.066	0.6	94.9
African	Unmatched	.67755	.65966	3.8	
	Matched	.67755	.67913	-0.3	91.1
Coloured	Unmatched	.26847	.24986	4.2	
	Matched	.26847	.27055	-0.5	88.8
Asian_Indian	Unmatched	.00883	.01521	-5.8	
	Matched	.00883	.00928	-0.4	93.0
White	Unmatched	.0292	.05894	-14.5	
	Matched	.0292	.026	1.6	89.2
Pseudo R2	Unmatched	0.022	Matched	0.001	

Table 12: t-Test after Logit Matching - Rural Households

Variable	Sample	Mean Treated	Mean Control	%bias	%reduct bias
ln_PRIM1	Unmatched	7.6174	7.513	11.7	
	Matched	7.6174	7.6002	1.9	83.5
hhsizer1	Unmatched	6.8904	5.918	29.1	
	Matched	6.8904	6.5104	11.4	60.9
meanAGE1	Unmatched	23.48	26.268	-28.0	
	Matched	23.48	23.304	1.8	93.7
adultage1	Unmatched	3.0541	2.8515	12.5	
	Matched	3.0541	2.8555	12.2	2.0
hhedu1	Unmatched	10.22	9.7146	18.4	
	Matched	10.22	10.162	2.1	88.6
African	Unmatched	.94705	.90515	16.1	
	Matched	.94705	.92346	9.0	43.7
Coloured	Unmatched	.0224	.07026	-22.9	
	Matched	.0224	.0205	0.9	96.0
Asian_Indian	Unmatched	.01103	.00698	4.3	
	Matched	.01103	.02234	-12.0	-179.2
White	Unmatched	.00458	.00727	-3.5	
	Matched	.00458	.00411	0.6	82.3
Pseudo R2	Unmatched	0.033	Matched	0.007	

Table 13: t-Test after Logit Matching - Urban Households

Variable	Sample	Mean Treated	Mean Control	%bias	%reduct bias
ln_PRIM1	Unmatched	8.0913	8.0701	2.0	
	Matched	8.0913	8.0849	0.6	69.9
hhsizer1	Unmatched	5.7691	4.8921	30.8	
	Matched	5.7691	5.7215	1.7	94.6
meanAGE1	Unmatched	26.039	27.988	-18.5	
	Matched	26.039	26.351	-3.0	84.0
adultage1	Unmatched	3.0967	2.7545	22.3	
	Matched	3.0967	3.1136	-1.1	95.1
hhedu1	Unmatched	11.08	10.803	11.6	
	Matched	11.08	11.132	-2.2	81.2
African	Unmatched	.67755	.65966	3.8	
	Matched	.67755	.69144	-3.0	22.4
Coloured	Unmatched	.26847	.24986	4.2	
	Matched	.26847	.25934	2.1	50.9
Asian_Indian	Unmatched	.00883	.01521	-5.8	
	Matched	.00883	.00957	-0.7	88.4
White	Unmatched	.0292	.05894	-14.5	
	Matched	.0292	.02297	3.0	79.0
Pseudo R2	Unmatched	0.022	Matched	0.001	

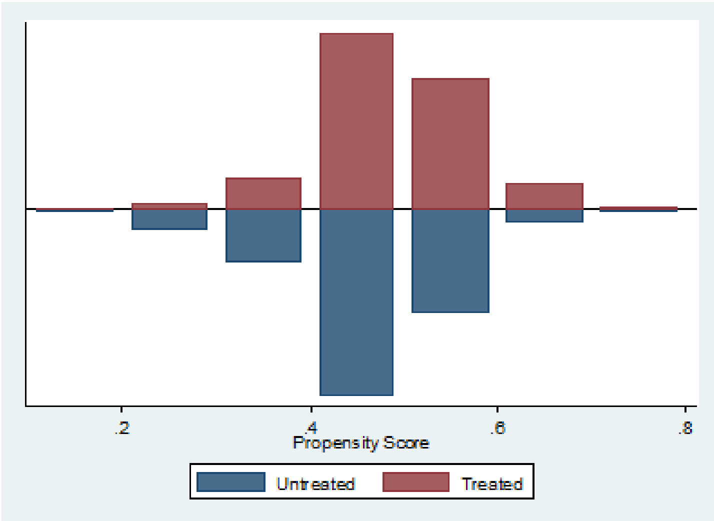

Figure 5: Nearest Neighbour Logit Matching - Rural Households

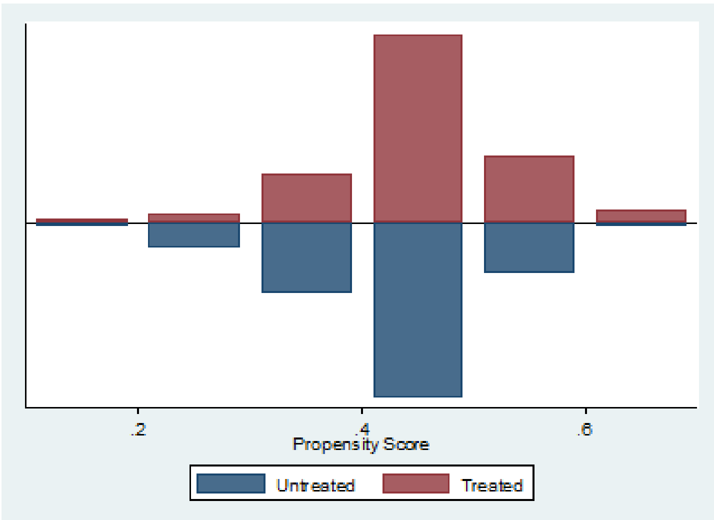

Figure 6: Nearest Neighbour Logit Matching - Urban Households