ESSAYS ON SOCIOECONOMIC INEQUALITY

A DISSERTATION SUBMITTED TO THE GRADUATE DIVISION OF THE UNIVERSITY OF HAWAIʻI AT MĀNOA IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

PHD

IN

ECONOMICS

MAY 2016

By

Jonathan R. Page

Dissertation Committee:

Timothy Halliday, Chairperson

Sumner La Croix
Inessa Love
Jeffrey Traczynski
Susanne Still

Keywords: income, educational attainment, sibling correlation, clustering

ACKNOWLEDGMENTS

I thank my committee for their invalualable comments. This dissertation is immeasurably better because of them. I thank my classmates for their support and suggestions. I thank my wife for giving me the time to work. The remaining errors and abuse of the English language are my own.

ABSTRACT

This dissertation studies several aspects of socioeconomic inequality. Chapter 1 compares the relative importance of inequality across neighborhoods in explaining overall inequality among black South Africans using variance decomposition methods. I find a statistically significant difference between the neighbor correlation in former bantustans and the former white South Africa. The neighborhood-level average consumption of black households in the former white South Africa exhibits a level of variation that is an order of magnitude greater than in the former bantustans. Urban status, region, and household-level education explain some of this difference, but this difference remains largely unexplained.

Chapter 2 replicates previous findings on sibling and neighbor correlations in educational attainment using the Panel Study of Income Dynamics. The replication establishes comparability across multiple methods and forms a solid foundation from which to compare the developed and developing world. Using dataset comparable to these foundational works, I show sibling and neighbor correlations in Indonesia closer to the estimates in the developed world. The similarity of correlations indicates a similar level of the importance of the family relative to outside factors. However, the variance of years of schooling in Indonesia is much higher than in the US. That is, the educational significance of the family and neighborhood effect is much higher in Indonesia than in the US.

Chapter 3 is the first attempt to group individual age-income profiles by the similarity of the profiles. Since many factors likely correlate with any chosen demographic feature, starting with emergent lifecycle patterns permits a more thorough analysis of likely causes. For example, previous papers compared income profiles across education groups, while this paper compares educational attainment across income profile groups. I find that individ-

uals with *high-growth* or *hump-shaped* age-income profiles have higher levels of education and lifetime income than those with *steady-declining* or *early-U* age-income profiles.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

THE STABILITY OF SEPARATION: NEIGHBORHOODS AND CONSUMPTION INEQUALITY AMONG BLACK HOUSEHOLDS IN POST-APARTHEID SOUTH AFRICA

1.1 Introduction

South Africa has the dubious honor of having exceptionally high inequality in absolute terms and given its GDP per capita (see figure B.1). In 1994 this fact would be unsurprising given the policies of the Apartheid regime which had maintained white control of South Africa through forced relocations and influx controls. Unfortunately, this high level of inequality has persisted in the post-Apartheid period without the accompanying growth advantage suggested by Kuznets (1955) (see figure B.2).[1]

Forced relocations by the Apartheid regime between the 1960s and the mid 1980s moved an estimated 20% of the black African native population (Platzky and Walker, 1985). Entire neighborhoods were uprooted and trucked to remote areas in the former reserves setup by the white settlers. These reserves were called bantustans (and later labeled as "Homelands" by the Apartheid regime in an attempt to legitimize their discriminatory policies) and were a tool of the Apartheid regime to revoke black citizenship in "white South Africa." They were used to ethnically and linguistically divide the majority black population to facilitate white-minority rule following the age-old maxim "divide and rule." Removing poor and informal black settlements from land desired by the "white" minority exacerbated the differences between "white South Africa" and the bantustans.[2] The "black spots," areas identified for removal by the Apartheid regime, tended to contain poor residents that supplemented their infrequent wage income with subsistence farming. Not all

[1] While unfortunate, as noted and modeled in Acemoglu and Robinson (2008), institutions are resilient to changing political climates.

[2] Many "black spots" that were removed had significant coal deposits (Platzky and Walker, 1985).

"black spots" were removed, but wealthy whites who had the most to gain were able to push their pet projects through.[3]

By removing a relatively poor neighborhood, the removal of a "black spot" raises the region's average income. These removals raised the relative importance of inequality across neighborhoods in explaining overall inequality (i.e., the between-neighborhood variance) in "white South Africa" and lowered it in the bantustans.[4] The divergent effects of forced relocation erodes the ability of provincial (or national) inequality measures to adequately describe income disparities in South Africa.

Households experience inequality at multiple levels. Inequality among households within neighborhoods may affect household decision making differently than inequality among neighborhoods within a region. A better understanding of the way inequality affects household decision-making will help identify the welfare effects of redistribution. An important step in this analysis is a thorough description of the existing multilevel structure of inequality.

This paper seeks to address the lack of evidence on the multilevel structure of inequality in South Africa. I decompose the variance of household consumption into household and neighborhood components to identify the relative importance of these components on a measure of inequality. I show that the between-neighborhood variance (i.e., the neighborhood component) explains explains a larger portion of the overall variance of consumption in former "white South Africa" than in the former bantustans.[5] I then test several candidate explanations for this difference and find that the household's urban status, magisterial district, and level of education are statistically significant explanatory factors for this difference. However, even after controlling for these factors, a statistically significant difference remains.

[3](Platzky and Walker, 1985) report that batustan governments also forced small farmers off their land in favor of "prestige projects," though to a much smaller extent than the forced removals perpetrated by the Apartheid regime.

[4]To see why the between-neighborhood variances would diverge in response to forced removals.

[5]I find a statistically significant difference between the neighborhood component in former "white South Africa" and the former bantustans

1.2 Spatial Inequality

There is a large literature on spatial inequality in both developed and developing economies.[6] This literature largely depends on the additive decomposability of the Theil L and Theil T inequality measures, which are generalized entropy measures of dispersion. Using these measures, between-county inequality explains between 40 and 50 percent of the overall household-level inequality in rural China (Gustafsson and Shi, 2002).[7] Between-district inequality explains between 34 and 36 percent of overall household-level inequality in Vietnam (Minot et al., 2006). For rural areas in Vietnam, between-district (between-commune) inequality explains 12 (14) percent of overall inequality (Epprecht et al., 2009). In Mozambique, between-group inequality for small administrative units explains between 25 and 29 percent of overall inequality (Simler et al., 2005).[8]

Elbers et al. (2005) combine census and household survey data to calculate inequality measures for small areas.[9] They find between-neighborhood inequality in per capita expenditures in Ecuador explained 14% of overall household-level inequality for rural areas and 23% in urban areas. In Madagascar, between-neighborhood inequality explained 18% of overall inequality in rural areas and 23% in urban areas. The former "white South Africa" is more developed and relatively more urban than the former bantustans. Thus, I expect the neighborhood component to be relatively smaller in the bantustans.

Additively decomposable inequality measures are affected by the overall population size and the relative groups sizes. This size effect complicates cross-population and cross-partition comparisons. To resolve this issue, Elbers et al. (2008) introduce a between-group inequality measure that indicates the proportion of the maximum possible between-group inequality. They produce estimates for several countries and various population groups. Though they do not explore population partitions based on geography, they do include estimates for South Africa based on racial partitions. They find that between-group inequality is relatively greater between Whites and non-Whites than for any other partition based on race, ethnicity or language. This suggests my focus on Black South Africans does not ig-

[6] See Shorrocks and Wan (2005) for a survey of spatial inequality research. Novotný (2007) tabulate the between-group inequality estimates for a large set of developed and developing world studies.

[7] See Gustafsson and Shi (2002) for a review of spatial inequality studies focused on China.

[8] In Simler et al. (2005), the small administrative units are smaller than an administrative post, but larger than a village or city block. They are called *localidade* in rural areas and *bairro* in urban areas.

[9] This method of small area estimation was introduced in Elbers et al. (2003). See Tarozzi and Deaton (2009) for a critique of the method's requirement that areas be homogeneous.

nore major inequalities between ethno-linguistic partitions of Black South Africans. The main inequality measures presented in this paper avoid the comparability issue of the additively decomposable measures since they are independent of the overall population size and the relative groups sizes.

An alternative approach is to directly measure the proportion of the variance explained by differences across groups. This approach of calculating the intraclass correlation[10] is used in Gräb and Grimm (2009) to measure the community effect on overall inequality in Burkina Faso using cross-sectional household surveys. They find that variance across communities explains between 21 and 27 percent of overall variation in income. The current study expands on this research by using household-survey panels in South Africa and tools from the sibling correlation literature to remove the transitory component of household income.

1.3 Post-Apartheid South Africa

The modern state of South Africa was shaped by the white-supremacist government that controlled it for much of the twentieth century. The 1913 Land Act, the 1950 Group Areas Act, and the era of Grand Apartheid from 1960-1985 evicted Black South Africans from desirable, or at least racially integrated, land, resettled them on reserves, and revoked their South African citizenship. These reserves were called bantustans and later "homelands", which in some cases were given nominal independence by the Apartheid regime.[11] This racial separation resulted in a separation of areas that is more distinct than the typical rural-urban divide in transition economies. The former "White South Africa", especially in urban areas, resembles a developed economy, while the former bantustans resemble developing economies.

In 1994 the election of Nelson Mandela, the black South African resistance leader who had spent 27 years in prison, signaled to the world the end of Apartheid. This new birth of economic and political freedom held great promise. Unfortunately, since 1994 inequality and headcount measures of poverty increased (cf., Özler 2007). More households were below 200% of the household subsistence line (HSL) in 1998 than in 1993 when comparing the income distribution in KwaZulu-Natal across these two periods. A review of the joint

[10] This is the same method used in the sibling correlation literature discussed below.

[11] Unsurprisingly, this "independence" was never internationally recognized.

distribution of income in 1993 and 1998 show the poor falling ever more behind (Carter and May, 2001). Transition matrices, whether using endogenous income quintiles (Woolard and Klasen, 2005) or exogenous income groups based on percent of HSL (Carter and May, 2001) reveal a society with significant mobility, though much of that mobility entails poor households becoming poorer.

Through 2007 the former bantustans were the most deprived in terms of income, employment, education, and living environment according to the South African Index of Multiple Deprivation (Noble and Wright, 2012). This separate development can be seen in figure B.3 which shows the lingering spatial effect of Apartheid in the KwaZulu-Natal province as of 2001. The development literature offers initial household size, education, asset endowment, employment access (Woolard and Klasen, 2005), and a highly segmented labor market (Özler, 2007) as reasons for the increase in socioeconomic inequality. Here I add to this literature by examining the importance of the neighborhood on the measurement and understanding of income inequality.[12] Alderman et al. (2002) measure the importance of the neighborhood effect at the provincial level in South Africa using 1995 and 1996 household surveys. They find the neighborhood effect is strongest in KwaZulu-Natal explaining more than half of the variation across households. This study extends their analysis by separating the former bantustans from the provinces which contain them and by controlling for variation over time.

1.4 Statistical Model

The model of household income employed here has been alternately referred to as a nested-error component model, a random effects model, a multilevel model, a mixed model, a variance components model, or a hierarchical model (see Snedecor and Cochran, 1980; Deaton, 1997).[13] The notation here mirrors that found in studies of sibling correlations

[12] I am aware of one paper with a similar approach. Gräb and Grimm (2009) examine the spatial inequality of income in Burkina Faso. For a look at community effects on sub-Saharan African education and fertility see Kravdal (2002). Community effects are those factors of income common to households in a community such as role models, social connections, exposure to violence, and discrimination.

[13] Montmarquette and Mahseredjian (1989) use a two-way nested-error component model to study the impact of a student's class and school on their educational achievement. Antweiler (2001) provides a succinct history of nested error models and discusses an application estimating the variances with maximum likelihood (ML).

(Solon, 1999). The natural logarithm of adult-equivalent[14] monthly expenditure, y_{ch}, for cluster c and household h is modeled as

$$y_{ch} = \mathbf{x}'_{ch}\beta + \varepsilon_{ch}. \tag{1.1}$$

$\mathbf{x}'_{ch}\beta$ includes an intercept, the number of children and the number of pensioners in order to control for key household life-cycle effects.[15] The residual, ε_{ch}, represents the effects of household-specific factors unrelated to neighborhood factors. I decompose ε_{ch} as follows:

$$\varepsilon_{ch} = a_c + v_{ch}, \tag{1.2}$$

where a_c is the component common to all households in community c, v_{ch} is the idiosyncratic component for household h.

By construction, the variance, σ^2_ε, equals

$$\sigma^2_\varepsilon = \sigma^2_a + \sigma^2_v. \tag{1.3}$$

Thus, the share of variance in income due to community background, and also the income correlation of randomly drawn pairs of households in a given community is

$$\rho = \frac{\sigma^2_a}{\sigma^2_a + \sigma^2_v} \tag{1.4}$$

[14]Here I use the adult equivalent scale, ϕ, used by Carter and May (1999) and common throughout the literature on South African household income. $\phi = (A + 0.5K)^{0.9}$, where A is the number of adults and K is the number of children. This structure reflects children's lower consumption relative to adults and assumes economies of scale.

[15]This follows the covariate setup in Solon et al. (1991) and Mazumder (2008) with relevant changes for this paper's household setting.

1.5 Estimation Procedure

I follow Mazumder (2008) and estimate the variance components using restricted maximum likelihood (REML).[16] While the ANOVA approach[17] to calculate ρ is straightforward and provides minimum variance estimator for balanced clusters, the same is not true for unbalanced clusters (Corbeil and Searle, 1976). Solon et al. (1991) introduce four weighting schemes to test robustness of results to various corrections for this imbalance. Another approach is using Theil's measure of inequality exploiting its additive decomposability.[18] REML has the advantage, even in the unbalanced case, of consistency, asymptotic normality, and a known asymptotic sampling dispersion matrix. Simulations by Browne and Draper (2006) indicate that bias is likely to be low when using REML for the number of clusters and households used in this study given the assumption that the log of adult equivalent expenditure is normally distributed.[19] Since the household survey data is unbalanced (i.e., each cluster is not restricted to the same number of households), I select REML as the preferred method in this case to estimate ρ (Mazumder, 2008).

1.5.1 Comparing regions

For each of the variance components and the neighbor correlation, I test for statistically significant differences between KwaZulu and Natal, and between tribal and non-tribal areas. After finding significant differences, I explore whether several candidate covariates can explain these differences. To do this, I recalculate equation (1.1), adding the observed variable (e.g., the mean years of education for resident adults) to X. Define the neighborhood-level variation from this new calculation σ_a^{2*}.

[16] This method is alternately referred to as residual maximum likelihood. I estimate REML through the *xtmixed* command in Stata.

[17] In the analysis of variance (ANOVA) approach ICC is simply the ratio of the between subjects (here survey clusters) variance to the total variance.

[18] See appendix C for a comparison of the neighbor correlations calculated using the Theil method.

[19] Following the relevant literature, I make this assumption. I produce quantile-quantile (Q-Q) plots for KwaZulu and Natal as a visual check of normality in figures B.4a and B.4b. Based on the Q-Q plots the distributions appear normal. The Shapiro-Wilk test for normality does not reject the null hypothesis that logged expenditure is normally distributed for the case of KwaZulu, though it does reject normality for Natal. This rejection of normality suggests that REML may be in fact be a biased estimate of the variances relative to ML. However, appendix D demonstrates that both procedures generate similar results.

I then measure the percent change in the differences due to the addition of the relevant covariate, which indicates the ability of the covariate to explain some of the difference. To see how robust the differences are to inclusion of these covariates, I also present the remaining difference after covariate inclusion.

While it would be convenient to measure the σ_a^{2*} using REML, Corbeil and Searle (1976) and Robinson (1987) note that REML, in contrast to maximum likelihood (ML), includes degrees of freedom in the estimation of the variance components.[20] As a result, it is possible with REML to have $\sigma_a^{2*} > \sigma_a^2$ when adding additional fixed effects to the model.[21] Because of this issue, it seems wise to diverge from the procedure in Mazumder (2008) by using ML to calculate the contributions to ρ. I present the contributions using REML for consistency with Mazumder (2008), since in this case the difference between the results are negligible. I present the ML results corresponding to tables A.6 and A.9 in tables D.1 and D.2 as demonstration of how close the results are.

1.6 Data

To compare the patterns of inequality in the bantustans and "white South Africa," I use data from the KwaZulu-Natal Income Dynamic Study (KIDS) and the National Income Dynamic Study (NIDS). KIDS[22] is a panel household survey covering the KwaZulu-Natal province of South Africa that was conducted in 1993, 1998, and 2004.[23,24] The first round of KIDS formed part of the national Project for Statistics on Living Standards and De-

[20] See appendix D.

[21] In fact, this is the case when estimating the contributions presented later in this paper when using REML.

[22] This is the same data used by other studies of household income mobility covering this period in South Africa (e.g, Klasen, 2000; Leibbrandt and Woolard, 2001; Woolard and Klasen, 2005; May et al., 2007).

[23] The roughly 5-year gaps between observations satisfy the prescription from Naschold and Barrett (2011) that long periods of examination are needed to accurately measure structural mobility (as opposed to short-term fluctuations), while the gaps low frequency of the observations increases measurement error.

[24] The KwaZulu-Natal Income Dynamics Study (KIDS) was a collaborative project between researchers at the University of KwaZulu-Natal, the University of Wisconsin, London School of Hygiene & Tropical Medicine, International Food Policy Research Institute (IFPRI), the Norwegian Institute of Urban and Regional Studies and the South African Department of Social Development. In addition to support from these institutions, the following organizations provided financial support: UK Department for International Development; the United States Agency for International Development (USAID); the Mellon Foundation; and National Research Foundation/Norwegian Research Council grant to the University of KwaZulu-Natal.

velopment in 1993, which itself was based off of the World Bank Living Standards Measurement Surveys. White and coloured[25] households were excluded from the resurvey since they were deemed to be non-representative. Thus, the resurvey only includes African and Indian households. This study focuses solely on the African households. The KwaZulu-Natal regions was chosen for resurvey because of the relatively high quality of the location data.(May et al., 2000) In 1996, KwaZulu-Natal it had the largest population of any South African province with 8.4 million inhabitants, roughly 20.7% of the country's population (Africa, 1996).

The first KIDS survey in 1993 uses the Apartheid era province names. The current-day province of KwaZulu-Natal is composed of the 1993 "white South Africa" province of Natal and the bantustan KwaZulu. I use the 1993 province variable to identify which clusters were located within a bantustan.

NIDS is an ongoing panel household survey of the entire nation of South Africa with data available for waves in 2008, 2010/2011 and 2012 .[26] The NIDS questionnaires cover the same general topics that were covered in KIDS. A major advantage to the the NIDS surveys is their inclusion of all race classifications. I restrict my analysis to the African households and to those within the KwaZulu-Natal province. Unfortunately, NIDS does not contain data on whether or not a cluster was located within a bantustan under Apartheid. I use the NIDS indicator for tribal areas as a proxy for bantustan membership since the two categories are correlated.

[25]Coloured is a race classification used under Apartheid. The other classifications being white, African, and Indian.

[26]Wave 1 (2008) Data: Southern Africa Labour and Development Research Unit. National Income Dynamics Study 2008, Wave 1 [dataset]. Version 5.2. Cape Town: Southern Africa Labour and Development Research Unit [producer], 2009. Cape Town: DataFirst [distributor], 2014

Wave 2 (2010-2011) Data: Southern Africa Labour and Development Research Unit. National Income Dynamics Study 2010-2011, Wave 2 [dataset]. Version 2.2. Cape Town: Southern Africa Labour and Development Research Unit [producer], 2014. Cape Town: DataFirst [distributor], 2014

Wave 3 (2012) Data: Southern Africa Labour and Development Research Unit. National Income Dynamics Study 2012, Wave 3 [dataset]. Version 1.2. Cape Town: Southern Africa Labour and Development Research Unit [producer], 2014. Cape Town: DataFirst [distributor], 2014

1.7 Results

1.7.1 Neighbor Correlations

For the former bantustans, the variance component assigned to the household is larger than the neighborhood component, while for former "white South Africa" the opposite is true (see tables A.1, A.2 and A.3). As a result, the neighbor correlation significantly differs between these two historical administrative spheres.

Possible explanations for this difference include industry location and the forced relocations of entire neighborhoods. The Apartheid regime located industries (e.g., coal mining) within the bounds of "white South Africa." Where the regime could not choose the location of industries, it changed the borders of "white South Africa" to contain them. The concentration of places of employment within the boundaries of "white South Africa" permitted economies of agglomeration of laborers that was relatively absent in the former bantustans. Forced relocations themselves mechanically contribute to the difference in variance structure.

1.7.2 Contributing Factors

The proportion of household income derived from agriculture, the level of household education and neighborhood location in the region (i.e., the district location and urban status) are observables which may help explain the difference between the bantustans and former "white South Africa." Bantustan territory was selected from the most infertile land in South Africa as the Apartheid regime sought to consolidate the fertile land into large white-owned farms. The more dependent households are on agricultural output the greater we may expect the difference between these regions to be. Tables A.4, A.5, and A.6 show that the two location variables and household education measures explain a statistically significant percentage of the difference between these areas in the neighborhood component. However, statistically significant differences remain (see tables A.7, A.8, and A.9).

1.8 Conclusion

20 years after democratization a gap persists between consumption patterns for black South Africans in former bantustans and those in the former "white South Africa." Per capita expenditures are lower among blacks in former bantustans and community average expenditures are more homogeneous in the former bantustans.

Household-level education, urban status, and the magisterial district explain some of this difference, though a statistically significant gap persists between these two regions.

This persistent gap in the structure of household-level performance may be disappointing given the great promise of ending Apartheid for those black South Africans that had been marginalized and in some cases forcibly relocated.[27] This gap is reinforced by the continued existence of the tribal authorities, which were established to legitimate the Apartheid regime's creation of ethnic homelands through manipulation of tribal structures.[28] On one hand, the continued existence of tribal authorities provides certain government services to otherwise underserved remote areas. On the other hand, their continued existence reinforces historic inequalities.

To understand inequality in South Africa, a country that frequently ranks among the most unequal in the world, it is important to know how this inequality is composed. Real differences exist between tribal authority areas (the former bantustans) and non-tribal areas. Across these region types, the relative importance of neighborhoods in explaining consumption differences varies.

Reducing inequality continues to be an important and much debated subject of public policy in South Africa. This study helps build a more complex picture of the structure of inequality. Future work will need to identify mechanisms to reduce these inequalities.

[27] While disappointing, unchanging economic systems in the face of political change is nothing new and is modeled in detail by Acemoglu and Robinson (2008).

[28] The effects of colonial manipulation of traditional power structures is explored in Acemoglu et al. (2014).

CHAPTER 2

SIBLING AND NEIGHBOR CORRELATIONS IN EDUCATIONAL ATTAINMENT IN INDONESIA

2.1 Introduction

Sibling and neighbor correlations in education attainment measure the importance of factors common to a family or neighborhood. A high sibling or neighbor correlation indicates low socioeconomic mobility since relatively little is left to individual decisions or capabilities. In the developed world, Solon et al. (2000) estimate the sibling correlation in educational attainment for the US to be approximately 50%. They found a relatively small neighbor correlation that disappears after controlling for observable family factors. In the developing world, Emran and Shilpi (2015) estimate the sibling correlation in educational attainment in India to be 40% and neighbor correlations to be 20%. The methods employed vary across these studies and weakens cross-country and developed vs. developing world comparisons. This study replicates the procedure in Solon et al. and applies it to the developing world context. My results support the hypothesis that social mobility is lower and the neighborhood effect is stronger in the developing world than in the developed world.

In the developing world an understanding of educational attainment is important for the understanding of intergenerational wealth mobility. In Bangladesh, the primary channel explaining low intergenerational wealth mobility is low intergenerational mobility of years of schooling (Asadullah, 2012). By providing estimates of sibling and neighbor correlations in years of schooling in Indonesia, this paper contributes to the understanding of overall wealth mobility in the developing world.

There is a growing literature on neighborhood effects in developing countries. Neighbors influence farmer's choices on the adoption of new agricultural technologies in Indonesia (Case, 1992) and Ghana (Conley and Udry, 2010); the neighborhood affects children's schooling (Asadullah, 2011) and neighbors pool food security risk in Bangladesh (Park, 2006). This paper adds to the literature on neighborhood effects in the developing world

by extending the work of Solon et al. (2000) on sibling and neighbor correlations in educational attainment to Indonesia.

This paper extends the small literature in the developing world on sibling correlations. Emran and Shilpi (2015) study sibling correlations in years of schooling in India. Due to their large sample (nearly 40,000 siblings in 2006), they are able to separately analyze correlations across genders and urban/rural status. They estimate sibling correlations (pooling brothers and sisters) in educational attainment to be approximately 39% when controlling for neighborhood fixed effects.

To measure the neighbor correlation, Emran and Shilpi (2015) use the difference between the sibling correlations with and without the neighborhood fixed effects. Their estimate of the neighbor correlation in educational attainment, 20%, is larger than similar measures in the developed world.

This paper extends the relatively new literature on sibling and neighborhood correlations in the developing world. I provide estimates of these correlations in educational attainment using two of the established methodologies: the ANOVA approach used in the foundational work of Solon et al. (2000) and the REML approach used more recently by Mazumder (2008).[1]

2.2 Neighbor Correlations

The study of neighborhood effects on years of schooling has a long history. Datcher (1982) regresses years of schooling on family and neighborhood factors, but laments the limitation of using observed factors as proxies for the full group-level effects. Subsequent correlation studies attempt to measure these group-level effects directly. Solon et al. (2000) estimate the neighbor correlation in years of schooling is not statistically significant in the US using the PSID.

In the developed world, neighborhood correlations in other measures are also non significant. Oreopoulos (2003) report non significant neighborhood correlations in income, earning, and years on welfare for a variety of public housing projects in Toronto using the

[1] Mazumder (2008) also compared results from the ANOVA and REML approaches in the NLSY. This paper repeats this comparison for the dataset used in Solon et al. (2000), the PSID, and for the Indonesian Family Life Survey.

procedure used by Solon et al. (2000). Page and Solon (2003) find that urban and region dummies can explain the apparent neighbor correlation in earnings in the PSID.

To control for the observable family characteristics, Solon et al. (2000) used the years of schooling for the household heads and their spouses. Rosenzweig (2003) provides some guidance to measuring this characteristic in the Indonesian context. He finds the maximum years of schooling for adult household members in 1982 in Indonesia significantly affected the years of schooling for males in 2000 who were younger than 30 in 1982, controlling for household income and land ownership. Using data from the Ghana Living Standards Survey covering 1988-1989, Jolliffe (2002) explores the relative performance of household educational attainment proxies in Mincer-type wage equations. They reject the use of the household head's attainment or that of the adult with the least education in explaining education. They find some support for the use of the maximum and median years of education to proxy the household's level of education.[2]

There is a vast literature on sibling correlations in economic outcomes (surveyed by Solon, 1999 and Björklund and Jäntti, 2009) and a smaller literature on neighborhood correlations in economic outcomes (e.g., Solon et al., 2000; Page and Solon, 2003; Lindahl, 2011). These studies focus on the developed world. The only study of which I am aware that examines sibling and neighbor correlations in the developing world is Emran and Shilpi (2015), which measures correlations in educational attainment in India. The current study contributes to this nascent literature by measuring sibling and neighbor correlations in educational attainment in Indonesia.

2.3 Data

The replication I conduct of Solon et al. (2000) uses educational data from the Panel Study of Income Dynamics (PSID). The PSID is a longitudinal survey in the US covering a broad spectrum of social and economic indicators for individuals and households. It has been conducted by the Survey Research Center of the University of Michigan. The survey's sample follows a cluster design, began in 1968, has continued with annual resurveys through

[2] Jolliffe (2002) considers each household member that is at least 15 years old to be an adult for the purpose of estimating household education.

1997 and biennial surveys from 1997 to the present. Solon et al. chose the 1985 survey to measure years of schooling because of its detailed educational questionnaire.

Survey clusters are used as a proxy for neighborhood identity because of the close proximity of survey households in the smallest unit in the sampling design, the Survey Research Center (SRC) segments. While Solon et al. (2000) use the SRC segments, this study is restricted to the public use data. The measure of neighborhood in Solon et al. consists of groups of individuals that live in close proximity (e.g., in urban areas, Solon et al.'s segments may comprise a single block). The measure available to me is closer to the level of a county. My neighborhood units contain six to twenty (averaging eight) of the neighborhood units in Solon et al. (2000). If the ratio of the within-city-block variance to the between-city-block variance is higher than the ratio of the within-county variance to the between-county variance, then I expect my measure of neighborhood covariance to be biased down relative to the measure in Solon et al. (2000). Due to assortative matching of households to neighborhoods, I expect my neighborhood covariance, and correlation, to be lower than Solon et al (see appendix I for further discussion).

To extend the analysis to Indonesia, I make use of the Indonesian Family Life Survey (IFLS). The IFLS is a longitudinal survey that is representative of around 83% of the Indonesian population. The IFLS thoroughly tracks initial sample members and has a low attrition rate. The IFLS also follows a cluster design with the enumeration areas as small as 60 to 70 households. For Emran and Shilpi (2015), Indian neighborhoods contain a population of approximately 750 to 1000 individuals in urban areas and approximately 1500 in rural areas (Rao and Murthy, 2007). Their sample, from the National Family Health Survey in 2005/6, provides 24 households per neighborhood. I consider the IFLS cluster a reasonable proxy for neighborhood membership due to their relatively small size.

The first round of the IFLS, conducted in 1993, provides the family background data used in the estimation of the sibling and neighbor correlations. The most recently available round of the IFLS, conducted from late 2007 through early 2008, provides detailed educational attainment data that is comparable with the detail of schooling data in the 1985 PSID.

The sample restrictions in the current work follow the restrictions described in Solon et al. (2000). The pool of siblings is restricted to household members in 1968 who were between the ages of 25 and 33 in 1985. Solon et al. choose age 25 as a lower bound in their sam-

ple since they assume most schooling is completed by that age. I choose age 23 as a lower bound for the Indonesian sample following an analysis of school attendance probabilities in the US and Indonesian population surveys.[3]

For the Indonesian data, I restrict my sample to individuals between the ages of 23 and 31 at the time of the 2007/2008 survey who were members of 1993 survey households. To determine the bounds on the ages appropriate for this study, I follow the motivations of Solon et al. (2000). As in Solon et al., I also restrict my sample to an eight-year span of ages.[4] Appendix G motivates the selection of age 23 as the point at which most individuals would have completed their education. I exclude those with incomplete education responses in 2007/2008. See table F.1 for a summary of the PSID and IFLS samples.

PSID households in the Survey of Economic Opportunity (SEO), also known as the poverty sample, are excluded from this paper and from Solon et al. (2000) since they were only included if they were sufficiently poor. This restriction on sampled neighbors would bias the estimates of typical neighbor correlations.

Solon et al. (2000) pooled male and female observations due to their similarity of educational attainment. The differences between male and female mean years of schooling for both the PSID and IFLS samples are not significant (see table F.2).

In this paper I estimate years of schooling following the approach in Solon et al. (2000).[5] I use a table constructed by the Indonesian Ministry of Education and Culture (MOEC, 2013) to provide the mapping from years completed in various education levels to total years of schooling.[6]

Table F.3 compares the urban/rural divide in US and Indonesia. Page and Solon (2003) use urban areas to explain the neighbor correlations in income in the PSID. Emran and Shilpi (2015) split their sample by urban and rural subsamples and show differences in sibling correlations. Motivated by these results, I include estimates with and without controls for urban status.

[3]The analysis of schooling attendance by age in the US and Indonesia is presented in appendix G.

[4]For Björklund et al. (2010), the maximum possible birth year difference between siblings is nine years.

[5]Educational attainment could also be summarized by categories. For example, Barro and Lee (2013) rely on seven broad categories for cross-country comparison of educational attainment: no formal education, incomplete primary, primary, incomplete secondary, secondary, incomplete tertiary, and tertiary.

[6]I present the data in MOEC (2013) compared with the years-of-schooling estimation rules applied by Solon et al. (2000) in appendix H.

2.4 Econometric Model

To measure the importance of family and neighborhood background on subsequent educational attainment, I assume the following hierarchical model of educational attainment.

$$y_{cfs} = a_c + b_{cf} + e_{cfs}$$

The years of schooling,[7] y_{cfs}, for sibling in family in neighborhood depends on a linear combination of neighborhood-, a_c, family-, b_{cf}, and individual-specific effects, e_{cfs}. The effects in this model are composed of both observed and unobserved characteristics. The individual effect is assumed to be uncorrelated with the neighborhood effect and the family effect.

In the model presented in Solon et al. (2000), the neighborhood and family effects are assumed to be correlated. This assumption implies the estimates of the neighborhood correlations are upper bounds on the importance of neighborhood background. The sibling correlation in years of schooling is given by

$$\rho_f = \frac{Cov(y_{cfs}, y_{cfs'})}{Var(y_{cfs})}.$$

Where the sibling covariance of years of schooling is

$$\begin{aligned} Cov(y_{cfs}, y_{cfs'}) &= Cov(a_c + b_{cf} + e_{cfs}, a_c + b_{cf} + e_{cfs'}) \\ &= Cov(a_c, a_c) + 2Cov(a_c, b_{cf}) + Cov(a_c, e_{cfs'}) + Cov(b_{cf}, b_{cf}) \\ &+ Cov(b_c f, e_{cfs'}) + Cov(e_{cfs}, a_c) + Cov(e_{cfs}, b_{cf}) + Cov(e_{cfs}, e_{cfs'}) \end{aligned}$$

The individual effect is assumed to be random and uncorrelated with the neighborhood and family effects. Since each covariance involving the individual effect is then zero, the sibling covariance simplifies to the following equation.

[7] Years of schooling is assumed to be the deviation from mean schooling (i.e., $y_{cfs} \sim N(0, \sigma_y)$).

$$Cov(y_{cfs}, y_{cfs'}) = Var(a_c) + Var(b_{cf}) + 2Cov(a_c, b_{cf})$$

and the total variance of years of schooling is

$$Var(y_{cfs}) = Var(a_c) + Var(b_{cf}) + 2Cov(a_c, b_{cf}) + Var(e_{cfs}).$$

Similarly, the neighborhood correlation is given by

$$\rho_c = \frac{Cov(y_{cfs}, y_{cf's})}{(Var(y_{cfs}))}.$$

Where the covariance of years of schooling for two neighbors from different families is

$$Cov(y_{cfs}, y_{cf's}) = Var(a_c) + Cov(b_{cf}, b_{cf'}) + 2Cov(a_c, b_{cf}). \tag{2.1}$$

The covariance of neighbors' years of schooling, $Cov(y_{cfs}, y_{cf's})$, is an upper bound on the overall neighborhood effect on educational attainment. The covariance of family effects within the same neighborhood, $Cov(b_{cf}, b_{cf'})$, should not be entirely attributed to the neighborhood. Also, the covariance between family effects and their neighborhood effects, $Cov(a_c, b_{cf})$, is assumed to be positive due to assortative matching of advantaged families to advantaged neighborhoods.

Solon et al. (2000) tighten the upper bound on the neighborhood influence on years of schooling by subtracting off an estimate of the observable component of $Cov(b_{cf}, b_{cf'})$. To calculate the observable component, I regress years of schooling on family characteristics like race, parental income, and parental education and a neighborhood fixed effect. The addition of the neighborhood fixed effect controls for observed and unobserved neighborhood effects. This allows the estimation of the observed family effect to be free from all neighborhood effects.

The predicted observed family component from the above regression is my estimate of the observed component of years of schooling for each individual. The neighbor covariance of this observed family component is an element of $Cov(b_{cf}, b_{cf'})$. I subtract this component from the overall neighbor covariance of years of schooling to produce a tighter upper

bound on the influence of the neighborhood on educational attainment. The results of performing this procedure appear in tables 4 and 5 below.

2.5 Observable Family Characteristics

Observable family characteristics help tighten the upper bound on the neighborhood correlation. In the case of the PSID, I follow Solon et al. (2000) and control for whether the household head in 1968 was black, the log of family income in 1967 (as reported in the 1968 survey), whether a male head was present in 1968, his years of schooling, whether a female head or wife was present in 1968, and her years of schooling. For the IFLS, I control for the maximum level of adult education and the log of per capita expenditure,[8] in the first survey year (1993). As in Björklund et al. (2010) my interest is in those family characteristics which increase sibling similarity. Households missing family background were excluded from both the PSID and IFLS samples.

Instead of relying on another level of hierarchy in their multi-level model, Emran and Shilpi (2015) use neighborhood-level fixed effects to measure the neighborhood effect. This follows the process in Mazumder (2008) and Björklund et al. (2010).

2.6 ANOVA Estimation

I estimate the neighbor and sibling covariance with the maximum likelihood estimator of the lower-level covariances (among unrelated neighbors and siblings, respectively) averaged to produce aggregate estimates.[9] I begin by regressing years of education on gender and age via OLS. The residuals from this regression (labelled y_{cfs}) permit a simple estimator of the neighbor covariance:

[8] Though per capita income includes information on family size and income, Maralani (2008) found family size and years of schooling were not consistently correlated for households in the IFLS. Black et al. (2005) find birth-order, but no family size, effects on years of schooling for the entire population of Norway. Angrist et al. (2010) find the same using two Israeli census samples.

[9] The procedure I describe here is identical to the one in Solon et al. (2000) using their weighting scheme 1.

$$\widehat{Cov}(y_{cfs}, y_{cf's}) = \hat{E}(y_{cfs}, y_{cf's})$$
$$= \sum_{c=1}^{C} \left(\sum_{f \neq f' \in F_c} \left(\frac{\sum_{s=1}^{S_{cf}} \sum_{s'=1}^{S_{cf'}} y_{cfs} y_{cf's'}}{S_{cf} S_{cf'}} \right) \Big/ F_c \left(F_c - 1\right) \right) \Big/ C$$

The sibling covariance is similar:

$$\widehat{Cov}(y_{cfs}, y_{cfs'}) = \hat{E}(y_{cfs}, y_{cfs'})$$
$$= \sum_{c=1}^{C} \left(\sum_{f=1}^{F_c} \left(\frac{\sum_{s \neq s' \in S_{cf}} y_{cfs} y_{cfs'}}{S_{cf} \left(S_{cf} - 1\right)} \right) \Big/ F_c \right) \Big/ C$$

The variance of years of schooling is estimated by the sample variance. When calculating the sibling correlation, I calculate the variance (i.e., the denominator) from the sample restricted to families with at least two siblings. When calculating the neighbor correlation, I follow Solon et al. (2000) and calculate the variance of the full sample.

To determine the amount to be subtracted from the neighbor covariance in the adjusted neighbor correlation calculation, I first regress on the observed family characteristics and neighborhood fixed effects. I use the resulting coefficients on the family characteristics to predict the observed component of the family effect for each sibling. The neighbor covariance estimator above is then applied to these observed family effect estimates as an estimator of the observable component of $Cov(b_{cf}, b_{cf'})$. This portion of the within-neighborhood family covariance is subtracted from the previously estimated neighbor covariance to provide a tighter upper bound on the neighbor correlations.

Besides Solon et al. (2000), the above procedure is applied to correlations in educational attainment by Raaum et al. (2006).

2.7 REML Estimation

Unfortunately, the above covariance estimates weight each sibling, family, and neighborhood equally. Solon et al. (2000) propose several alternative weights, but these weighting schemes are admittedly ad hoc. Following the sibling correlation literature (e.g., Björklund and Jäntti, 2012; Mazumder, 2008 etc.), I assume the neighborhood and family effects are uncorrelated and random:

$$Cov(b_{cf}, b_{cf'}) = Cov(a_c, b_{cf}) = 0$$

This dramatically simplifies the calculation of the neighbor and sibling correlations, but at the cost of assuming no assortative matching. The simplified neighbor correlation is

$$\rho_c^{REML} = \frac{Var(a_c)}{Var(y_{cfs})}$$

and the simplified sibling correlation is

$$\rho_f^{REML} = \frac{Var(a_c) + Var(b_{cf})}{Var(y_{cfs})}.$$

I use restricted maximum likelihood (REML) to estimate this simplified mixed effects models.[10] REML estimates the variances of the random effects directly as opposed to estimating the effects themselves as would be the case for a fixed effects model. REML has the advantage of being unbiased for unbalanced group sizes and is preferred over ANOVA and ML for the analysis of variance components (Corbeil and Searle, 1976).

This solves one problem by replacing it with another. In the ANOVA procedure, the covariances are biased due to the choice of weight for each sibling, family, and neighborhood. In the REML procedure, the covariances are biased due to the assumption that the hierarchical level effects are uncorrelated. The neighbor effect can best be represented by $Var(a_c) + 2Cov(a_c, b_{cf})$.

[10] REML is used to calculate sibling and neighbor correlations in Mazumder (2008) and Lindahl (2011).

The neighbor covariance calculated in the ANOVA procedure includes the correlation of household effects within each neighborhood ($Cov(b_{cf}, b_{cf'})$), while the REML modelling assumptions omit the sorting of advantaged households into advantaged neighborhoods. The bias due to the REML procedure is undefined due to the intractability of the corresponding optimization. Simulations can be shed some light on the expected bias.

To remove some of the within-neighborhood family covariance, family characteristics are added to the fixed effect component of the mixed effects model. The results of the adjusted model are listed in the tables along the unadjusted results. The adjustment for family characteristics here corresponds to the adjustment made in the ANOVA procedure above.

I calculate 95% confidence intervals about my estimates using the percentiles of 200 bootstrap replications.[11] Each replication draws with replacement from the pool of primary sampling units and repeats each step in the above calculations.

2.8 Results

My replication of the procedure in Solon et al. (2000) produces similar estimates of each correlation (see tables 4 and 5). The 95% bootstrapped confidence intervals around my ANOVA estimates include the estimates in Solon et al. (table F.4). The lower estimate of the neighbor correlation is expected due to the size of my neighborhoods compared to the segments in Solon et al. (see appendix I). The consistency across my ANOVA estimates and those in Solon et al. supports the following comparisons with the results of the REML procedure and those using the IFLS.

The results of the REML procedure applied to the PSID sample are within the expected ranges (see table F.4). The sibling correlation confidence interval contains the estimate provided in Solon et al. (2000). The neighbor correlation is lower, as expected (see appendix I), and is not statistically different from zero. For both Solon et al. and my ANOVA estimates, the neighbor correlation adjusted for observable family characteristics is not statistically different from zero. For Solon et al., the estimate of the unadjusted neighbor cor-

[11] Raaum et al. (2006) use 250 replications to estimate standard errors for their estimates of sibling and neighbor correlations in educational outcomes.

relation is statistically different from zero, while for my ANOVA and REML estimates, the unadjusted neighbor correlation is not.

The variances for Indonesia are much larger than those in the US[12][13] (see table F.5). This is driven in part by the difference in compulsory schooling. Indonesia instituted six years of compulsory education beginning in 1984 (including two years of kindergarten) and raised this minimum to nine years in 1994 (MOEC, 2013). In 1993, the first round of the IFLS, approximately half of the siblings may have completed the old compulsory minimum.[14] The measure of years of schooling used in the models here is the residual of a regression on age and sex. This regression assumes a linear relationship between age and years of schooling. Though not explored in this paper, one could model the relationship between years of schooling and age recognizing the potential discontinuity provided by the 1994 policy change. The level of compulsory education in the US varies by state, but is typically nine years excluding kindergarten.

While the variances are much larger in Indonesia than in the US, the sibling correlation is remarkably similar. The larger variances indicate that the similarity across siblings has an even more dramatic impact in Indonesia. The similarity of the sibling correlation indicates that family has a similar relative impact on schooling in Indonesia and in the US. This similarity does not indicate a closeness in terms of inequality between the developed and the developing world. To the contrary. The scale of the variances indicates that the transmission of schooling advantage has a larger real impact in perpetuating educational inequality in Indonesia than in the US.

The neighbor correlation in Indonesia is larger than the neighbor correlation in the US (see table F.5). This difference is largely driven by the importance of the neighborhood effect in rural areas. Rural areas have lower expected educational attainment (see table F.3) and higher neighbor and sibling correlations of educational attainment than urban areas in Indonesia.

Due to the structure of the schooling system in Indonesia, this difference in the neighborhood effect may be driven by the relatively high heterogeneity of educational options

[12]Direct comparisons between the US and Indonesia throughout this section are intended to provide context to the scale of the variance and correlations in educational attainment in Indonesia.

[13]A comparison of the ANOVA procedure applied to the PSID and the IFLS is presented in table F.5.

[14]This was calculated by seeing how many siblings were older than 11, the age for the seventh year of schooling (MOEC, 2013). See the table in appendix H for a full listing of ages and years of schooling.

in rural areas. The schooling system in Indonesia includes public, private, and Islamic schools. The non-formal Islamic boarding schools, called *pesantren*, exist primarily in rural areas. The *madrasah*, Islamic educational institutions, tend to serve poor rural areas and are largely unsupported by public funds (MOEC, 2013).

The REML results (see table F.6), suggest the sibling correlation is lower in Indonesia than in the US. The scale of the estimated variance components is such that the real impact of the educational inequalities is higher in Indonesia than in the US. The inclusion of family observables has a larger impact on the correlation estimates in the US than in Indonesia (see table F.7). Cross-country comparison is made more difficult due to the difference in the relevant family observables. While whether or not the head of household is black is important in the US. The relevant class distinctions in the IFLS sample are more difficult to identify.

2.9 Comparison to Emran and Shilpi

Emran and Shilpi (2015) report estimates of the sibling and neighbor correlations in years of schooling larger than those reported here. Bias in the sibling and correlation estimates from Emran and Shilpi (2015) has a variety of potential sources. The age range of siblings in Emran and Shilpi (2015) is between 16 and 27, where individuals are included in my IFLS sample if they were household members between the ages of 8 and 16. My sample was much more likely to be living with the household in which they were raised than the sample in Emran and Shilpi. Nearly half of my sample (44%) in 2007 between the ages of 23 and 27 were no longer in their childhood home, as measured by their 1993 household.

Their estimates are positively biased due to omitting the children who no longer live with the surveyed household. Emran and Shilpi (2015) use a rural survey in India to estimate the bias due to the omission of individuals no longer living with the household.[15] They report there is no significant bias, but this bias is not precisely indicated.

The measurement of years of schooling was taken between the ages of 16 and 27 for Emran and Shilpi (2015), but between the ages of 23 and 31 for my sample. As a result, my mea-

[15] Emran et al. (2015) use the same data, the Rural Economic and Demographic Survey conducted by the National Council for Applied Economic Research in India, to measure co-resident bias of the intergenerational correlation of years of schooling. They estimate the co-resident sample selection bias is 10.4%.

sure of years of schooling is more representative of eventual educational attainment than the measure in Emran and Shilpi.

2.10 Conclusion

A previous study indicates neighbor correlations are larger in the developing world. Unfortunately, this study establishes only weak links to studies in the developed world. If the family/neighborhood background impact on socioeconomic mobility for the developed is different than in the developing world, policy makers need to consider the neighborhood background when generalizing policy recommendations across contexts. My paper is the first to establish a more concrete comparison to the developing world.

I find that in Indonesia the sibling correlation is smaller and the neighbor correlation is larger than in the US. This difference in not always statistically significant. This apparent similarity in correlations suggests the proportion of educational inequality affected by the family is relatively constant regardless of the level of development.

However, the similarity in correlations is not the end of the story. The variance of educational attainment is much larger in Indonesia than in the US. Simply comparing correlation measures across countries understates economically significant differences in mobility. In Indonesia, living in a better neighborhood (or with a better family) has a similar proportional impact on educational attainment as in the US, while the schooling mobility (or risk) in terms of years of schooling is much higher in Indonesia.

CHAPTER 3
WHOSE INCOME IS HUMP SHAPED?

3.1 Introduction

The average age profile of income is "hump" shaped,[1] but not every individual is average. The average age profile of income suggests that borrowing early in life is consumption smoothing, but, without higher income later in life, early borrowing can be disastrous. Previous research has decomposed the aggregate age profile of income by defining subpopulations on the basis of level of education and occupation.[2] In contrast, this study defines subpopulations on the basis of the similarity of individual age profiles of income.

This distinction can be compared to asking whether the less educated are relatively poor (previous studies) vs. asking whether the relatively poor are less educated (this study). Since lack of education is both a cause and an effect of poverty, an understanding of both analytical approaches provides a more nuanced perspective on the nature of poverty. In the same way, I describe the demographic characteristics of income-profile groups, where previous studies have described the income-profile characteristics of demographic groups.

I create income-profile groups using standard time-series clustering techniques applied to the PSID and find that the average *hump-shaped* pattern can be decomposed into a set of distinct income-profile patterns. I label these patterns *high-growth*, *hump-shaped*, *declining*, and *flat*. The largest portion of the sample is assigned to a hump-shaped group (40%), followed by the flat group (34%), then declining (21%), and finally high-growth (5%).

The demographic features of these income-profile groups shows that the hump-shaped and high-growth groups are more highly educated and have higher lifetime incomes relative to the other groups. The high-growth group has the lowest starting incomes, suggesting

[1] The hump-shaped feature of average income has inspired many studies from Thurow (1969) through to Aguiar and Hurst (2013). Lee et al. (2008) suggest the hump-shaped aggregate age-income profile is a robust observation across countries, independent of the level of development.

[2] For examples of age profiles of income and consumption disaggregated by level of education, see Attanasio et al. (1999); Gourinchas and Parker (2002); Cagetti (2003); Fernández-Villaverde and Krueger (2007); Guvenen (2007). See Gourinchas and Parker (2002) for an example of disaggregation by occupation.

that much of the hump shape observed in the aggregate is driven by staggered engagement into the workforce across individuals rather than a gradual increase in individual income. Wealthy parents may enable their children to postpone work through their 20s, while, for others, market conditions may drive the timing of their exit from higher education.

The shape of average income and consumption profiles drives understanding of lifecycle savings decisions. In general, the perceived hump shape is fundamental to the study of the lifecycle income. This study is step towards a more nuanced understanding of typical lifecycle patterns of income.

3.2 Review of Related Literature

Lifecycle profiles of wealth, consumption, and income are the focus of a large literature. Lee and Mason (2011b) present several relevant chapters focused on incomes over the lifecycle. Lee and Mason (2011a) show the average age-income profile is hump-shaped for rich, poor, and hunter-gatherer economies. Lee and Ogawa (2011) present age-income profiles for 23 countries, providing further evidence that the aggregate age-income profile is hump-shaped regardless of the level of development. In contrast, private consumption profiles exhibit distinctive heterogeneity across economies (An-Chi Tung, 2011). Lee et al. (2011) show that, in the US, the aggregate age-income profile has grown steeper from 1960 to 2003. These papers demonstrate the hump-shaped aggregate age-income profile is a consistent feature across economies and birth-year cohorts. I extend this literature by identifying groups for which this aggregate pattern does not hold.

Aguiar and Hurst (2013) show that there is "substantial heterogeneity" in lifecycle consumption patterns when looking at consumption subcategories. In particular, they show that the apparent decline in non-durable consumption later in life is driven by decreases among a select group of subcategories (food, nondurable transportation, and clothing/personal care).

Attanasio et al. (1999) decompose age profiles of income according to high or low education subsamples. They show incomes are both higher and steeper for the more highly educated subsample. Previous studies have also classified earners by a variety of occupation categories (Masson, 1986), blue-collar vs. white-collar workers (Burbidge and Robb, 1985), immigration status (Shamsuddin and DeVoretz, 1998; Hao, 2004), and race-ethnicity (Hao,

2004). In contrast, this paper seeks to explain the "hump" in lifecycle income through decomposing income by subpopulations defined by emergent lifecycle patterns rather than predefined subcategories.

This paper also extends work by Flachaire and Nuñez (2007) to disaggregate the cross-sectional distribution of incomes into homogenous subpopulations. They model the heterogeneous population distribution as a mixture of homogeneous subpopulation income distributions. This paper explores the possibility that the heterogeneity of lifecycle income patterns may be concealing homogeneity of subpopulation income patterns.

Deaton (1992) characterizes the two dominant theories of income over the past half century. In the Modigliani and Brumberg (1954) lifecycle hypothesis of savings, savings decisions are determined by the value of lifetime resources (Modigliani, 1986). In Friedman's permanent income theory of consumption, consumption is determined by permanent income which could be interpreted as the average or expected income (Friedman, 1957). The focus of this paper on the relationship between income and age is closely aligned with the body of research on the lifecycle hypothesis. The permanent income theory of consumption focuses on the consumption response to changes in income. An exploration of this relationship, while of great interest, is beyond the scope of this present work.

There is a large literature on lifecycle profiles of wealth, consumption, and income. Lee and Mason (2011b) present several relevant chapters on incomes over the lifecycle. Lee and Mason (2011a) show the average age-income profile is hump-shaped for rich, poor, and hunter-gatherer economies. Lee and Ogawa (2011) present age-income profiles for 23 countries, providing further evidence that the aggregate age-income profile is hump-shaped regardless of the level of development. What does vary is the age at which the peak of the hump occurs. Lee et al. (2011) show that, in the US, the aggregate age-income profile has grown steeper from 1960 to 2003. These papers demonstrate the hump-shaped aggregate age-income profile is a consistent feature across economies and birth-year cohorts. I extend this literature by identifying groups for which this aggregate pattern does not hold.

3.3 Data

3.3.1 PSID

To analyze incomes over the life course, I use the Panel Survey of Income Dynamics (PSID), the longest running longitudinal survey of income. As in Aguiar and Hurst (2013), I control for cohort and time effects,[3]. I restrict my sample to males who have ever been household heads, and who were 25 during a year on which they reported income (or at least family income). This study includes income for years 1967 through 2010 adjusted for inflation to 2014 USD. Before analyzing the age profiles, I regress income on a set of cohort and year dummy variables to control for effects that may be independent of age. The following model clarifies the starting assumptions of this process.

$$ln\ y_{iact} = \beta + \gamma_c + \psi_t + \epsilon_{iact} \tag{3.1}$$

I assume a time-invariant cohort effect γ_c and a year effect ψ_t on income for individual i at age a in time period t. I assume that the true data generating process is given by the following:

$$ln\ y_{iakct} = \beta + \gamma_c + \psi_t + \alpha_{ak} + u_{iakct} \tag{3.2}$$

where k indicates the group to which individual i belongs. I also assume no *a priori* knowledge of the group-level age effects α_{ak}. The group-level age effects cannot be separately estimated from the residual without first identifying the group assignments. I begin with estimating the residuals from equation 3.1 and use time-series clustering to group the individuals based on their age profiles.

I use the needs estimates described in Grieger et al. (2009) to construct family-income-to-needs ratios, which are used to describe the resulting groups.[4] The US Census Bureau reports poverty thresholds estimated from the March Current Population Survey. Grieger

[3]Following Deaton (1997, pp. 123-127) I attribute growth to age and cohort effects, and assume the year effects measure the business cycle, are orthogonal to a time trend, and sum to 0 over the long run.

[4]The needs estimate and family income are both determined in part by family size. As a result, differences in the income-to-needs ratio cannot be separately assigned to a family size or income effect. Despite this drawback, the family-to-income needs ratio provides a simple measure of equivalent income to facilitate welfare comparisons across individuals.

et al. use the detailed poverty-threshold matrices from 1972 through the present. They use weighted averages based on non-farm family sizes in the pre-1972 period. They demonstrate their needs measure produces poverty rates that are more correlated with the official US Census Bureau poverty rates than the alternatives typically used in the literature. This measure is not without its complications (e.g., the codetermination of family income, family needs, and family size), but it does indicate a family's eligibility for welfare and thus captures one notion of poverty.

3.3.2 Summary of data

Table J.1 summarizes the number of observations available at different ages. While attrition does occur, the dramatic decline of observations based on age is driven by the continued addition of individuals throughout the PSID. That is, while the survey began in 1968, individuals have been continuously added to the sample. For example, sample individuals born in 1987 could report incomes for 2012 but would not be represented in any other age group.[5] I calculate the clustering assignment using ages 25-45 since this range maximizes my input data under the algorithm's constraint of a balanced panel. Results are presented through age 60 to show more of the life cycle. I do not present data for ages greater than 60 due to the greatly reduced sample at these ages.

Plotting the real income of household heads (relative to the same at age 25) presents the standard "hump" in income over the life course (figure K.1). The observed paths deviate sufficiently from the mean to inspire the clustering exercises presented in this paper. While there may be individuals who have the mean experience, they appear to be a slim minority.

Not all individuals have the same income experience over the life course. However, this heterogeneity can still be summarized by representative profiles. The next section outlines methods of identifying profiles that are more nuanced than the mean experience, helping explain some of the variation around it.

[5] 2012 is the last year for which incomes are recorded at the time of this writing. The latest available survey year, 2013, allows individuals to report last year's labor income.

3.4 Methodology

The techniques for clustering a time-series can be divided into raw-data-based, feature-based, and model-based algorithms.[6] Raw-data-based algorithms group like series without first summarizing them. Feature-based algorithms reduce each series to a set of characteristic features (e.g., minimum, maximum, etc.) and group like features. The prime advantage of the feature-based approach is a reduction in computation and complexity by reducing dimensionality. Model-based algorithms estimate a model (e.g., ARIMA) and group like model parameters.[7]

This paper uses a raw-data-based algorithm. I take the income series from age 25 through 45, classify each year as a dimension, and cluster individuals based on proximity in the resulting 20 dimensional space. In the next section I define the specific clustering algorithm I use (i.e., k-means clustering).

3.4.1 k-Means Algorithm

I begin with some basic definitions to setup the definition of the k-means algorithm. Assume $z^n = \{x_1, \ldots, x_n\}$, where $x \in \mathbb{R}^p \ \forall i \in \{1, \ldots, n\}$, are drawn from an unknown distribution $p(x)$.[8] In general, a clustering algorithm, $\Psi(z^n, k)$, generates a mapping, $\psi(x) : \mathbb{R}^p \to \{1, \ldots, k\}$, from each sample observation, $x \in \mathbb{R}^p$, to its cluster, k.

The k-means algorithm randomly selects k observations $r_c^0 \in \mathbb{R}^p$, $c \in \{1, \ldots, k\}$ to stand as representatives of each cluster. Each observation, x_i, is assigned to share the cluster of its closest representative r_c^0.[9] The centroid of each cluster is calculated and assigned to the representatives, r_c^j,of the current iteration, j, of the algorithm. Given the new representatives, observations are assigned to their closest iteration-specific representative. The algorithm continues to recalculate centroids and reassign observations to clusters until cluster membership does not change.

[6]This characterization comes from Liao (2005). For a more recent survey of the time-series clustering literature, see Fu (2011).

[7]For example, D'Urso and Maharaj (2009) use autocorrelation coefficients to perform fuzzy clustering of CO emissions time series. De Luca and Zuccolotto (2011) use tail dependence coefficients to cluster time series of market indices. This approach tends to cluster series facing the same set of shocks (e.g., market indices within the same geographic region).

[8]In this section I follow the notation in Wang (2010).

[9]Closeness is determined using Euclidean distance.

The convergence of this algorithm depends on the initial positions of the cluster representatives, and the number of iterations. I use 25 restarts of the algorithm and allow for a maximum of 100 iterations.[10] The ending assignment of $x_i \rightarrow \{1, \ldots, k\}$ is the cluster mapping $\psi(x)$.

3.4.2 Choosing the number of clusters

There exists no universal method of identifying the appropriate number of clusters in the absence of meaningful age profile of income categories. That is, it is not known from how many distributions the age profiles of income of individuals are drawn. To select an optimal number of clusters, the first step is to choose a criterion over which to optimize. This optimal number is then only optimal insofar as the criterion is of interest.

One promising criterion is the stability of cluster assignment. That is, given the correct number of clusters, it is assumed that members of the same subpopulation will always be grouped together, and vice versa.

In this way, I identify the appropriate number of clusters k by minimizing the cluster-assignment instability across choices of k as in Wang (2010).[11] The first step is estimating the cluster-assignment instability $s(\Psi, k, n)$ for a given choice of k. A clustering algorithm $\Psi(\cdot; k)$ is stable if, for every two samples of size n, the corresponding cluster mappings $\psi_1(x)$ and $\psi_2(x)$ are identical. We can measure the distance between two cluster mappings as the probability that two independent samples X and Y drawn from $p(x)$ are in the same cluster under one mapping but not the other as

$$d(\psi_1, \psi_2) = Pr[I\{\psi_1(X) = \psi_1(Y)\} + I\{\psi_2(X) = \psi_2(Y)\} = 1]$$

Wang (2010) defines cluster-assignment instability as

[10] The k-means algorithm I use is detailed in Hartigan and Wong (1979). The algorithm is restarted if a cluster is empty after the initial assignment. Multiple restarts increase the probability that all clusters are nonempty.

[11] Fang and Wang (2012) propose a similar method based on the bootstrap. Unfortunately, as noted by Lange et al. (2004), the bootstrap samples are non-disjoint and may artificially lower the dissimilarity measure. In response to this concern, I use the method proposed in Wang (2010).

$$s(\psi, k, n) = E[d\{\psi(Z^n; k), \psi(Z^{n*}; k)\},]$$

where Z^n and Z^{n*} are two independent samples of size n drawn from $p(x)$.

Lower values of $s(\psi, k, n)$ indicate a more stable cluster assignment. The degenerate case, $k = 1$, is excluded from the analysis.

3.4.3 Leave-many-out Cross-validation with Averaging

In order to estimate the instability of clustering assignment, Wang (2010) proposes the following leave-many-out cross-validation with averaging algorithm:

Step 1 Permute the empirical distribution $(x_1, \ldots, x_n)$ and label the new distribution $(x_1^c, \ldots, x_n^c)$.

Step 2 Create three disjoint subsets of the data such that the first two have the same number of observations m with $2m < n$: $z_1^c = (x_1, \ldots, x_m)$, $z_2^c = (x_{m+1}, \ldots, x_{2m})$, and $z_3^c = (x_{2m+1}, \ldots, x_n)$. This provides two training sets and one validation set respectively.

Step 3 Generate $\psi_1^c = \Psi(z_1^c, k)$ and $\psi_2^c = \Psi(z_2^c, k)$.

Step 4 Calculate the distance between these cluster mappings as follows:

$$d^c = \frac{2}{(n-2m)(n-2m-1)} \sum_{2m+1 \leq i < j \leq n} I[I\{\psi_1^c(x_i) = \psi_1^c(x_j)\} + I\{\psi_2^c(x_i) = \psi_2^c(x_j)\} = 1]$$

That is, estimate the expected probability that randomly selected pairs of observations (e.g., x_i and x_j) are assigned to the same cluster in one cluster mapping (e.g., ψ_1), but not the other (e.g., ψ_2).

Repeat Steps 1-4 for each permutation $c \in (1, \ldots, C)$. Then the estimated $s(\Psi, k, n)$ is given by

$$\hat{s}(\Psi, k, n) = C^{-1} \sum_{(} c = 1)^{C} d^{c}.$$

So, instability is an estimate of the expected distance between cluster mappings constructed using disjoint subsamples. This could also be viewed as the consistency of the clustering algorithm $\Psi(\cdot, k)$ for a given choice of k. That is,

$$s(\Psi, k, n) - 2\hat{\sigma}_k \leq s(\Psi, k', n) \text{ for any } k' < k.$$

3.4.4 Cluster Instability Results

Following Wang (2010), I set the number of permutations, C, to 100 in order to estimate the clustering instability, $\hat{s}(\Psi, k, n)$. The clustering algorithm, $\Psi(\cdot, k)$, is the k-means clustering algorithm specified above. In the estimation of cluster assignment, each individual in the sample is represented by the five-year moving averages of income at ages 25 through 45, where income is the residual from equation 1. Requiring an observation at each year in this range restricts the sample to 2,236 individuals.[12]

Figure K.2 displays the instability estimates, $\hat{s}$, for the methods described in Wang (2010). There are local minima at five clusters and nine clusters.[13]

3.5 Results

I now present the results of the k-means algorithm for the two candidate values of k (i.e., five and nine). Figure K.3 displays the cluster assignments for the k-means algorithm when the number of clusters is five. The clusters are labeled in descending order of sub-

[12] I use five-year moving averages for age-income profiles following the approach in Deaton (1997) and Page and Solon (2003). This allows me to include individuals with up to 4 consecutive years of missing income data between ages 25 and 45.

[13] I have also calculated the cluster instability setting C to 1000. The resulting graph had local minima at five and nine clusters as well.

sample size. The dark line in each plot traces the mean values of income. The volatility of the mean series increases over time as individuals drop out of the survey.[14,15]

For figures K.3-K.6, I keep the y scale fixed within a row of clusters, but allow it to vary across rows. This allows comparison of patterns across rows where such a comparison is relatively natural. The alternative of keeping the scale fixed for all clusters obscured the patterns present in the larger clusters in order to show the patterns in the smaller, higher variance clusters.

When the number of clusters is 5, cluster 1, comprising 41% of the sample, exhibits a flat age profile of income with a decline from age 50. Cluster 2, comprising 35% of the sample, exhibits the typical "hump" shape, maintaining its peak from around 35 to 50, before returning to age-25 levels by age 60. Cluster 3, roughly 12% of the sample has a mean profile that suffers a massive decline from age 25 to 35. This cluster is characterized by a high level of volatility. Clusters 4 and 5, which combined comprise the remaining 12% of the sample, are characterized by high growth between 25 and 30 then plateau through age 60.

Figure K.4 indicates that the high-growth clusters (4 and 5) began with relatively low income. I label these groups *late-bloomers*. Due to their respective income patterns, I label cluster 1, *early-retirement*, cluster 2, *hump-shaped*, and cluster 3, *U-shaped*.

Figures K.5 and K.6 permit a similar characterization. In this case, clusters 4, 7, 8 match the *late-bloomers* pattern and cluster 1 matches the *early-retirement* pattern. Cluster 2 matches the *hump-shaped* pattern, clusters 3 and 6 I label *steady-decline.* Cluster 5 faces a "U"-shaped decline and recovery between ages 25 and 35, then roughly maintains its level through 60. I label this cluster *early-U*. Cluster 9 groups seven individuals that I label *outliers* since their distinctive feature is a massive negative spike in income at age 39.

3.6 Summary of cluster demographics

Figures K.7-K.10 display demographic distributions for the clusters when the number of clusters is nine. Relative to other clusters, *late-bloomers* on average are from older cohorts

[14]The most common way for individuals to drop out of the survey is by belonging to a cohort that does not reach 60 by the most recent survey year. For example, the 1982 birth-year cohort had only reached age 30 by 2012, the last recorded year of income.

[15]Appendix N presents the income variance by age in the sample broken down by cluster.

(figure K.7), have more years of education (figure K.8), have lower incomes at age 25 (figure K.9), and have relatively high lifetime family-income-to-needs ratios (figure K.10). The *early-retirement* group has relatively low family-income-to-needs ratios and relatively low education. This suggests the late drop in income may be related to involuntary reductions in wage income as opposed to voluntary early retirement.

The *steady-decline* clusters, on average, start with the highest levels at age 25 (figure K.9), but have relatively low lifetime family-income-to-needs ratios (figure K.10). The *hump-shaped* cluster is more educated than the other large clusters. Based on the selected demographics, there is no significant distinction between the *early-U* demographic averages and those for the steady-decline clusters.

Appendix L presents this demographic comparison when the number of clusters is five. Appendix M identifies the pairwise comparisons that are statistically significant using two-sided t-tests and the correction for type I errors in Holm (1979).

3.7 Comparison to Related Studies

This study confirms the findings of previous studies and expands upon them. Attanasio et al. (1999) and Cagetti (2003) find that individuals in the PSID with higher education levels have steeper age profiles of income. Gourinchas and Parker (2002) find the same when looking at the Consumer Expenditure Survey. This finding is echoed by the finding here that *late-bloomers* are more highly educated. That is, not only do those with higher education tend to have steeper age profiles of income, those with steeper age profiles of income also tend to be more highly educated.

Modigliani's lifecycle hypothesis predicts savings behavior based on assumptions of individuals' current income relative to their expected lifetime wealth Modigliani (1986). Assuming clusters identify different experiences (and different expectations) of income over the lifetime, the aggregate savings decision at a given age is composed of decisions based on these different types. The decrease in savings rate in the US since the 1970s may have been driven in part by a shift in types, since individuals with steadily declining incomes are much less likely to have hump-shaped savings than individuals with hump-shaped income. A thorough analysis of the lifecycle hypothesis under varying types is beyond the

scope of this paper. Future work, through incorporating lifecycle patterns in consumption and savings, can more directly address the implications of Modigliani's framework.

3.8 Conclusion

This paper is the first to group individual age-income profiles by the similarity of the profiles. I improved on previous attempts to decompose the age profile of income by starting free of assumptions on the causal mechanism affecting the shape of the age profile of income. Previous attempts first split their sample based on a demographic feature and attributed perceived differences in the profiles to the demographic feature (e.g., Attanasio et al. (1999) splitting based on level of education). Since many factors likely correlate with any chosen demographic feature, starting with emergent lifecycle patterns permits a more thorough analysis of likely causes. For example, previous papers compared income profiles across education groups, while this paper compares educational attainment across income profile groups.

The aggregate age-income profile is not a representative profile, grouping individuals by the similarity of their age-income profiles allowed the characteristic patterns of income to surface. Individuals with *high-growth* or *hump-shaped* age profiles of income have higher levels of education and lifetime income than those with *steady-declining* or *early-U* age profiles of income. Additionally, those assigned to the *high-growth* group had much lower starting incomes.

Everyone has a story about their past. These stories are filled with ups and downs. Among the highs we recount new jobs, promotions, and new family members. Job loss, missed opportunities, and the loss of loved ones lend our stories a sense of tragedy. These stories are how we create self-identity[16] and how we identify with others.[17] About stories, Kurt Vonnegut (2009) said, "stories have shapes which can be drawn on graph paper, and that the shape of a given society's stories is at least as interesting as the shape of its pots or spearheads." It is these shapes of the stories of our lives which I have attempted to identify in this paper.

[16] Haslam et al. (2011) show that even memories of personal facts (e.g., the name of your first job) as opposed to remembered experiences play an important role in self identity.

[17] Alea and Bluck (2007) provide experimental evidence supporting the intuition that sharing life stories builds intimacy.

Alternate clustering algorithms, such as model-based and feature-based approaches, can test the robustness of these results and extend the clustering process to address other questions. For example, one could ask what clusters would be generated if clustering by the autoregressive parameter. Clusters formed in this way would allow comparison of groups across which the persistence of income shocks varies. I intend to explore these and other open questions in future work.

APPENDIX A

TABLES FOR CHAPTER 1

Table A.1: Neighbor Correlation and Variance Components (KIDS)

	KwaZulu	Natal	Natal − KwaZulu
Neighbor Correlation	0.3835	0.8929	0.5094 ***
	[0.2403, 0.5169]	[0.7365, 0.9549]	[0.2773, 0.6785]
Household Component	0.0866	0.0968	0.0101
	[0.0639, 0.1113]	[0.043, 0.1571]	[-0.0584, 0.0816]
Neighbor Component	0.0539	0.8064	0.7525 ***
	[0.0311, 0.0783]	[0.3357, 1.2412]	[0.2616, 1.2131]
Transitory Component	0.2559	0.2177	-0.0382
	[0.2302, 0.2785]	[0.1606, 0.2579]	[-0.1035, 0.0149]
Households	578	175	
Neighborhoods	50	17	

Brackets define 95% confidence intervals calculated with 10,000 bootstrap samples.

Table A.2: Neighbor Correlation and Variance Components (NIDS-KZN)

	Tribal	Non-Tribal	Non-Tribal − Tribal
Neighbor Correlation	0.1414	0.5335	0.3921 *
	[0.0692, 0.2086]	[0.1448, 0.7408]	[-0.0128, 0.6275]
Household Component	0.1412	0.1828	0.0416
	[0.1095, 0.1766]	[0.1286, 0.2426]	[-0.0254, 0.107]
Neighborhood Component	0.0233	0.2091	0.1858 **
	[0.0107, 0.0359]	[0.0294, 0.4667]	[0.0039, 0.4473]
Transitory Component	0.3399	0.3172	-0.0227
	[0.2951, 0.3852]	[0.2708, 0.3593]	[-0.0947, 0.0456]
Households	950	373	
Neighborhoods	51	31	

Brackets define 95% confidence intervals calculated with 10,000 bootstrap samples. *, **, and *** indicate zero is not included in a 90%, 95%, and 99% confidence intervals respectively.

Table A.3: Neighbor Correlation and Variance Components (NIDS-RSA)

	Tribal	Non-Tribal	Non-Tribal − Tribal
Neighbor Correlation	0.1054	0.5503	0.4449 ***
	[0.0528, 0.1497]	[0.4419, 0.6209]	[0.3315, 0.5328]
Household Component	0.1798	0.2201	0.0403 **
	[0.1535, 0.207]	[0.1929, 0.2485]	[0.0015, 0.0796]
Neighborhood Component	0.0212	0.2694	0.2482 ***
	[0.0101, 0.031]	[0.1811, 0.3429]	[0.1602, 0.321]
Transitory Component	0.3701	0.3839	0.0138
	[0.3397, 0.3987]	[0.3561, 0.4112]	[-0.0279, 0.0538]
Households	2237	2122	
Neighborhoods	131	212	

Brackets define 95% confidence intervals calculated with 10,000 bootstrap samples. *, **, and *** indicate zero is not included in a 90%, 95%, and 99% confidence intervals respectively.

Table A.4: Percent Change in the Difference
Between KwaZulu and Natal (KIDS)

	Education	Agriculture	Urban	District
Neighbor Correlation	0.109	0.0054	-0.0379	0.0543
	[-0.1105, 0.2993]	[-0.0161, 0.0346]	[-0.6627, 0.2801]	[-1.0215, 0.5492]
Household Component	0.4643	0.0155	-0.1543	-0.0286
	[-4.8581, 4.2351]	[-0.6115, 0.583]	[-0.7019, 0.7813]	[-0.5072, 0.5201]
Neighborhood Component	-0.4412***	0.0221	-0.6419***	-0.6876***
	[-0.6285, -0.2826]	[-0.0043, 0.0846]	[-0.9171, -0.2351]	[-1.0004, -0.4446]
Transitory Component	0.1251	0.0802	-0.0058	0.0043
	[-1.0352, 1.5683]	[-0.3414, 0.726]	[-0.0681, 0.0472]	[-0.0247, 0.0398]

Brackets define 95% confidence intervals calculated with 10,000 bootstrap samples. *, **, and *** indicate zero is not included in a 90%, 95%, and 99% confidence intervals respectively.

Table A.5: Percent Change in the Difference
Between Tribal and Non-Tribal in KwaZulu-Natal (NIDS)

	Education	Agriculture	Urban	District
Neighbor Correlation	-0.1321	0.0022	-0.0655	-0.0299
	[-1.0322, 0.2826]	[-0.0296, 0.0307]	[-0.7757, 0.195]	[-1.3081, 0.709]
Household Component	-0.3592	-0.006	0	-0.0588
	[-2.1323, 1.2022]	[-0.1306, 0.1073]	[-0.0701, 0.0778]	[-0.6736, 0.4789]
Neighborhood Component	-0.4244**	-0.0026	-0.127	-0.1032
	[-0.8506, -0.1497]	[-0.0345, 0.0144]	[-0.6634, 0.1171]	[-1.0754, 0.4467]
Transitory Component	0.1562	-0.0147	0.001	0.0005
	[-1.7635, 1.7657]	[-0.1827, 0.1605]	[-0.0208, 0.0214]	[-0.0424, 0.0434]

Brackets define 95% confidence intervals calculated with 10,000 bootstrap samples. *, **, and *** indicate zero is not included in a 90%, 95%, and 99% confidence intervals respectively.

Table A.6: Percent Change in the Difference
Between Tribal and Non-Tribal in South Africa (NIDS)

	Education	Agriculture	Urban	District
Neighbor Correlation	-0.0871**	0.0030**	-0.0846***	0.0338
	[-0.1718, -0.0182]	[0.0000, 0.0078]	[-0.1602, -0.0294]	[-0.2387, 0.1186]
Household Component	-0.1705	-0.0055	0.0032	-0.0328
	[-0.7302, 0.5345]	[-0.055, 0.0199]	[-0.0267, 0.0359]	[-0.1675, 0.0474]
Neighborhood Component	-0.3562***	0.0006	-0.1494***	0.0100**
	[-0.4315, -0.2827]	[-0.0013, 0.0028]	[-0.2561, -0.0588]	[-0.4425, -0.0033]
Transitory Component	-0.2019	0.0308	-0.0001	-0.0008
	[-1.9395, 1.976]	[-0.346, 0.4047]	[-0.0109, 0.0113]	[-0.031, 0.0269]

Brackets define 95% confidence intervals calculated with 10,000 bootstrap samples. *, **, and *** indicate zero is not included in a 90%, 95%, and 99% confidence intervals respectively.

Table A.7: Remaining Difference
Between KwaZulu and Natal (KIDS)

	Education	Agriculture	Urban	District
Neighbor Correlation	0.5649***	0.5121***	0.4901**	0.5371*
	[0.3068, 0.7353]	[0.3096, 0.667]	[0.1296, 0.6836]	[-0.0102, 0.7374]
Household Component	0.0149	0.0103	0.0086	0.0099
	[-0.0366, 0.0739]	[-0.0478, 0.075]	[-0.0499, 0.0719]	[-0.0493, 0.0731]
Neighborhood Component	0.4205***	0.7691***	0.2695***	0.2351*
	[0.1467, 0.6961]	[0.3009, 1.2043]	[0.053, 0.4462]	[-0.0004, 0.3539]
Transitory Component	-0.0430	-0.0413	-0.0380	-0.0384
	[-0.1025, 0.0028]	[-0.1036, 0.0068]	[-0.0998, 0.0104]	[-0.1, 0.0098]

Brackets define 95% confidence intervals calculated with 10,000 bootstrap samples. *, **, and *** indicate zero is not included in a 90%, 95%, and 99% confidence intervals respectively.

Table A.8: Remaining Difference
Between Tribal and Non-Tribal in KwaZulu-Natal (NIDS)

	Education	Agriculture	Urban	District
Neighbor Correlation	0.3403*	0.3930**	0.3664*	0.3804
	[-0.0127, 0.5836]	[0.0057, 0.6128]	[-0.0196, 0.5835]	[-0.0556, 0.5423]
Household Component	0.0267	0.0414	0.0416	0.0392
	[-0.0206, 0.076]	[-0.0233, 0.1088]	[-0.0232, 0.1092]	[-0.0247, 0.1059]
Neighborhood Component	0.107**	0.1854**	0.1622**	0.1667*
	[0.0023, 0.2704]	[0.007, 0.4371]	[0.0017, 0.3662]	[-0.0076, 0.2505]
Transitory Component	-0.0263	-0.0224	-0.0227	-0.0227
	[-0.0925, 0.0364]	[-0.0884, 0.0414]	[-0.0887, 0.0412]	[-0.0892, 0.0382]

Brackets define 95% confidence intervals calculated with 10,000 bootstrap samples. *, **, and *** indicate zero is not included in a 90%, 95%, and 99% confidence intervals respectively.

Table A.9: Remaining Difference
Between Tribal and Non-Tribal in South Africa (NIDS)

	Education	Agriculture	Urban	District
Neighbor Correlation	0.4061***	0.4463***	0.4073***	0.4599***
	[0.2939, 0.4942]	[0.3331, 0.5344]	[0.2914, 0.4999]	[0.2948, 0.5202]
Household Component	0.0334**	0.0401**	0.0404**	0.0390**
	[0.001, 0.0661]	[0.0013, 0.0794]	[0.0025, 0.0791]	[0.0014, 0.0779]
Neighborhood Component	0.1598***	0.2483***	0.2111***	0.2506***
	[0.1027, 0.2082]	[0.1604, 0.3211]	[0.1328, 0.2782]	[0.1128, 0.2606]
Transitory Component	0.0110	0.0143	0.0138	0.0138
	[-0.0288, 0.051]	[-0.0274, 0.0544]	[-0.0261, 0.0537]	[-0.0261, 0.0537]

Brackets define 95% confidence intervals calculated with 10,000 bootstrap samples. *, **, and *** indicate zero is not included in a 90%, 95%, and 99% confidence intervals respectively.

APPENDIX B

FIGURES FOR CHAPTER 1

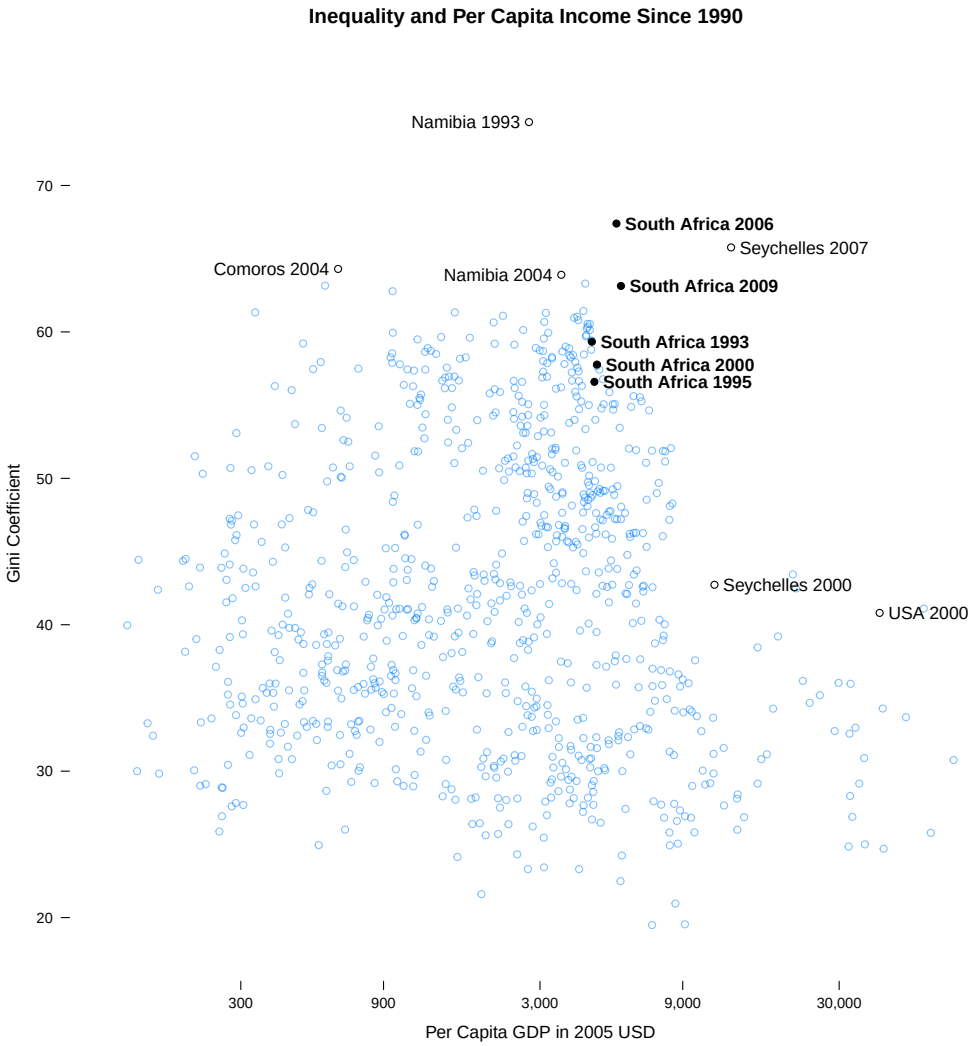


Figure B.1: Source: Author's calculations using World Bank data on Gini coefficients and per capita GDP.

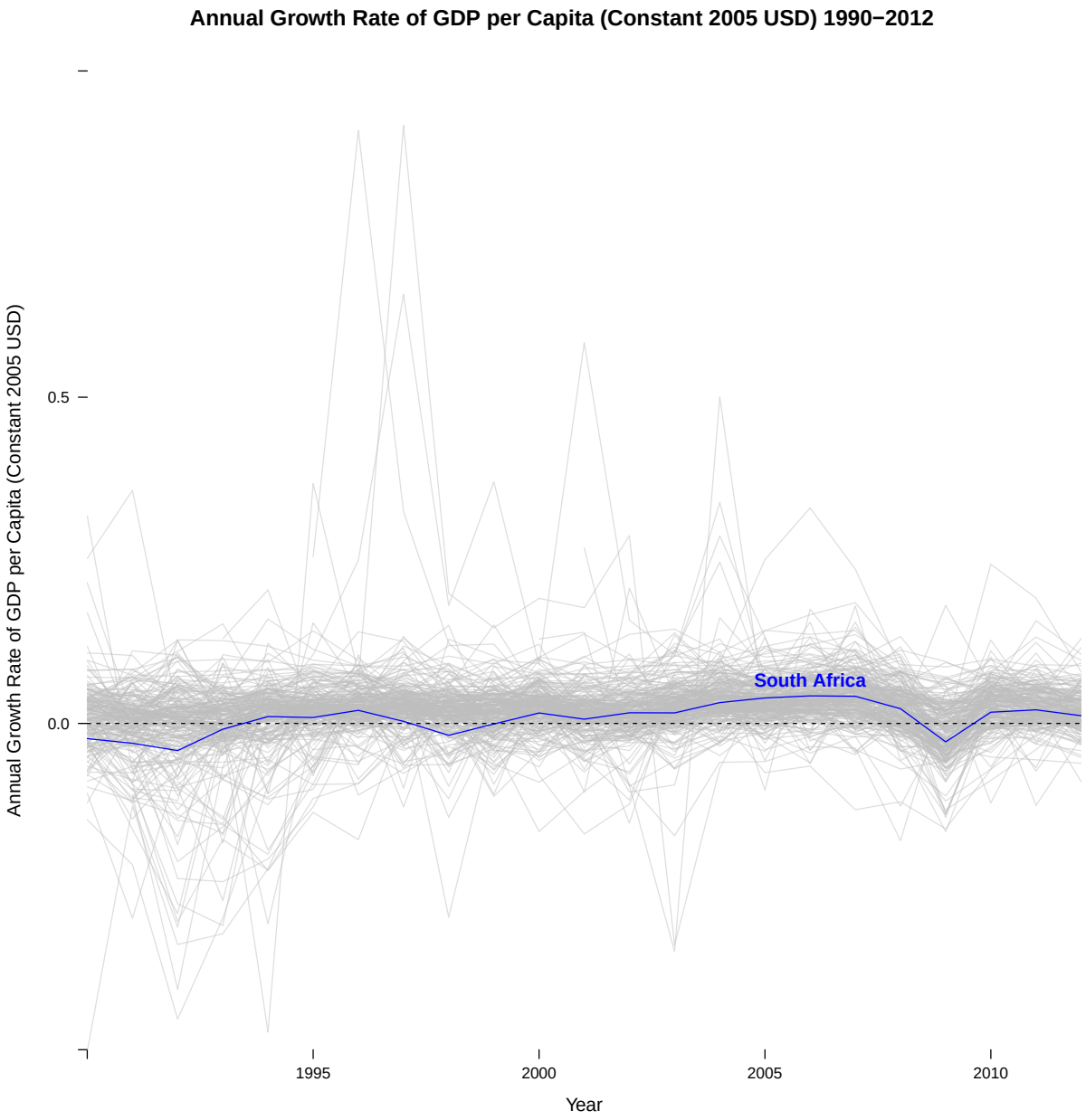


Figure B.2: Source: Author's calculations using World Bank data covering all countries for which data was available.

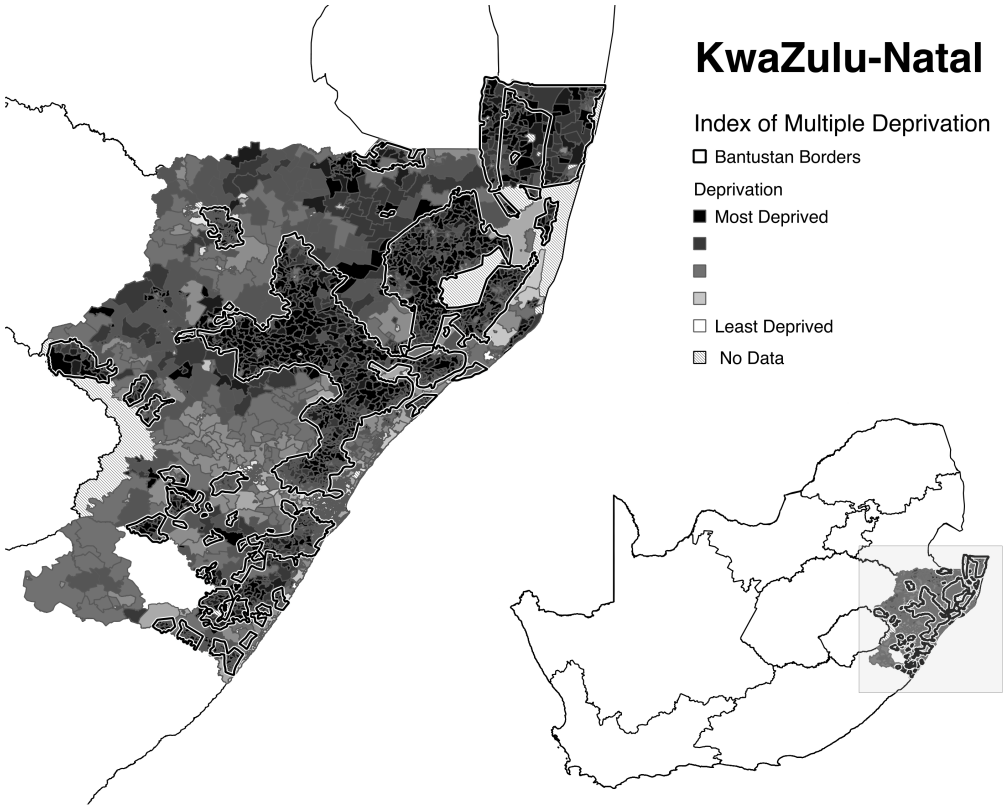
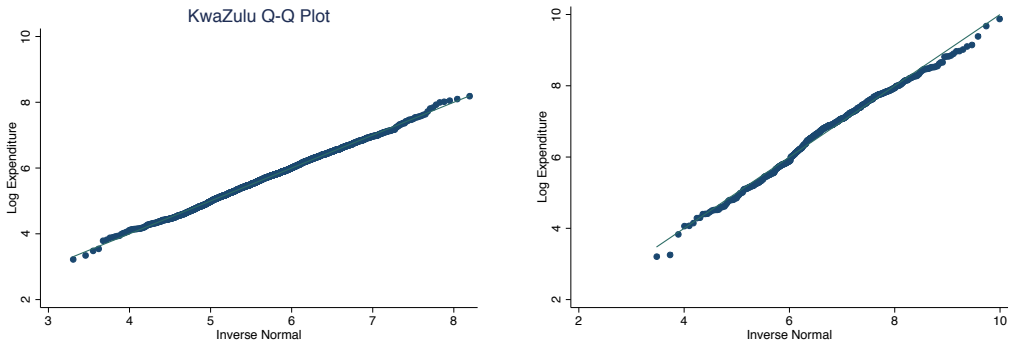


Figure B.3: Source: Author's calculations using data described in Noble et al. (2009). Each geographic division was designed to contain roughly 2,000 individuals. Darker areas are relatively more deprived and the former bantustan borders are highlighted. Areas in the former bantustans more consistently among the most deprived relative to former "white South Africa."



(a) KwaZulu (b) Natal

Figure B.4: Q-Q plots of logged monthly expenditure: KwaZulu and Natal

APPENDIX C
THEIL DECOMPOSITION

An alternate measure of inequality is the Theil coefficient. The Theil coefficient is additively decomposable into a between group inequality and within group inequality. This decomposition has been used to study the spatial dimension of inequality in China and Indonesia (Akita, 2003) and to decompose inequality within and between 46 countries (Novotný, 2007).

Define the Theil coefficient as follows:

$$T = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{y_i}{\bar{y}} \ln \frac{y_i}{\bar{y}} \right)$$

where y_i is the measure of income for household i, n is the number of households, and $\bar{y}$ is the mean income over n households. T measures the level of inequality in the population and is larger for more unequal populations.

Define k non-overlapping groups that contain all n members of the population. The Theil coefficient is additively decomposable into a component equal to the weighted average of the Theil coefficient within groups and a component equal to the Theil coefficient across group averages (i.e., inequality between groups).

$$T = \left(\sum_{j=1}^{k} \frac{n_j}{n} \frac{y_j}{\bar{y}} \ln \frac{y_j}{\bar{y}} \right) + \left(\sum_{j=1}^{k} \frac{1}{n} \frac{y_j}{\bar{y}} \sum_{i=1}^{n_j} \frac{y_{ij}}{y_j} \ln \frac{y_{ij}}{y_j} \right) \tag{C.1}$$

where

$$y_j = \frac{\sum_{i=1}^{n_j} y_{ij}}{n_j}$$

The first term in equation C.1 is the between-group coefficient and the second term is the weighted within-group coefficient. The ratio of the between-group coefficient to the overall Theil coefficient is analogous to the neighbor correlation defined in this paper. Table C.1 presents the neighbor correlations calculated using the Theil decomposition and the difference between tribal authority areas and the areas of former "white South Africa." These results confirm the significant differences in the neighbor correlations between these two

Table C.1: Theil Neighbor Correlations
Tribal and Non-Tribal Areas

	Tribal	Non-Tribal	Non-Tribal − Tribal
KwaZulu-Natal			
Neighbor Correlation	0.0791	0.3060	0.2269***
	[0.0611, 0.0981]	[0.2572, 0.3512]	[0.1750, 0.2761]
All South Africa			
Neighbor Correlation	0.0648	0.2910	0.2261***
	[0.0472, 0.0795]	[0.2323, 0.3293]	[0.1677, 0.2689]

Brackets define 95% confidence intervals calculated with 10,000 bootstrap samples. *, **, and *** indicate zero is not included in a 90%, 95%, and 99% confidence intervals respectively.

administrative designations found in the analysis using the REML variance decomposition above.

APPENDIX D

REML VS ML

The REML estimator is

$$\hat{\sigma}^2 = y'T'(THT')^{-1}Ty/(N-k)$$

where k is the number of fixed effects in the model (Corbeil and Searle, 1976), while the ML estimator is

$$\hat{\sigma}_i^2 = m_i^{-1}\{d_i'd_i + \sigma^2 tr(\Sigma_{22.1i}^{-1} + \gamma_i^{-1}I_{m_i})^{-1}\} \qquad (i = 1, 2, \ldots, c)$$

(Hartley and Rao, 1967). The incorporation of k causes the REML estimator to increase when adding a fixed effect to the model.

The Restricted Maximum Likelihood (REML) procedure and the Maximum Likelihood (ML) procedure produce similar results for the calculations presented in this paper. Tables D.1 and D.2 presents the output of the ML procedure that corresponds to the results in table A.6 and A.9.

Table D.1: Percent Change in the Difference (ML)
Between Tribal and Non-Tribal in South Africa (NIDS)

	Education	Agriculture	Urban	District	
Neighbor Correlation	-0.0934**	0.0030**	-0.0886***	-0.0749**	
	[-0.1799, -0.0228]	[0.0001, 0.0079]		[-0.1672, -0.0321]	[-0.3567, -0.0011]
Household Component	-0.1713	-0.0054	0.0042	-0.0138	
	[-0.724, 0.5249]	[-0.0541, 0.0184]	[-0.0259, 0.0401]	[-0.1288, 0.1485]	
Neighborhood Component	-0.3596***	0.0005	-0.1563***	-0.2607***	
	[-0.4359, -0.2852]	[-0.0012, 0.0029]	[-0.2635, -0.0653]	[-0.5906, -0.2394]	
Transitory Component	-0.1997	0.0302	-0.0001	-0.0025	
	[-1.9212, 1.9688]	[-0.3214, 0.3353]	[-0.0116, 0.0117]	[-0.0377, 0.0344]	

Brackets define 95% confidence intervals calculated with 10,000 bootstrap samples. *, **, and *** indicate zero is not included in a 90%, 95%, and 99% confidence intervals respectively.

Table D.2: Remaining Difference (ML)
Between Tribal and Non-Tribal in South Africa (NIDS)

	Education	Agriculture	Urban	District
Neighbor Correlation	0.3966***	0.4388***	0.3987***	0.4047***
	[0.2817, 0.4856]	[0.3234, 0.5255]	[0.2828, 0.4906]	[0.2426, 0.4619]
Household Component	0.0336**	0.0403**	0.0407**	0.0399**
	[0.0019, 0.0662]	[0.0025, 0.0791]	[0.0028, 0.0794]	[0.0026, 0.0789]
Neighborhood Component	0.1496***	0.2337***	0.1971***	0.1727***
	[0.0947, 0.196]	[0.1498, 0.305]	[0.1231, 0.2598]	[0.0778, 0.1889]
Transitory Component	0.0110	0.0142	0.0138	0.0137
	[-0.0285, 0.0512]	[-0.0256, 0.0541]	[-0.0261, 0.0536]	[-0.0261, 0.0535]

Brackets define 95% confidence intervals calculated with 10,000 bootstrap samples. *, **, and *** indicate zero is not included in a 90%, 95%, and 99% confidence intervals respectively.

APPENDIX E

ALTERNATE EDUCATION MEASURES AND THE NEIGHBORHOOD COMPONENT

The measure of education used in this paper is the mean years of education for the household members that are at least 21 years old. Table E.1 presents the results of exchanging the main measure of education with five alternate measures which vary the age threshold and whether the mean or the max is calculated. In each case, introducing the measure of education significantly lowers the difference in the neighborhood variance component between KwaZulu (the bantustan) and Natal (the former province of "white South Africa"). 99% confidence intervals calculated with 10,000 bootstrap samples did not include zero in any of the presented specifications.

Table E.1: Percent Change to the Neighborhood Component Using Various Education Measures

	Estimate	Lower Bound	Upper Bound
Mean Years of Education			
>= 16 years	-0.4682	-0.6575	-0.3261
>= 21 years	-0.4412	-0.6301	-0.2772
>= 26 years	-0.4032	-0.6050	-0.2622
Max Years of Education			
>= 16 years	-0.3627	-0.4949	-0.1648
>= 21 years	-0.3147	-0.4584	-0.0971
>= 26 years	-0.3227	-0.4934	-0.1806

Estimates are for percent changes to the neighborhood variance component calculated using the REML procedure.

APPENDIX F

TABLES FOR CHAPTER 2

Table F.1: Comparison of Samples.

	PSID	PSID		IFLS	
	Solon et al.	Author's Sample		Author's Sample	
	(2000)	Singletons	Siblings	Singletons	Siblings
Siblings	687	1296	976	2654	729
Families	379	593	315	1451	223
Clusters	144	85	63	281	87
Siblings per Family (mean)	1.8	2.2	3.1	1.8	3.3
Years of Schooling (mean)	13.3	13.6	13.7	11.1	11.8
Age at Measurement (mean)	29.1	29.1	29	25.9	26.1
Proportion Male	0.5	0.51	0.52	0.5	0.47
Proportion Black	0.1	0.07	0.06		
Proportion Urban		0.22	0.22	0.56	0.69

Comparison of Solon et al. (2000) sample to the PSID and IFLS sample used in this paper. The mean in the IFLS is inflated by two years due to the calculation of years of schooling in Indonesia, which is presented in appendix H.

Table F.2: Comparison of Male and Female Mean Years of Schooling.

	PSID	IFLS
Male	13.91	11.72
Female	13.41	11.95
Male - Female	0.500	-0.233
	[0.242, 0.725]	[-0.830, 0.364]

Numbers in square brackets are the 2.5 and 97.5 percentiles of 200 bootstrap replications. The samples compared are families with at least two siblings.

Table F.3: Comparison of Urban and Rural Mean Years of Schooling.

	PSID	IFLS
Urban	14.33	12.39
Rural	13.48	10.59
Urban - Rural	0.847	1.808
	[0.432, 1.234]	[1.114, 2.486]

Numbers in square brackets are the 2.5 and 97.5 percentiles of 200 bootstrap replications. The samples compared are families with at least two siblings.

Table F.4: Replication of Solon et al. (2000) and comparison with results from REML.

	Solon et al. (2000)	PSID* with method from Solon et al. (2000)	PSID* with REML
Sibling Correlation	0.513	0.416	0.484
	(0.045)	[0.299, 0.513]	[0.384, 0.548]
Neighbor Correlation	0.153	0.067	0.012
	(0.055)	[-0.010, 0.161]	[0.000, 0.061]
Adjusted Neighbor	0.083	0.016	0.000
Correlation	(0.067)	[-0.057, 0.097]	[0.000, 0.020]

PSID* is the PSID sample I constructed following the restrictions in Solon et al. (2000), but does not include the restricted use geographic data. This table compares data which appears in table 1 of Solon et al. , numbers in parenthesis are the standard errors reported by Solon et al. (2000). Numbers in square brackets are the 2.5 and 97.5 percentiles of 200 bootstrap replications.

Table F.5: Solon et al. (2000) Methodology Applied to PSID and IFLS.

	Solon et al. (2000)	PSID*	IFLS	Urban IFLS	Rural IFLS
Sibling Covariance	2.75‡	1.772	5.339	3.359	6.502
		[1.233, 2.282]	[3.911, 6.901]	[2.076, 4.686]	[3.854, 9.168]
Neighborhood Covariance	0.832‡	0.296	3.146	0.711	3.386
		[-0.042, 0.713]	[1.780, 4.743]	[0.026, 1.624]	[1.153, 5.965]
Neighborhood Covariance	0.386‡	0.226	0.54	0.255	0.483
of Family Observables		[0.102, 0.400]	[0.354, 0.848]	[0.132, 0.452]	[0.178, 1.004]
Total Variance	5.42	4.4	13.681	12.069	14.686
(with singletons)		[3.931, 4.835]	[12.00, 15.25]	[10.12, 13.75]	[12.06, 17.30]
Total Variance	5.37	4.256	13.883	12.361	13.848
(without singletons)		[3.754, 4.731]	[13.04, 14.87]	[11.48, 13.17]	[12.23, 15.15]
Sibling Correlation	0.513	0.416	0.39	0.278	0.443
	(0.045)	[0.299, 0.513]	[0.306, 0.471]	[0.182, 0.358]	[0.280, 0.582]
Neighbor Correlation	0.153	0.067	0.227	0.058	0.245
	(0.055)	[-0.010, 0.161]	[0.131, 0.330]	[0.002, 0.127]	[0.093, 0.432]
Adjusted Neighbor Correlation	0.083	0.016	0.188	0.037	0.210
	(0.067)	[-0.057, 0.097]	[0.094, 0.284]	[-0.025, 0.102]	[0.060, 0.384]

PSID* is the PSID sample I constructed following the restrictions in Solon et al. (2000). ‡ indicates values that were calculated from reported values in Solon et al., numbers in parenthesis are the estimated standard errors reported in Solon et al., numbers in square brackets are the 2.5 and 97.5 percentiles of 200 bootstrap replications.

Table F.6: REML Applied to PSID and IFLS.

	PSID Educational Attainment	PSID with Adjustment	IFLS Educational Attainment	IFLS with Adjustment
Family Variance Component	2.017	1.218	3.398	2.793
	[1.491, 2.477]	[0.762, 1.568]	[1.525, 4.714]	[1.198, 3.593]
Neighborhood Var. Comp.	0.053	0	0.579	0.082
	[0.000, 0.257]	[0.000, 0.069]	[0.000, 1.918]	[0.000, 1.323]
Residual Var. Comp.	2.204	2.212	9.836	9.672
	[1.921, 2.528]	[1.928, 2.523]	[8.472, 10.91]	[8.293, 10.73]
Sibling Correlation	0.484	0.355	0.288	0.229
	[0.384, 0.548]	[0.239, 0.424]	[0.212, 0.347]	[0.161, 0.281]
Neighbor Correlation	0.012	0	0.042	0.007
	[0.000, 0.061]	[0.000, 0.020]	[0.000, 0.151]	[0.000, 0.116]

Numbers in square brackets are the 2.5 and 97.5 percentiles of 200 bootstrap replications. The adjustment in columns two and four is the addition of the family observable factors.

Table F.7: Effect of Family Observables on REML Correlation Estimates.

	PSID	IFLS
Sibling Correlation Adjustment	0.129	0.059
	[0.095, 0.189]	[0.018, 0.115]
Neighbor Correlation	0.012	0.035
Adjustment	[0.000, 0.058]	[-0.012, 0.078]

Effects are the difference between correlations in the adjusted and unadjusted model specifications. Numbers in square brackets are the 2.5 and 97.5 percentiles of 200 bootstrap replications.

APPENDIX G
EXPECTED AGE OF SCHOOLING COMPLETION

Solon et al. (2000) restrict their sample based on the assumption that for most individuals in the United States complete their schooling by age 25, but it is not obvious what year should be the cutoff in the Indonesian context. To compare the expectations of schooling completion in the US and Indonesia I estimated probabilities of school attendance by sex and age. For Indonesia, I also stratify by urban status since Maralani (2008), Takahashi (2011), and others have identified significant differences in enrollment rates between urban and rural areas (see table F.3).

The argument for the lower bound on the age of the sample depends on the expected age of schooling completion in the year of the relevant survey. The Solon et al. (2000) study of educational attainment focuses on survey responses for the 1985 PSID. Their argument for the lower age bound for their sample is that those who were 25 years old in 1985 would be expected to have completed their education. This paper focuses on the fourth round of the IFLS that was conducted in late 2007 and early 2008. To form expectations on the age of schooling completion I compare data from the 1980 and 1990 US Census to the 2005 Intercensal Population Survey and the 2010 Population Census of the Indonesian Central Bureau of Statistics (see figures G.1-G.4).

For males, the age at which education can be expected to be completed is lower in Indonesia than in the US. In 1980, the census year before the 1985 PSID from which the education values in Solon et al. (2000) are estimated, the expected probability of school attendance for males at age 25 was 13.8% (see figure G.1). In 2005, the year of the intercensal population survey before the 2007/2008 IFLS from which the education values in this paper are estimated, males in rural (urban) areas had a school attendance probability below that threshold by age 19 (23) at 13.0% (10.4%) (see figure G.1). In 1990, the census year after the 1985 PSID, the expected probability of school attendance for males at age 25 was 17.4% (see figure G.2). In 2010, the year of the population census after the 2007/2008 IFLS, males in rural (urban) areas had a school attendance probability below that threshold by age 20 (23) at 11.7% (12.8%) (see figure G.2).

The age of average schooling completion is lower in Indonesia than in the US and is also lower for females. In 1980, the census year before the 1985 PSID, the expected probability of school attendance for females at age 25 was 10.9% (see figure G.3). In 2005, the year of the intercensal population survey before the 2007/2008, females in rural (urban) areas had a school attendance probability below that threshold by age 20 (23) at 5.2% (7.9%) (see figure G.3). In 1990, the census year after the 1985 PSID, the expected probability of school attendance for females at age 25 was 16.1% (see figure G.4). In 2010, the year of the population census after the 2007/2008 IFLS, females in rural (urban) areas had a school attendance probability below that threshold by age 20 (23) at 11.2% (10.2%) (see figure G.4).

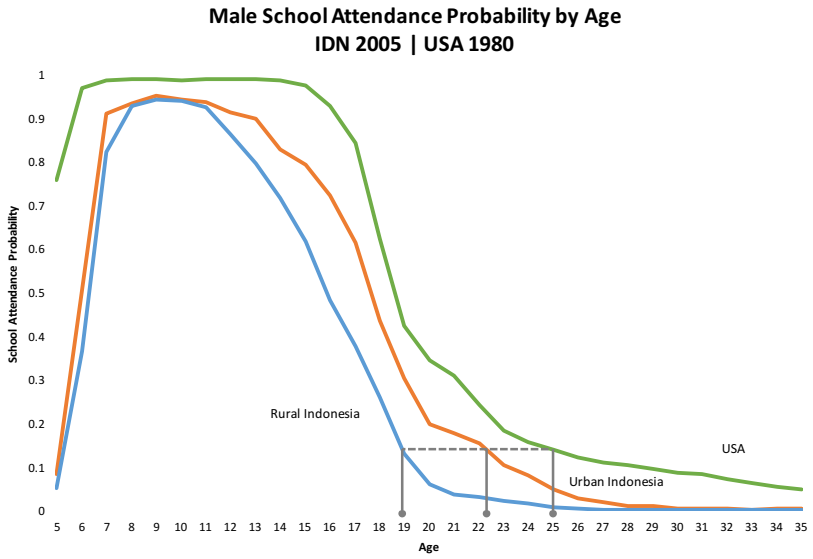


Figure G.1: Male school attendance by age before the relevant education survey. Source: Authors calculations of the 1980 US Census Bureau Census of Population and Housing, and the 2005 Indonesia Central Bureau of Statistics Intercensal Population Survey.

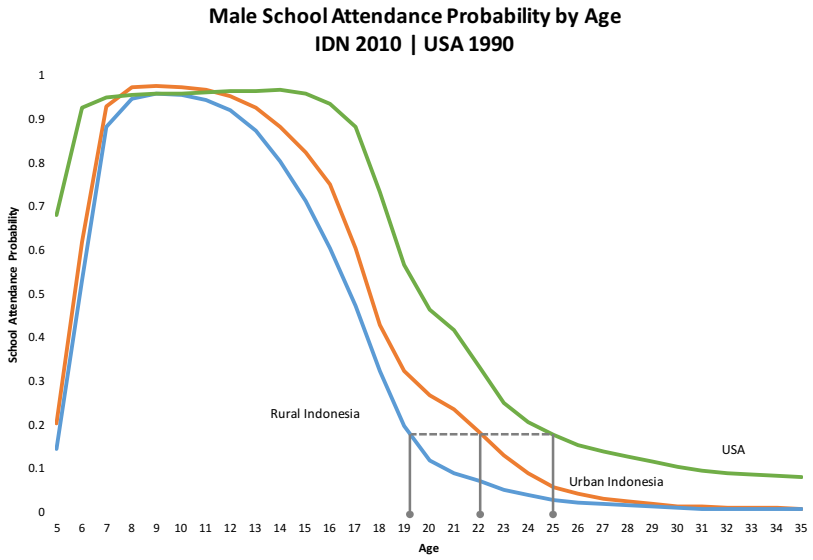


Figure G.2: Male school attendance by age after the relevant education survey. Source: Authors calculations of the 1990 US Census Bureau Census of Population and Housing, and the 2010 Indonesia Central Bureau of Statistics Population Census.

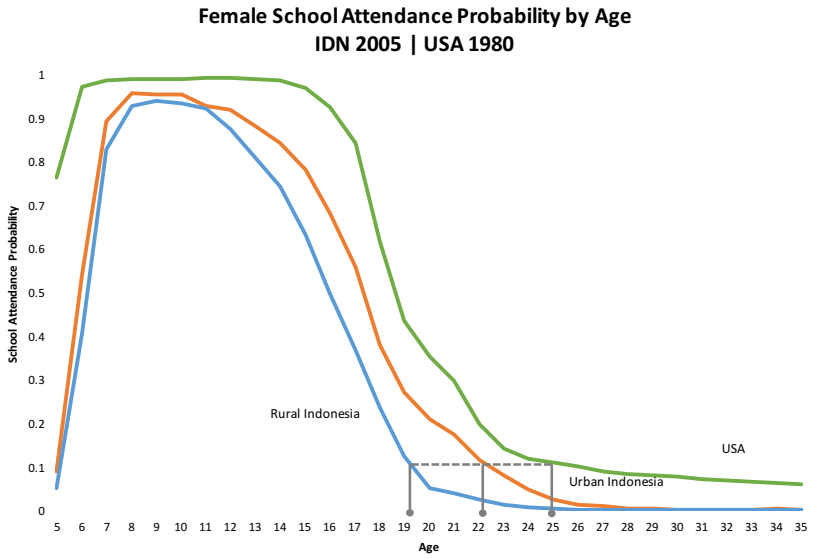


Figure G.3: Female school attendance by age before the relevant education survey. Source: Authors calculations of the 1980 US Census Bureau Census of Population and Housing, and the 2005 Indonesia Central Bureau of Statistics Intercensal Population Survey.

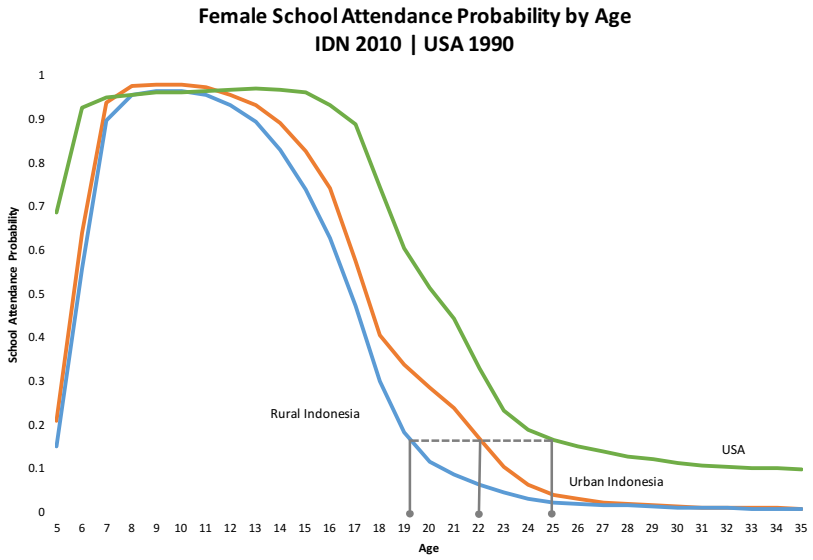


Figure G.4: Female school attendance by age after the relevant education survey. Source: Authors calculations of the 1990 US Census Bureau Census of Population and Housing, and the 2010 Indonesia Central Bureau of Statistics Population Census.

APPENDIX H
YEARS OF EDUCATION ESTIMATION

The education system in Indonesia is composed of a variety of general, vocational, centralized and Islamic schools. Table H.1 identifies the years of schooling and age estimates provided by the Indonesian Ministry of Education and Culture (MOEC, 2013). MOECs years of schooling are two years ahead of similar programs in the US due to the counting of two years of kindergarten. In this study, I subtract two years of schooling from the years reported by MOEC. In the IFLS survey participants are able to specify if they attended college, but may only report up to 7 years of post secondary education. The years of education are thus top-coded at 21 years.

Table H.1: Years of Schooling, Age, and Grade in the US and Indonesia.

MOEC (2013) Age Estimate	MOEC (2013) Years of Schooling	My Assignment IFLS Years of Schooling	Education Levels	Solon et al. (2000) PSID Years of Schooling
Above 22	23	21	Doctoral	
	22	20		20
	21	19		19
	20	18	Master	18
	19	17		17
22	18	16	Undergraduate	16
21	17	15		15
20	16	14		14
19	15	13		13
18	14	12	Senior High	12*
17	13	11		
16	12	10		
15	11	9	Junior High	
14	10	8		
13	9	7		
12	8	6	Elementary	
11	7	5		
10	6	4		
9	5	3		
8	4	2		
7	3	1		

* In Solon et al. (2000) years of schooling are only adjusted when information about high school graduation and above are provided.

APPENDIX I

COMPARISON OF CLUSTER AND NEIGHBORHOOD COVARIANCE

I assume the cluster covariance (and correlation) I calculate will be smaller than the neighborhood covariance (and correlation) calculated by Solon et al. (2000). Assume the neighborhood effect in the model above is composed of a cluster effect, ρ_c, and neighborhood effect, α_{cn}, (with neighborhood n nested within cluster c). The model of years of schooling becomes

$$y_{cnfs} = \rho_c + \alpha_{cn} + b_{cnf} + e_{cnfs}$$

As above, I assume the individual component, e_{cnfs}, is uncorrelated with the remaining effects in the model.

With this expanded model, the neighbor covariance in Solon et al. (2000) becomes

$$\begin{aligned} Cov(y_{cnfs}, y_{cnf's}) = & Var(\rho_c) + 2Cov(\rho_c, \alpha_{cn}) + Var(\alpha_{cn}) \\ & + 2Cov(\rho_c, b_{cnf}) + 2Cov(\alpha_{cn}, b_{cnf}) + Cov(b_{cnf}, b_{cnf'}). \end{aligned}$$

The cluster covariance I calculate using the PSID becomes

$$\begin{aligned} Cov(y_{cnfs}, y_{cn'fs}) = & Var(\rho_c) + 2Cov(\rho_c, \alpha_{cn}) + Cov(\alpha_{cn}, \alpha_{cn'}) \\ & + 2Cov(\rho_c, b_{cnf}) + 2Cov(\alpha_{cn}, b_{cnf}) + Cov(b_{cnf}, b_{cn'f}). \end{aligned}$$

The difference between the neighbor covariance estimated in Solon et al. and the one estimated in this paper is

$$
\begin{aligned}
Cov(y_{cnfs}, y_{cnf's}) - Cov(y_{cnfs}, y_{cn'fs}) = Var(\alpha_{cn}) - Cov(\alpha_{cn}, \alpha_{cn'}) \\
+ Cov(b_{cnf}, b_{cnf'}) - Cov(b_{cnf}, b_{cn'f})
\end{aligned}
$$

I analyze the sign of the first difference, $Var(a_{cn}) - Cov(a_{cn}, a_{cn'})$, by exploring the corresponding neighborhood variance and covariance for a particular cluster (i.e., $Var(a_{cn}) - Cov(a_{cn}, a_{cn'})$). Analyzing the sample analog of the relevant expectation provides intuition on the above sign. This simpler difference can be estimated as follows

$$
\begin{aligned}
Var(a_n) - Cov(a_n, a_{n'}) &= E(a_n^2) - E(a_n a_{n'}) \\
&= \frac{\sum_{i=1}^{N} a_i^2}{N} - \frac{\sum_{i=1}^{N} \sum_{j \neq i} a_i a_j}{N\,(N-1)} \\
&= \frac{(N-1) \sum_{i=1}^{N} a_i^2 - \sum_{i=1}^{N} \sum_{j \neq i} a_i a_j}{N\,(N-1)} \\
&= \frac{\sum_{i=1}^{N} a_i \left[(N-1)\, a_i - \sum_{j \neq i} a_j \right]}{N\,(N-1)} \\
&= \frac{\sum_{i=1}^{N} a_i \left[N a_i - \sum_{i=1}^{N} a_i \right]}{N\,(N-1)}
\end{aligned}
$$

I assume the neighbor effects are random with mean zero, so that the above difference is approximately $\frac{N}{N-1} Var(a_n)$.

Thus, the first difference is nonnegative and is strictly positive when the variance of the neighborhood effect is nonzero.

The second difference is positive when the family effect is more similar within neighborhoods than within clusters (which may be as large as counties). I assume the assortative matching of advantaged families into advantaged neighborhoods is stronger than the assortative matching of advantaged families into advantaged counties due to the relative ease of relocation across neighborhoods vs. across counties. Thus, I assume the second difference to be positive. This implies my estimate of the neighbor covariance (and hence the neighbor correlation) will be smaller than the measures reported in Solon et al. (2000).

APPENDIX J

TABLES FOR CHAPTER 3

Table J.1: Number of Observations by Age.

Age	Observations
25	7646
30	5716
35	4497
40	3456
45	2515
50	1678
55	999
60	455
65	93

APPENDIX K

FIGURES FOR CHAPTER 3

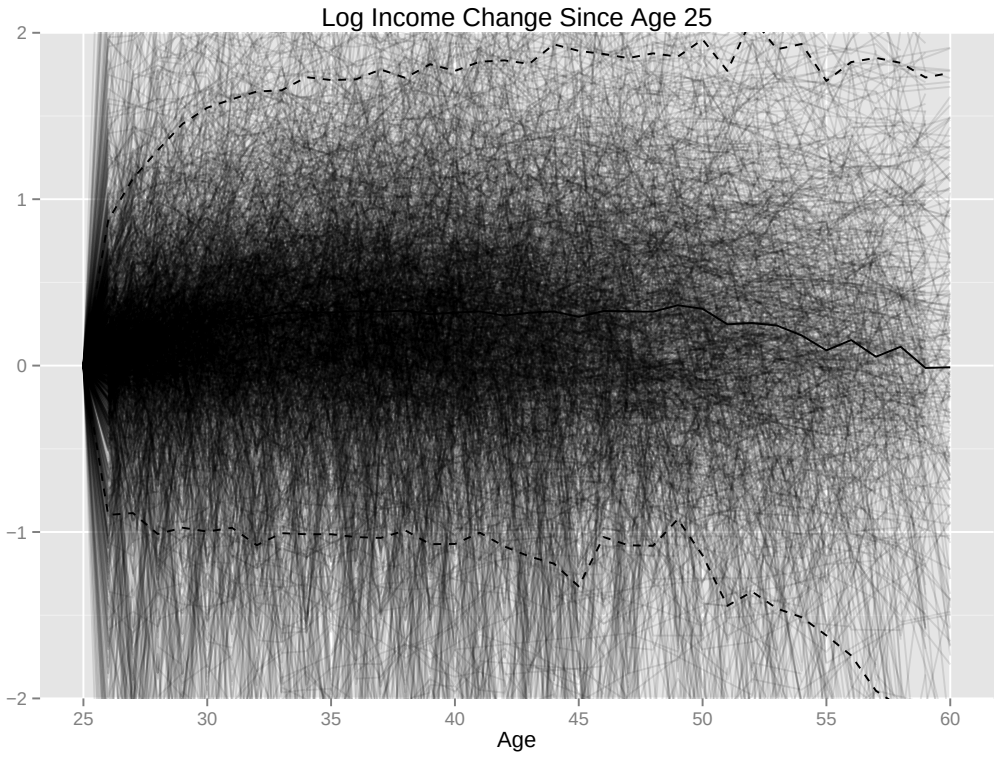


Figure K.1: Individual age profiles of log income change since age 25. The y axis is the difference of the log household-head income at the given age and the log household-head income at age 25. Individual profiles are shown with transparency. The solid line is the mean. The dotted lines represent the 5th and the 95th percentiles.

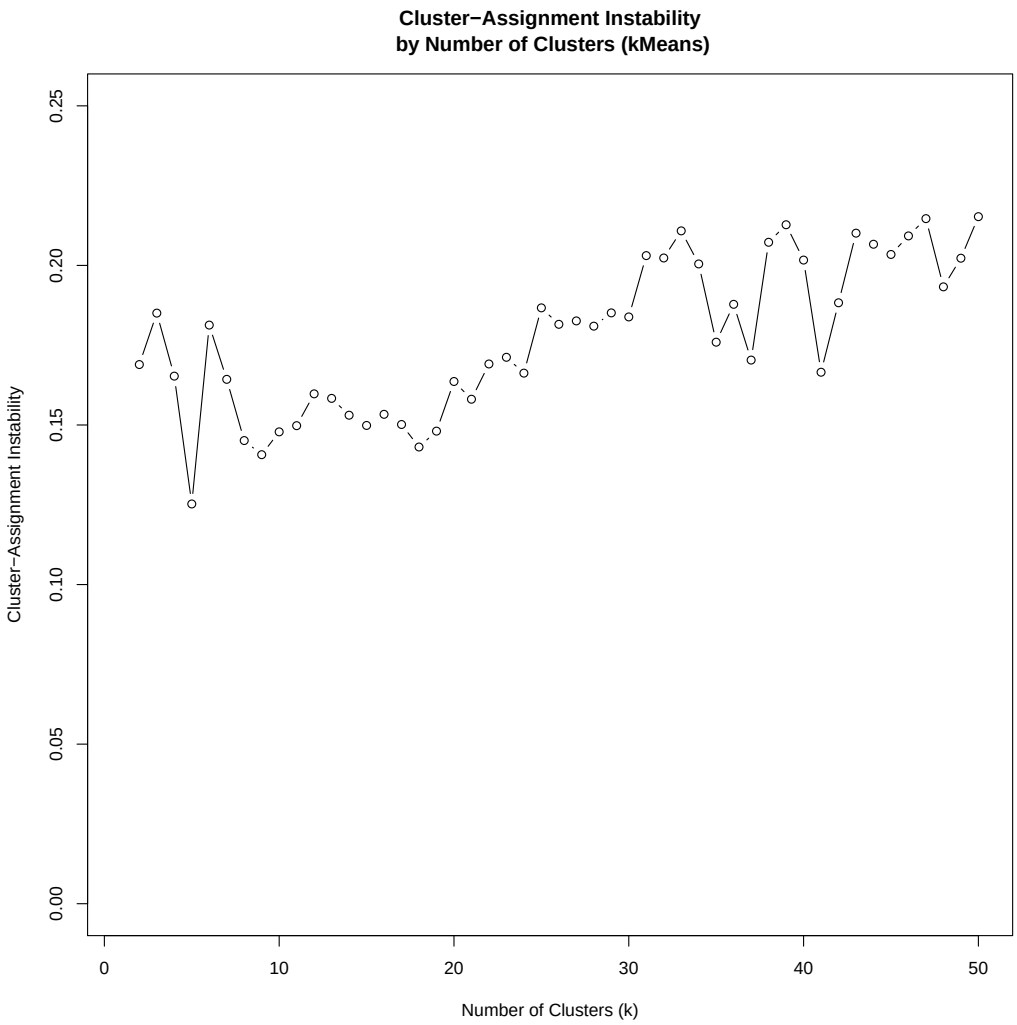


Figure K.2: Clustering assignment instability. The first point is where the number of clusters, k, is 2. There are local minima at 5 clusters and 9 clusters.

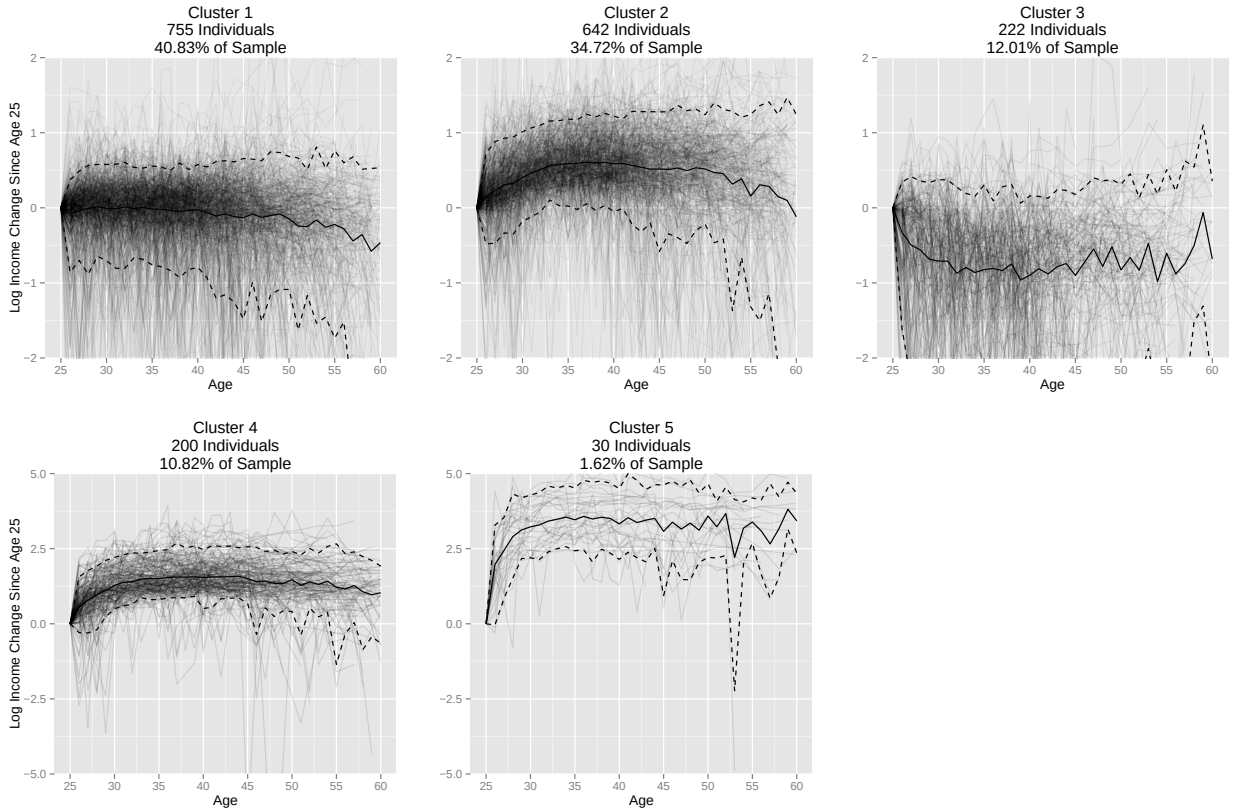


Figure K.3: Age profiles of log income change since 25 by cluster ($k = 5$). The solid line indicates the mean age profile of income. The dashed lines indicate the 5th and 95th percentile.

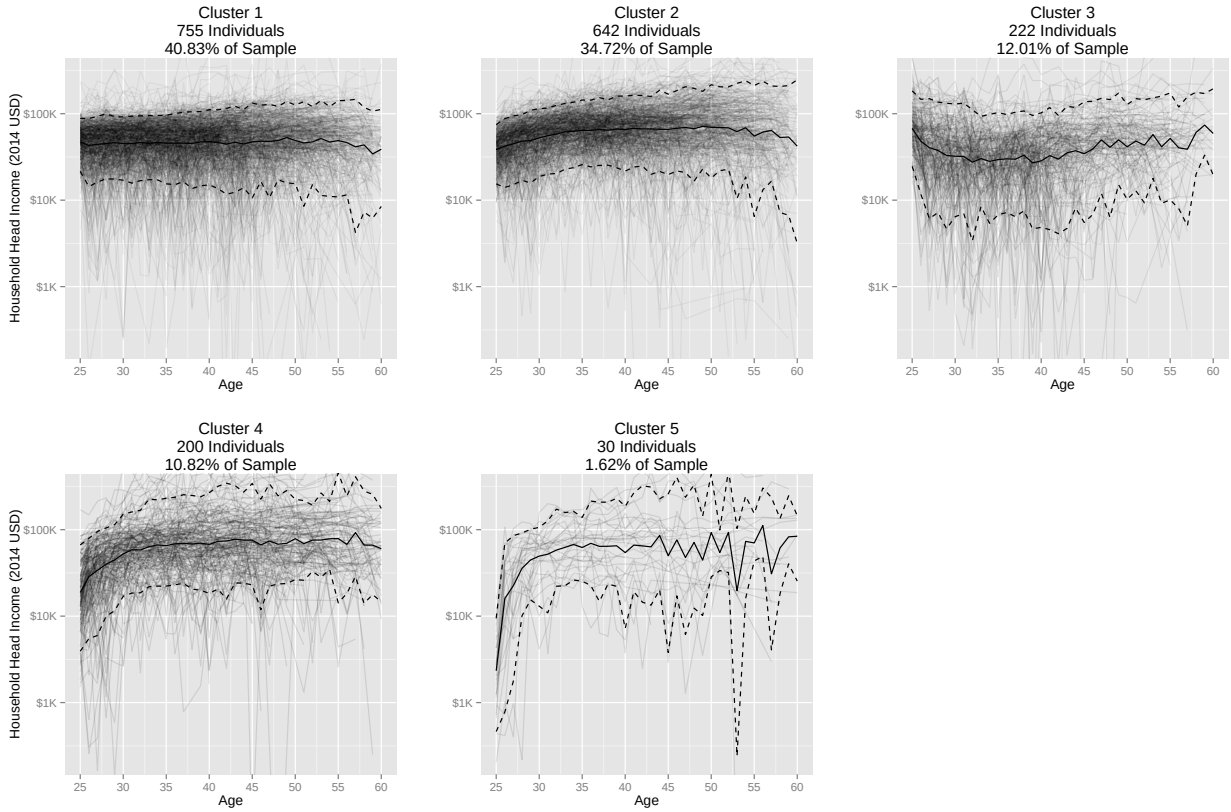


Figure K.4: Age profiles of log income by cluster ($k = 5$). The solid line indicates the mean age profile of income. The dashed lines indicate the 5th and 95th percentile.

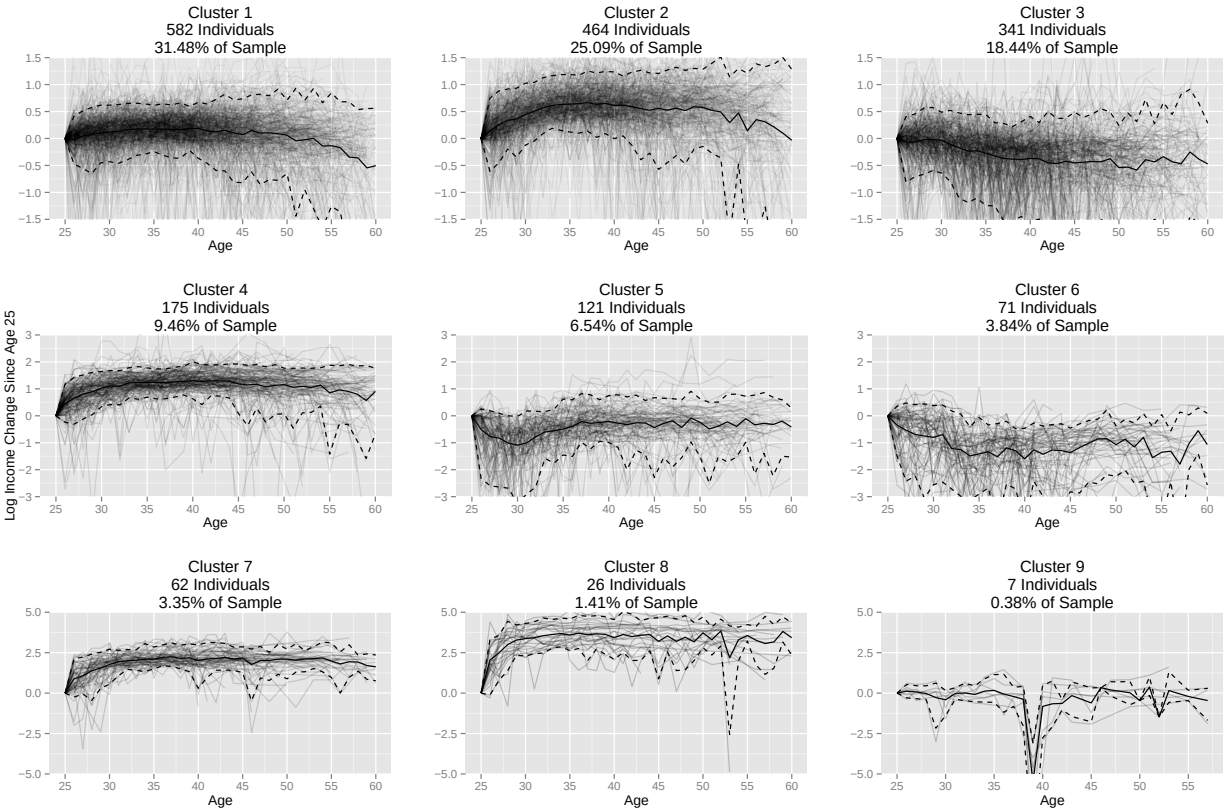


Figure K.5: Age profiles of log income change since 25 by cluster ($k = 9$). The solid line indicates the mean age profile of income. The dashed lines indicate the 5th and 95th percentile.

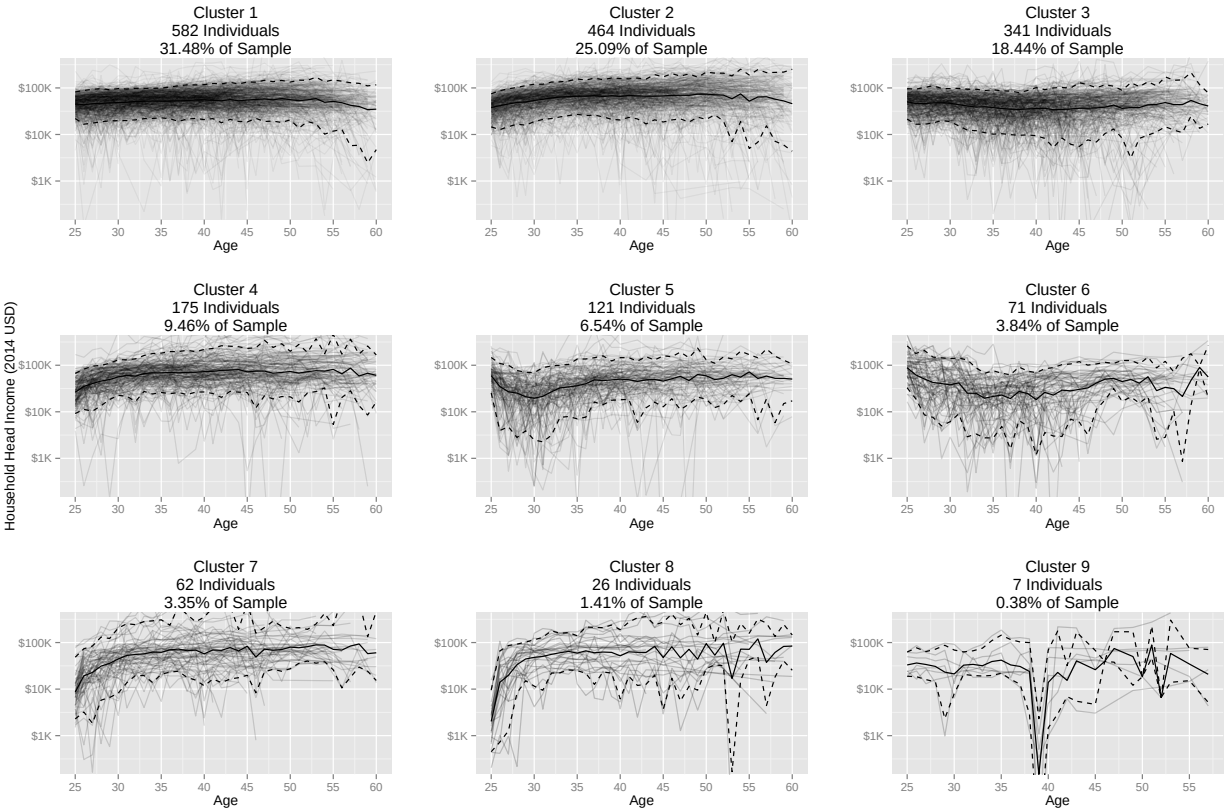


Figure K.6: Age profiles of log income by cluster ($k = 9$). The solid line indicates the mean age profile of income. The dashed lines indicate the 5th and 95th percentile.

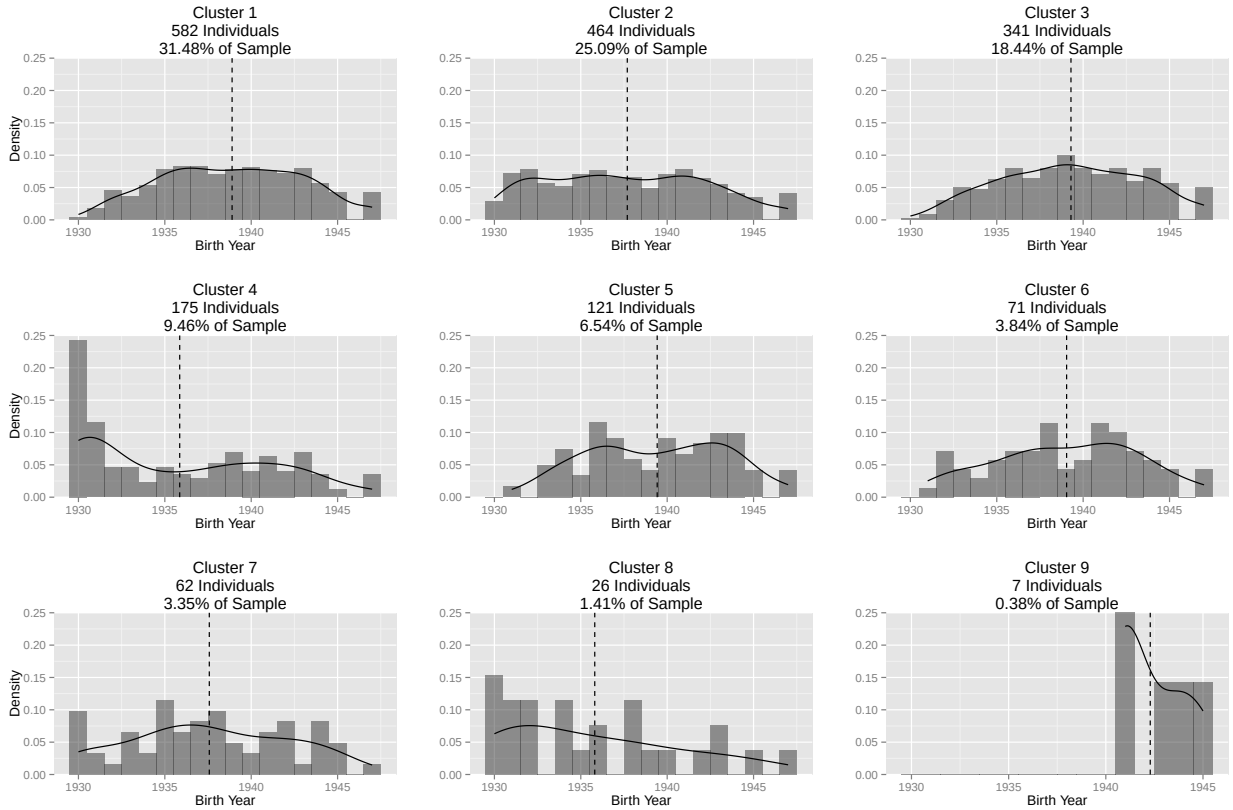


Figure K.7: Birth year distributions by cluster ($k = 9$).

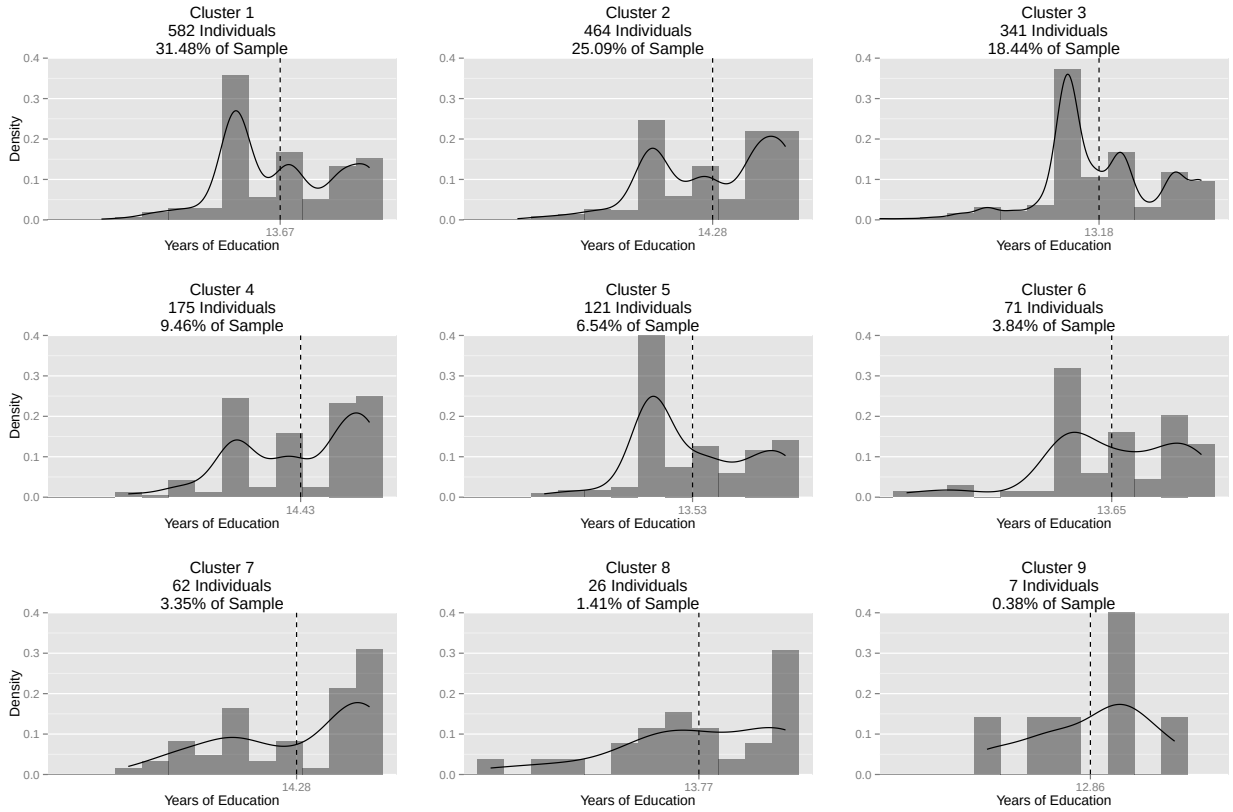


Figure K.8: Years of education distribution by cluster ($k = 9$).

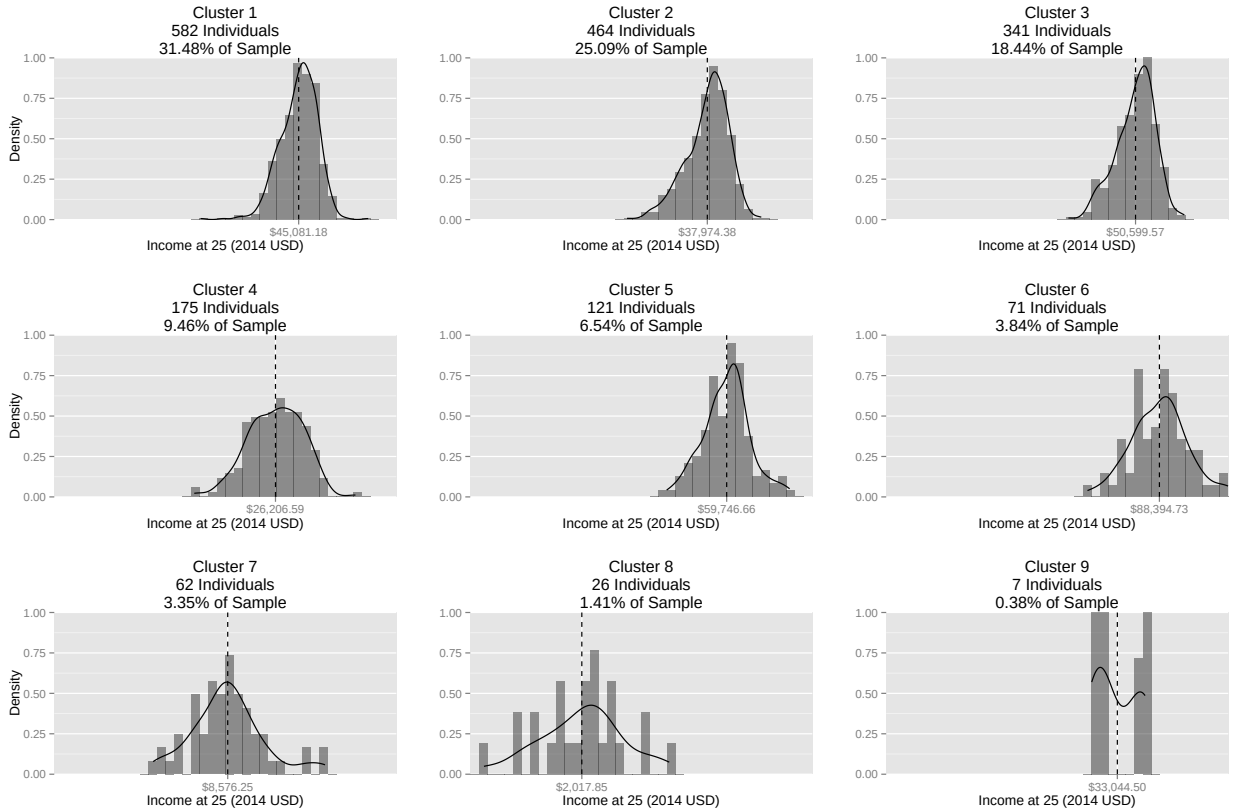


Figure K.9: Income of Household Head at Age 25 by Cluster ($k = 9$). The x axis is in logarithmic scale.

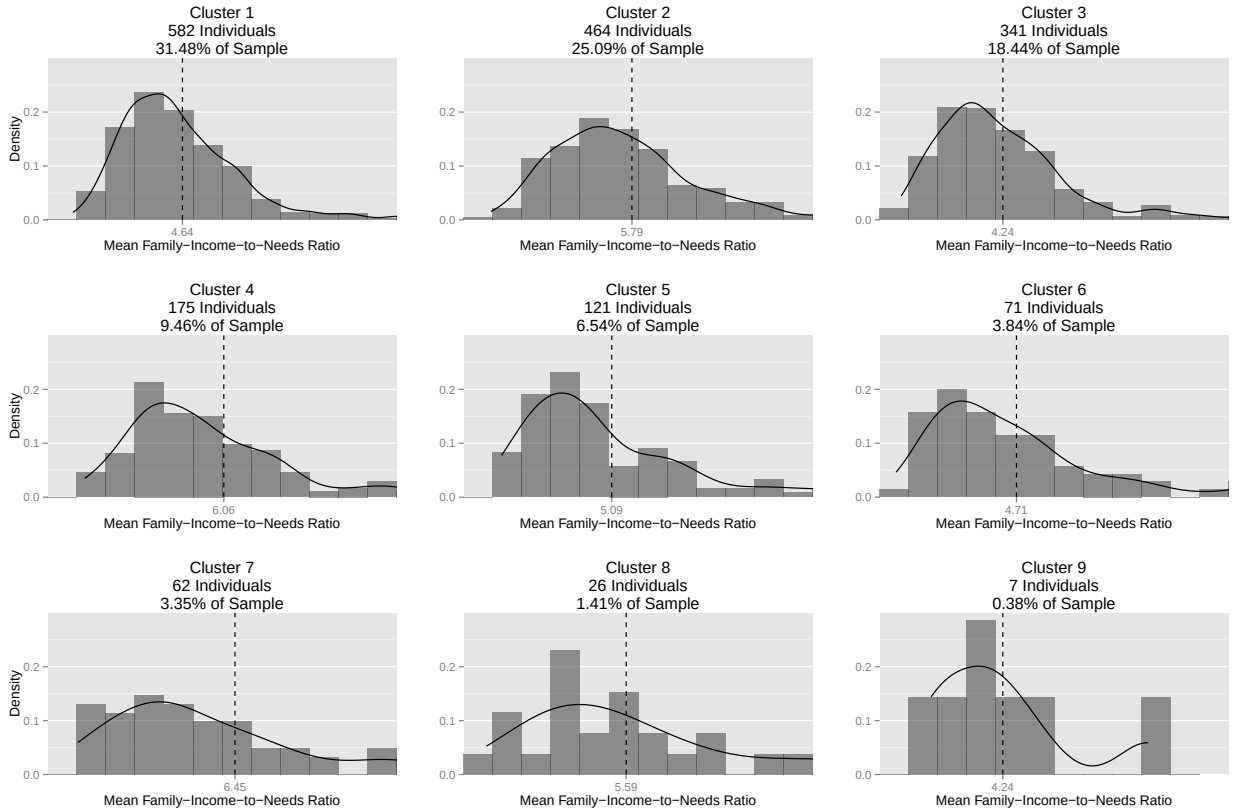


Figure K.10: Mean Family-Income-to-Needs Ratio by Cluster ($k = 9$).

APPENDIX L

SUMMARY OF CLUSTERS WHEN THE NUMBER OF CLUSTERS IS FIVE

For ease of discussion when setting the number of clusters to five (see figures 3 and 4 above), I label cluster 1 *early retirement*, cluster 2 *hump shaped*, cluster 3 *U shaped*, and clusters 4 and 5 *late bloomers*. Figure 11 shows the late bloomers (clusters 4 and 5) have a higher than average concentration of individuals born in 1930. In figure 12, the late bloomers (4 and 5) and hump shaped (2) clusters are relatively more highly educated than the early retirement (1) and U shaped (3) clusters. Again, figure 13 shows late bloomers (4 and 5) have the lowest starting incomes. Finally, figure 14 confirms the finding above with early retirement (1) and U shaped (3) groups have lower lifetime family-to-income-needs ratios.

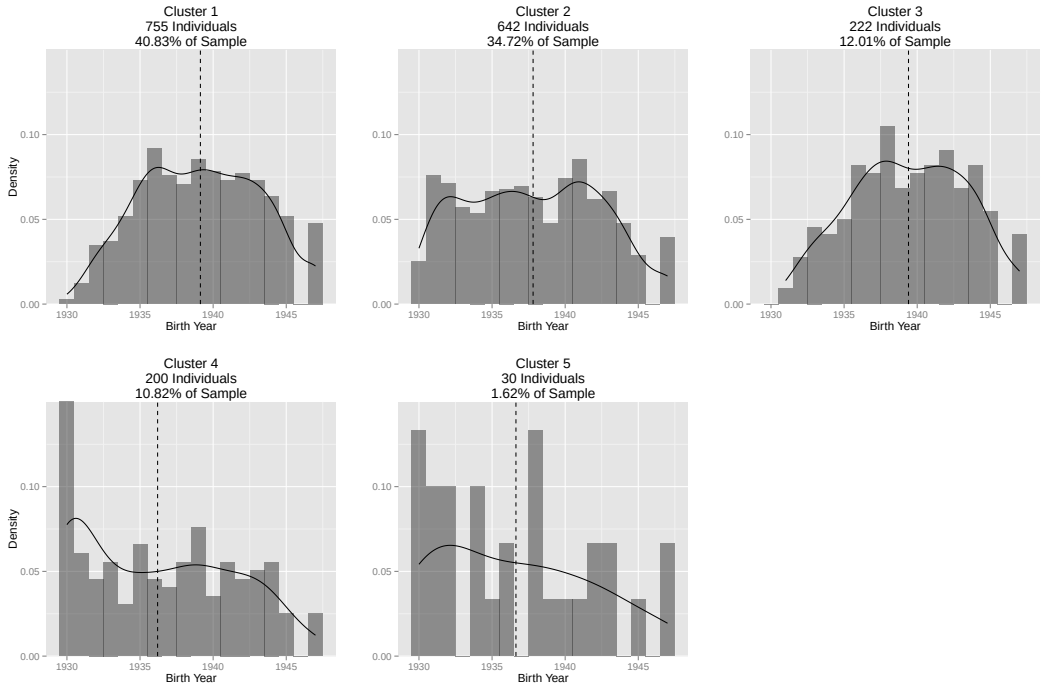


Figure L.1: Birth year distributions by cluster ($k = 5$).

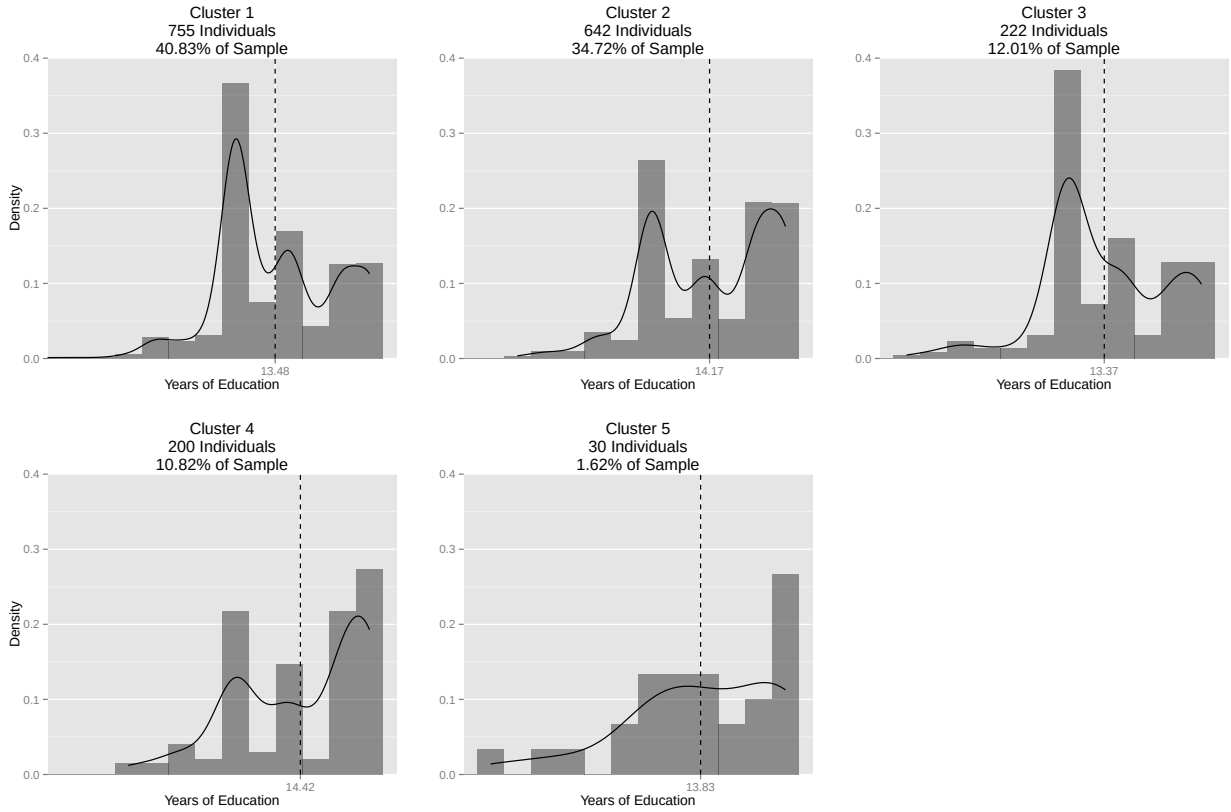


Figure L.2: Years of education distribution by cluster ($k = 5$).

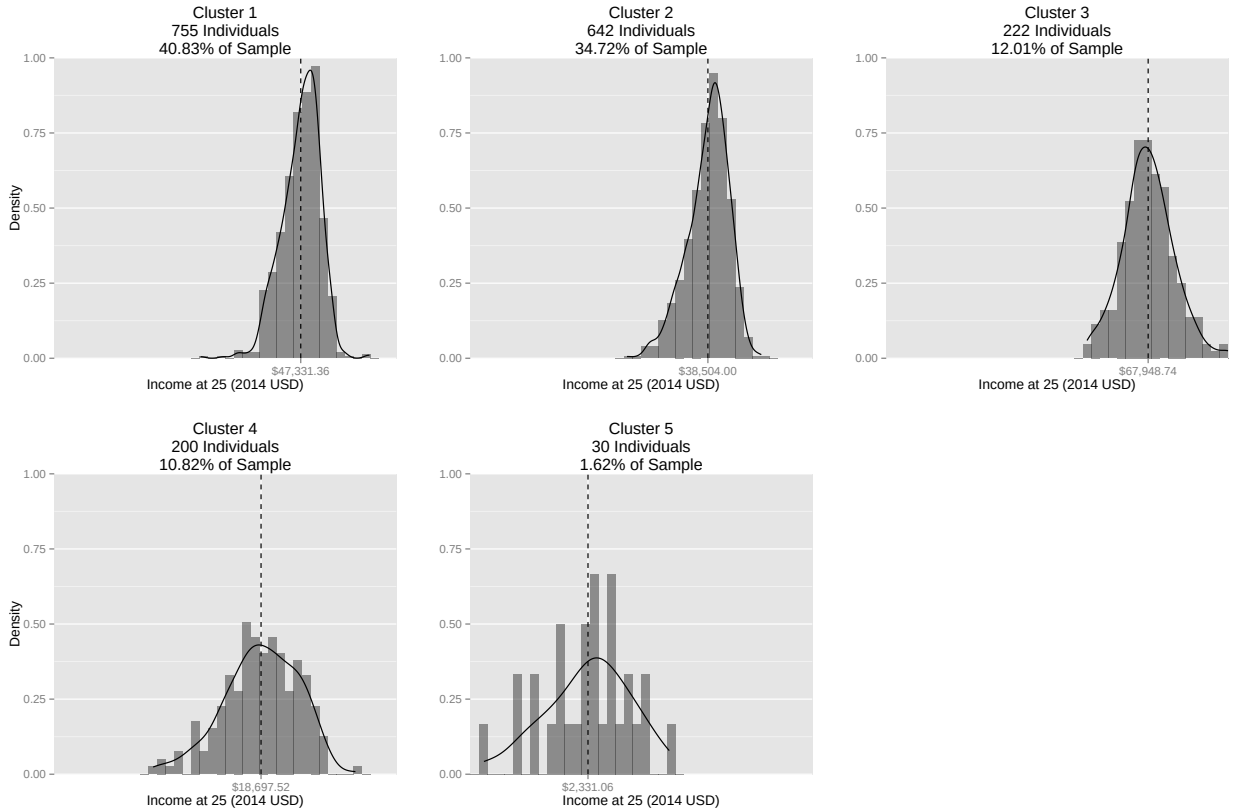


Figure L.3: Income of Household Head at Age 25 by Cluster ($k = 5$). The x axis is in logarithmic scale.

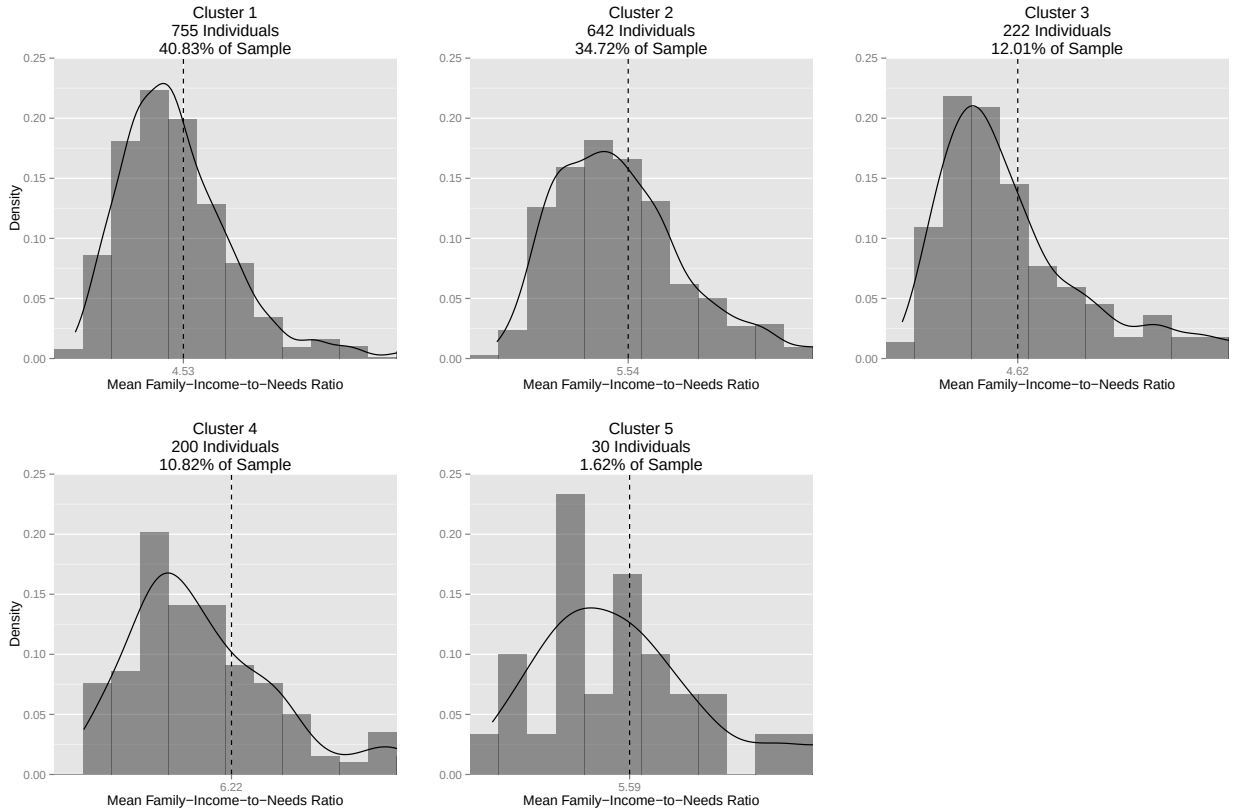


Figure L.4: Mean Family-Income-to-Needs Ratio by Cluster ($k = 5$).

APPENDIX M

PAIRWISE COMPARISONS OF CLUSTER DEMOGRAPHICS

The following tables present pairwise comparisons of cluster demographics using pooled standard deviations and using the method of correction for Type I errors in Holm (1979). The tables show that many of the observed differences between cluster-level means are statistically significant when setting α to 0.05. To aid comparison, I have highlighted all statistically significant comparisons.

Table M.1: Birth Year Comparisons.

Clusters	1	2	3	4	5	6	7	8
2	**0.000***							
3	1	**0.000***						
4	**0.000***	**0.000***	**0.000***					
5	1	**0.003***	1	**0.000***				
6	1	0.275	1	**0.000***	1			
7	0.404	1	0.104	0.156	0.140	0.720		
8	0.011	0.493	**0.003***	1	**0.004***	0.030	0.850	
9	0.582	0.124	0.809	**0.004***	0.850	0.740	0.140	0.012

Table M.2: Income at Age 25 Comparisons.

Clusters	1	2	3	4	5	6	7	8
2	**0.000***							
3	**0.000***	**0.000***						
4	**0.000***	**0.000***	**0.000***					
5	**0.000***	**0.000***	1	**0.000***				
6	**0.000***	**0.000***	**0.000***	**0.000***	**0.000***			
7	**0.000***	**0.000***	**0.000***	**0.000***	**0.000***	**0.000***		
8	**0.000***	**0.000***	**0.000***	**0.000***	**0.000***	**0.000***	**0.000***	
9	1	0.205	1	**0.001***	1	**0.000***	**0.000***	**0.000***

Table M.3: Levels of Education Comparisons.

Clusters	1	2	3	4	5	6	7	8
2	**0.001***							
3	0.055	**0.000***						
4	**0.004***	1	**0.000***					
5	1	**0.042***	1	**0.030***				
6	1	0.879	1	0.469	1			
7	1	1	**0.018***	1	0.983	1		
8	1	1	1	1	1	1	1	
9	1	1	1	1	1	1	1	1

Table M.4: Mean Family-Income-to-Needs Ratio Comparisons.

Clusters	1	2	3	4	5	6	7	8
2	**0.000***							
3	1	**0.000***						
4	**0.000***	1	**0.000***					
5	1	1	0.433	0.419				
6	1	0.346	1	0.140	1			
7	**0.002***	1	**0.000***	1	0.311	0.102		
8	1	1	1	1	1	1	1	
9	1	1	1	1	1	1	1	1

APPENDIX N

INCOME VARIANCE BY CLUSTER BY AGE

The following figures display the age-specific variance in the log change of income since age 25. I present this as an illustration of the goodness of fit of the clustering assignment.

The solid lines trace out the cluster specific variances while the dashed line traces out the variance across all individuals. The variance at age 25 is zero since all income is measured as a log deviation from income at age 25.

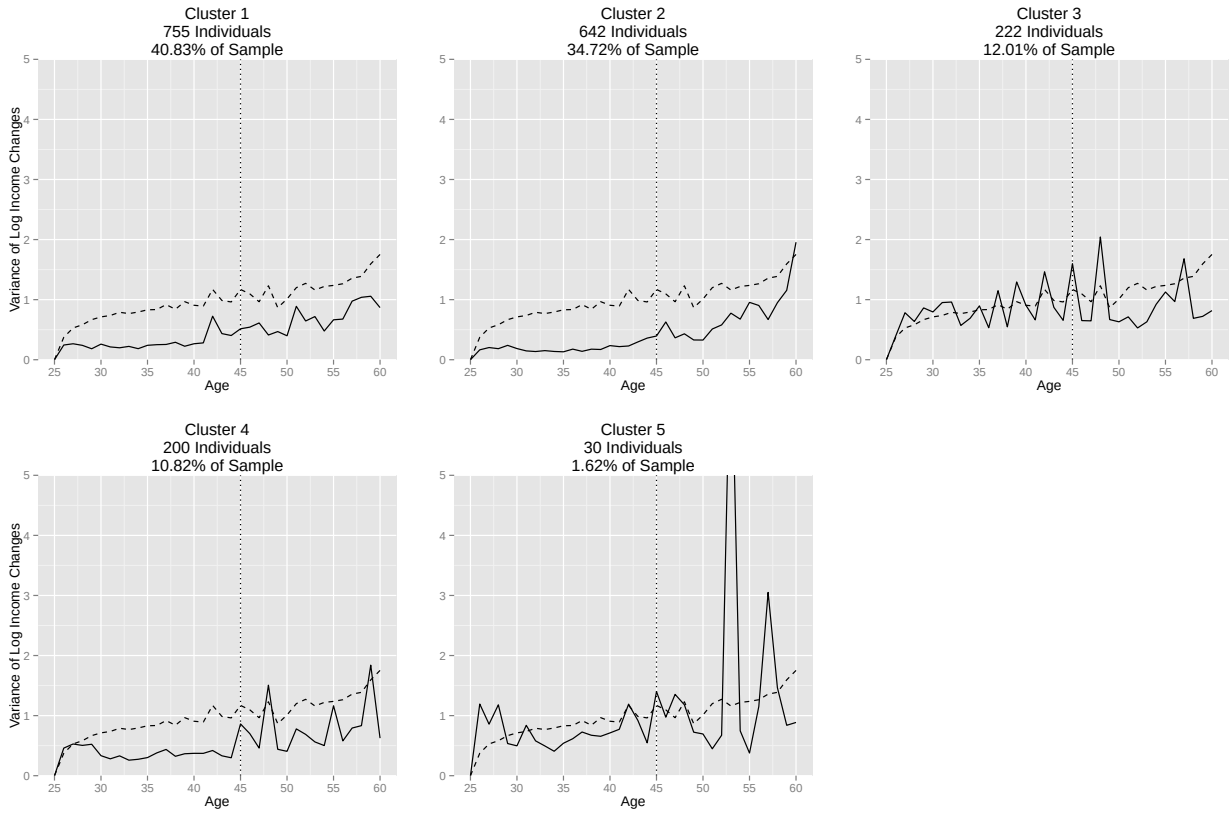


Figure N.1: Variance of log income change by cluster ($k = 5$). The solid line traces the variance of each cluster at each age. The dashed line traces the variance over all clusters at each age. The vertical dotted line indicates the last year that is used in the classification of clusters.

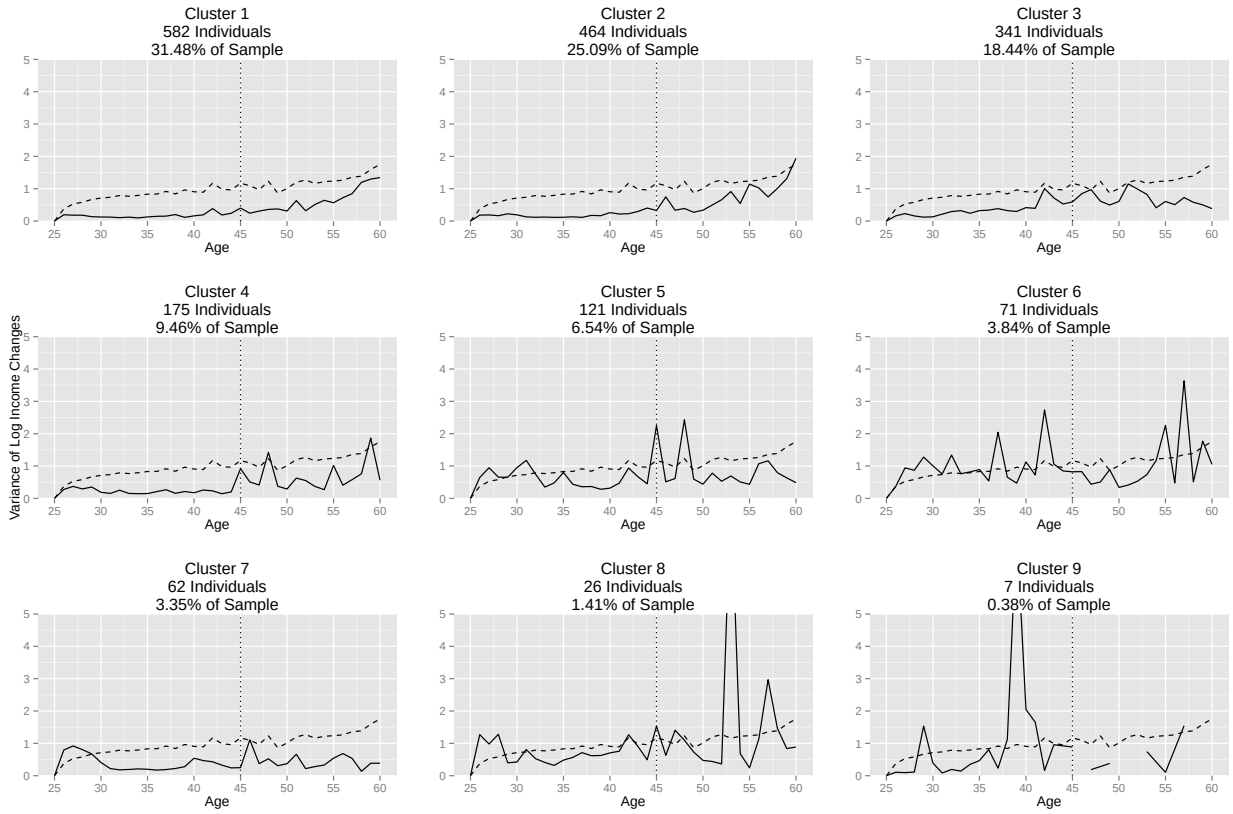


Figure N.2: Variance of log income change by cluster ($k = 9$). The solid line traces the variance of each cluster at each age. The dashed line traces the variance over all clusters at each age. The vertical dotted line indicates the last year that is used in the classification of clusters.

BIBLIOGRAPHY

Acemoglu, D., Reed, T., and Robinson, J. A. (2014). Chiefs: Economic Development and Elite Control of Civil Society in Sierra Leone. *Journal of Political Economy*, 122(2):pp. 319–368.

Acemoglu, D. and Robinson, J. A. (2008). Persistence of Power, Elites, and Institutions. *American Economic Review*, 98(1):267–293.

Africa, S. S. (1996). The People of South Africa, Population census, 1996. Technical report.

Aguiar, M. and Hurst, E. (2013). Deconstructing Life Cycle Expenditure. *Journal of Political Economy*, 121(3):pp. 437–492.

Akita, T. (2003). Decomposing regional income inequality in China and Indonesia using two-stage nested Theil decomposition method. *The Annals of Regional Science*, 37(1):55–77.

Alderman, H., Babita, M., Demombynes, G., Makhatha, N., and Özler, B. (2002). How Low Can You Go? Combining Census and Survey Data for Mapping Poverty in South Africa. *Journal of African Economies*, 11(2):169–200.

Alea, N. and Bluck, S. (2007). I'll keep you in mind: the intimacy function of autobiographical memory. *Applied Cognitive Psychology*, 21(8):1091–1111.

An-Chi Tung (2011). Consumption over the lifecycle: an international comparison. In Lee, R. and Mason, A., editors, *Population Aging and the Generational Economy: A Global Perspective*, pages 136–160. Edward Elgar Publishing, Cheltenham, UK.

Angrist, J., Lavy, V., and Schlosser, A. (2010). Multiple Experiments for the Causal Link between the Quantity and Quality of Children. *Journal of Labor Economics*, 28(4):773–824.

Antweiler, W. (2001). Nested random effects estimation in unbalanced panel data. *Journal of Econometrics*, 101(2):295–313.

Asadullah, M. N. (2011). Intra- and inter-household externalities in children's schooling: evidence from rural residential neighbourhoods in Bangladesh. *Applied Economics*, 43(14):1749–1767.

Asadullah, M. N. (2012). Intergenerational Wealth Mobility in Rural Bangladesh. *Journal of Development Studies*, 48(9):1193–1208.

Attanasio, O. P., Banks, J., Meghir, C., and Weber, G. (1999). Humps and Bumps in Lifetime Consumption. *Journal of Business & Economic Statistics*, 17(1):22–35.

Barro, R. J. and Lee, J. W. (2013). A new data set of educational attainment in the world, 1950–2010. *Journal of Development Economics*, 104:184 – 198.

Björklund, A. and Jäntti, M. (2009). Intergenerational income mobility and the role of family background. *Oxford Handbook of Economic Inequality, Oxford University Press, Oxford*.

Björklund, A. and Jäntti, M. (2012). How important is family background for labor-economic outcomes? *Labour Economics*, 19(4):465–474.

Björklund, A., Lindahl, L., and Lindquist, M. J. (2010). What more than parental income, education and occupation? an exploration of what swedish siblings get from their parents. *The BE Journal of Economic Analysis & Policy*, 10(1):–.

Black, S. E., Devereux, P. J., and Salvanes, K. G. (2005). The More the Merrier? The Effect of Family Size and Birth Order on Children's Education. *The Quarterly Journal of Economics*, 120(2):669–700.

Browne, W. J. and Draper, D. (2006). A comparison of Bayesian and likelihood-based methods for fitting multilevel models. *Bayesian Analysis*, 1(3):473–514.

Burbidge, J. and Robb, A. (1985). Evidence on Wealth-Age Profiles in Canadian Cross-Section Data. *The Canadian Journal of Economics / Revue canadienne d'Economique*, 18(4):pp. 854–875.

Cagetti, M. (2003). Wealth Accumulation Over the Life Cycle and Precautionary Savings. *Journal of Business & Economic Statistics*, 21(3):339–353.

Carter, M. and May, J. (2001). One kind of freedom: Poverty dynamics in post-apartheid South Africa. *World Development*, 29(12):1987–2006.

Carter, M. R. and May, J. (1999). Poverty, livelihood and class in rural South Africa. *World Development*, 27(1):1–20.

Case, A. (1992). Neighborhood influence and technological change. *Regional Science and Urban Economics*, 22(3):491–508.

Conley, T. G. and Udry, C. R. (2010). Learning about a New Technology: Pineapple in Ghana. *American Economic Review*, 100(1):35–69.

Corbeil, R. R. and Searle, S. R. (1976). Restricted Maximum Likelihood (REML) Estimation of Variance Components in the Mixed Model. *Technometrics*, 18(1):31–38.

Datcher, L. (1982). Effects of Community and Family Background on Achievement. *The Review of Economics and Statistics*, 64(1):32–41.

De Luca, G. and Zuccolotto, P. (2011). A tail dependence-based dissimilarity measure for financial time series clustering. *Advances in Data Analysis and Classification*, 5(4):323–340.

Deaton, A. (1992). *Understanding Consumption*. Clarendon Press.

Deaton, A. (1997). *The analysis of household surveys: a microeconomic approach to development policy*. Johns Hopkins University Press.

D'Urso, P. and Maharaj, E. A. (2009). Autocorrelation-based fuzzy clustering of time series. *Fuzzy Sets and Systems*, 160(24):3565 – 3589. Theme: Non-Linear Systems and Fuzzy Clustering.

Elbers, C., Lanjouw, J. O., and Lanjouw, P. (2003). Micro-Level Estimation of Poverty and Inequality. *Econometrica*, 71(1):355–364.

Elbers, C., Lanjouw, P., Mistiaen, J., Özler, B., and Simler, K. (2005). Are neighbours equal? Estimating local inequality in three developing countries. *Spatial inequality and development*, pages 37–60.

Elbers, C., Lanjouw, P., Mistiaen, J. A., and Özler, B. (2008). Reinterpreting between-group inequality. *The Journal of Economic Inequality*, 6(3):231–245.

Emran, M. S., Greene, W. H., and Shilpi, F. (2015). When measure matters: coresident sample selection bias in estimating intergenerational mobility in developing countries.

Emran, M. S. and Shilpi, F. (2015). Gender, Geography, and Generations: Intergenerational Educational Mobility in Post-Reform India. *World Development*, 72:362–380.

Epprecht, M., Müller, D., and Minot, N. (2009). How remote are Vietnam's ethnic minorities? An analysis of spatial patterns of poverty and inequality. *The Annals of Regional Science*, 46(2):349–368.

Fang, Y. and Wang, J. (2012). Selection of the number of clusters via the bootstrap method. *Computational Statistics & Data Analysis*, 56(3):468–477.

Fernández-Villaverde, J. and Krueger, D. (2007). Consumption over the Life Cycle: Facts from Consumer Expenditure Survey Data. *Review of Economics and Statistics*, 89(3):552–565.

Flachaire, E. and Nuñez, O. (2007). Estimation of the income distribution and detection of subpopulations: An explanatory model. *Computational Statistics & Data Analysis*, 51(7):3368–3380.

Friedman, M. (1957). The Permanent Income Hypothesis. In *A Theory of the Consumption Function*, pages 20–37. Princeton University Press, Princeton, NJ.

Fu, T.-c. (2011). A review on time series data mining. *Engineering Applications of Artificial Intelligence*, 24(1):164 – 181.

Gourinchas, P.-O. and Parker, J. A. (2002). Consumption Over the Life Cycle. *Econometrica*, 70(1):47–89.

Gräb, J. and Grimm, M. (2009). Spatial inequalities explained: evidence from Burkina Faso. ISS Working Papers - General Series 1765018725, International Institute of Social Studies of Erasmus University (ISS), The Hague.

Grieger, L. D., Schoeni, R. F., and Danziger, S. (2009). Accurately Measuring the Trend in Poverty In the United States Using The Panel Study of Income Dynamics. *Journal of Economic and Social Measurement*, 34(2):105–117.

Gustafsson, B. and Shi, L. (2002). Income inequality within and across counties in rural China 1988 and 1995. *Journal of Development Economics*, 69(1):179–204.

Guvenen, F. (2007). Learning Your Earning: Are Labor Income Shocks Really Very Persistent? *The American Economic Review*, 97(3):pp. 687–712.

Hao, L. (2004). Wealth of Immigrant and Native-Born Americans. *International Migration Review*, 38(2):518–546.

Hartigan, J. A. and Wong, M. A. (1979). Algorithm AS 136: A K-Means Clustering Algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):100–108.

Hartley, H. O. and Rao, J. N. K. (1967). Maximum-Likelihood Estimation for the Mixed Analysis of Variance Model. *Biometrika*, 54(1/2):pp.93–108.

Haslam, C., Jetten, J., Haslam, S. A., Pugliese, C., and Tonks, J. (2011). 'I remember therefore I am, and I am therefore I remember': Exploring the contributions of episodic and semantic self-knowledge to strength of identity. *British Journal of Psychology*, 102(2):184–203.

Holm, S. (1979). A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics*, 6(2):65–70.

Jolliffe, D. (2002). Whose Education Matters in the Determination of Household Income? Evidence from a Developing Country. *Economic Development and Cultural Change*, 50(2):287–312.

Klasen, S. (2000). Measuring Poverty and Deprivation in South Africa. *Review of Income and Wealth*, 46(1):33–58.

Kravdal, O. (2002). Education and fertility in sub-Saharan Africa: Individual and community effects. *Demography*, 39(2):233–250.

Kuznets, S. (1955). Economic Growth and Income Inequality. *The American Economic Review*, 45(1):1–28.

Lange, T., Roth, V., Braun, M. L., and Buhmann, J. M. (2004). Stability-Based Validation of Clustering Solutions. *Neural Computation*, 16(6):1299–1323.

Lee, R., Donehower, G., and Miller, T. (2011). The changing shape of the economic lifecycle in the United States, 1960 to 2003. In Lee, R. and Mason, A., editors, *Population Aging and the Generational Economy: A Global Perspective*, pages 313–326. Edward Elgar Publishing, Cheltenham, UK.

Lee, R., Lee, S.-H., and Mason, A. (2008). Charting the Economic Life Cycle. *Population and Development Review*, 34:208–237.

Lee, R. and Mason, A. (2011a). Lifecycles, support systems, and generational flows: patterns and change. In *Population Aging and the Generational Economy: A Global Perspective*, pages 79–106. Edward Elgar Publishing, Cheltenham, UK.

Lee, R. and Mason, A. (2011b). *Population Aging and the Generational Economy: A Global Perspective*. Edward Elgar Publishing, Cheltenham, UK.

Lee, S.-H. and Ogawa, N. (2011). Labor income over the lifecycle. In Lee, R. and Mason, A., editors, *Population Aging and the Generational Economy*, pages 109–135. Edward Elgar Publishing, Cheltenham, UK.

Leibbrandt, M. and Woolard, I. (2001). The labour market and household income inequality in South Africa: existing evidence and new panel data. *Journal of International Development*, 13(6):671–689.

Liao, T. W. (2005). Clustering of time series data - a survey. *Pattern Recognition*, 38(11):1857 – 1874.

Lindahl, L. (2011). A comparison of family and neighborhood effects on grades, test scores, educational attainment and income: evidence from Sweden. *The Journal of Economic Inequality*, 9(2):207–226.

Maralani, V. (2008). The changing relationship between family size and educational attainment over the course of socioeconomic development: Evidence from Indonesia. *Demography*, 45(3):693–717.

Masson, A. (1986). A Cohort Analysis of Wealth-Age Profiles Generated by a Simulation Model in France (1949-75). *The Economic Journal*, 96(381):pp. 173–190.

May, J., Carter, M. R., Haddad, L., and Maluccio, J. A. (2000). KwaZulu-Natal Income Dynamics Study (KIDS) 1993–98: A longitudinal household database for South African policy analysis. *Development Southern Africa*, 17(4):567–581.

May, J. D., Agüero, J., Carter, M. R., and Tim?us, I. M. (2007). The KwaZulu-Natal Income Dynamics Study (KIDS) third wave: methods, first findings and an agenda for future research. *Development Southern Africa*, 24(5):629–648.

Mazumder, B. (2008). Sibling similarities and economic inequality in the US. *Journal of Population Economics*, 21(3):685–701.

Minot, N., Baulch, B., and Epprecht, M. (2006). *Poverty and inequality in Vietnam: Spatial patterns and geographic determinants*. International Food Policy Research Institute Washington, DC.

Modigliani, F. (1986). Life Cycle, Individual Thrift, and the Wealth of Nations. *The American Economic Review*, 76(3):297–313.

Modigliani, F. and Brumberg, R. (1954). Utility Analysis and the Consumption Function: an Interpretation of Cross-Section Data. In Kurihara, K. K., editor, *Post-Keynesian Economics*. Rutgers University Press, New Brunswick, NJ.

MOEC (2013). Overview of the Education Sector in Indonesia 2012: Acheivements and Challenges.

Montmarquette, C. and Mahseredjian, S. (1989). Does School Matter for Educational Achievement? A Two-Way Nested-Error Components Analysis. *Journal of Applied Econometrics*, 4(2):pp.181–193.

Naschold, F. and Barrett, C. B. (2011). Do Short-Term Observed Income Changes Overstate Structural Economic Mobility?. *Oxford Bulletin of Economics & Statistics*, 73(5):705–717.

Noble, M., Barnes, H., Wright, G., McLennan, D., Avenell, D., Whitworth, A., and Roberts, B. (2009). The South African Index of multiple deprivation 2001 at datazone level. Technical report, Pretoria: Department of Social Development.

Noble, M. and Wright, G. (2012). Using Indicators of Multiple Deprivation to Demonstrate the Spatial Legacy of Apartheid in South Africa. *Social Indicators Research*, ONLINE FIRST:1–15.

Novotný, J. (2007). On the measurement of regional inequality: does spatial dimension of income inequality matter? *The Annals of Regional Science*, 41(3):563–580.

Oreopoulos, P. (2003). The Long-Run Consequences of Living in a Poor Neighborhood. *The Quarterly Journal of Economics*, 118(4):1533–1575.

Özler, B. (2007). Not Separate, Not Equal: Poverty and Inequality in Post-apartheid South Africa. *Economic Development and Cultural Change*, 55(3):pp.487–529.

Page, M. E. and Solon, G. (2003). Correlations between Brothers and Neighboring Boys in Their Adult Earnings: The Importance of Being Urban. *Journal of Labor Economics*, 21(4):831–855.

Park, C. (2006). Risk Pooling between Households and Risk-Coping Measures in Developing Countries: Evidence from Rural Bangladesh. *Economic Development and Cultural Change*, 54(2):423–457.

Platzky, L. and Walker, C. (1985). *The surplus people: Forced removals in South Africa.* Johannesburg South Africa Ravan Press 1985.

Raaum, O., Salvanes, K. G., and Sørensen, E. Ø. (2006). The Neighbourhood is Not What it Used to be*. *The Economic Journal*, 116(508):200–222.

Rao, N. and Murthy, N. (2007). *Applied Statistics in Health Sciences.* Jaypee Brothers Publishers.

Robinson, D. L. (1987). Estimation and Use of Variance Components. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 36(1):pp.3–14.

Rosenzweig, M. R. (2003). Payoffs from Panels in Low-Income Countries: Economic Development and Economic Mobility. *The American Economic Review*, 93(2):pp.112–117.

Shamsuddin, A. F. M. and DeVoretz, D. J. (1998). Wealth Accumulation of Canadian and Foreign-Born Households in Canada. *Review of Income and Wealth*, 44(4):515–533.

Shorrocks, A. and Wan, G. (2005). Spatial decomposition of inequality. *Journal of Economic Geography*, 5(1):59–81.

Simler, K. R., Nhate, V., and others (2005). Poverty, inequality, and geographic targeting: Evidence from small-area estimates in Mozambique. *International Food Policy Research Institute, FCND Discussion Paper*, 192.

Snedecor, G. G. and Cochran, W. G. (1980). *Statistical Methods.* The Iowa State University Press, 7 edition.

Solon, G. (1999). Chapter 29 Intergenerational mobility in the labor market. volume 3, Part A of *Handbook of Labor Economics*, pages 1761 – 1800. Elsevier.

Solon, G., Corcoran, M., Gordon, R., and Laren, D. (1991). A Longitudinal Analysis of Sibling Correlations in Economic Status. *The Journal of Human Resources*, 26(3):509–534.

Solon, G., Page, M. E., and Duncan, G. J. (2000). Correlations between Neighboring Children in Their Subsequent Educational Attainment. *The Review of Economics and Statistics*, 82(3):383–392.

Takahashi, K. (2011). Determinants of Indonesian rural secondary school enrolment: gender, neighbourhood and school characteristics. *Bulletin of Indonesian Economic Studies*, 47(3):395–413.

Tarozzi, A. and Deaton, A. (2009). Using Census and Survey Data to Estimate Poverty and Inequality for Small Areas. *Review of Economics and Statistics*, 91(4):773–792.

Thurow, L. C. (1969). The Optimum Lifetime Distribution of Consumption Expenditures. *The American Economic Review*, 59(3):pp. 324–330.

Vonnegut, K. (2009). *Palm Sunday: An Autobiographical Collage*. Random House Publishing Group.

Wang, J. (2010). Consistent selection of the number of clusters via crossvalidation. *Biometrika*, 97(4):893–904.

Woolard, I. and Klasen, S. (2005). Determinants of Income Mobility and Household Poverty Dynamics in South Africa. *Journal of Development Studies*, 41(5):865–897.