Estimating selected disaggregated socio-economic indicators using small area estimation techniques

Inaugural-Dissertation zur Erlangung des akademischen Grades eines
Doktors der Wirtschaftswissenschaft des Fachbereichs
Wirtschaftswissenschaft der Freien Universität Berlin

vorgelegt von Noah Cheruiyot Mutai, Master of Science
aus Nandi County, Kenya

March 20, 2023

Dekan: Prof. Dr. Giacomo Corneo
Erstgutachter: Prof. Dr. Timo Schmid
Zweitgutachter: Prof. Dr. Ulrich Rendtel
Tag der Disputation: 19. Juli 2022

Acknowledgments

I am deeply indebted to Prof. Dr. Timo Schmid, for accepting to supervise my doctorate. I thank him for supporting my application for a scholarship with the German Academic Exchange Service (DAAD) and lastly for his invaluable guidance and understanding during my PhD studies.

I am also grateful to Prof. Dr. Ulrich Rendtel, for accepting to be my second reviewer and providing profound input for this thesis.

Special thanks to the German Academic Exchange Service (DAAD) and the National Research Fund (Kenya) for funding my studies. I thank Taita Taveta University for granting me a study leave to pursue doctoral studies.

Further, I am very thankful to my colleagues and friends at the Chair of Applied Statistics and the Statistical Consulting Unit *fu:stat*, for making my life pleasant in Berlin, and for providing useful suggestions and guidance throughout my PhD studies. Also my colleagues at Taita Taveta University for support and encouragement during this period

My family has sacrificed a lot to see me through my studies. I sincerely thank them for their understanding. Life is back to normal now that I am coming home. Last but not the least, I thank all my friends and all those people not mentioned above who contributed to my success in one way or another.

Publication List

The publications and working papers listed below are the result of the research carried out in this thesis titled: Estimating selected disaggregated socio-economic indicators using small area estimation techniques.

(a) Mutai, N. C. and Skarke, F (2023). Estimation of disaggregated poverty and inequality indicators with application to the Kenya Integrated Household Budget Survey, Working paper, to be submitted.

(b) Mutai, N. C. (2022). Small area estimation of health insurance coverage for Kenyan counties, AStA Wirtschafts- und Sozialstatistisches Archiv 16, 231-254.

(c) Mutai, N. C. (2022). Estimating county-level overweight prevalence in Kenya using small area methodology, South African Statistical Journal 56(1), 1-19.

Contents

Introduction

In 2015, the United Nations (UN) set up 17 Sustainable Development Goals (SDGs) to be achieved by 2030 (General Assembly, 2015). The goals encompass indicators of various socio-economic characteristics (General Assembly, 2015). To achieve them, there is need to reliably measure the indicators especially at disaggregated levels. National Statistical Institutes (NSI) collect data on various socio-economic indicators by conducting censuses or sample surveys. Although a census provides data on the entire population, it is only carried out every 10 years in most countries and it requires enormous financial resources. Sample surveys on the other hand are commonly used because they are cheaper and require a shorter time to collect (Särndal et al., 2003; Cochran, 2007). They are, therefore, essential sources of data on country's key socio-economic indicators, which are necessary for policy- making, allocating resources, and determining interventions necessary. Surveys are mostly designed for the national level and specific planned areas or domains. Therefore, the drawback is sample surveys are not adequate for data dis-aggregation due to small sample sizes (Rao and Molina, 2015). In this thesis, geographical divisions will be called areas, while other sub-divisions such as age-sex-ethnicity will be called domains in line with (Pfeffermann, 2013; Rao and Molina, 2015).

One solution to obtain reliable estimates at disaggregated levels, is to use small area estimation (SAE) techniques. SAE increases the precision of survey estimates by combining the survey data and another source of data, for example a previous census, administrative data or other passively recorded data such as mobile phone data as used in Schmid et al. (2017). The results obtained using the survey data only are called direct estimates, while those obtained using SAE models will be called model-based estimates. The auxiliary data are covariates related to the response variable of interest (Rao and Molina, 2015). According to Rao and Molina (2015), an area or domain is regarded as small if the area or domain sample size is not adequate to provide estimates of a desired accuracy. The field of SAE has grown substantially over the years mainly due to the demand from governments and private sectors. Currently, it is possible to

estimate several linear and non-linear target statistics such as the mean and the Gini coefficient (Gini, 1912), respectively. This thesis contributes to the wide literature of SAE by presenting three important applications using Kenyan data sources.

Chapter 1 is an application to estimate poverty and inequality in Kenya. The Empirical Best Predictor (EBP) of Molina and Rao (2010) and the M-quantile model of Chambers and Tzavidis (2006) are used to estimate poverty and inequality in Kenya. Four indicators are estimated, i.e. the mean, the Head Count Ratio, the Poverty Gap and the Gini coefficient. Three transformations are explored: the logarithmic, log-shift and the Box-Cox to mitigate the requirement for normality of model errors. The M-quantile model is used as a robust alternative to the EBP. The mean squared errors are estimated using bootstrap procedures. Chapter 2 is an application to estimate health insurance coverage in Kenyan counties using a binary M-quantile SAE model (Chambers et al., 2016) for women and men aged 15 to 49 years old. This has the advantage that we avoid specifying the distribution of the random effects and distributional robustness is automatically achieved. The MSE is estimated using an analytical approach based on Taylor series linearization. Chapter 3 presents the estimation of overweight prevalence at the county-level in Kenya. In this application, the Fay-Herriot model (Fay and Herriot, 1979) is explored with arcsine square-root transformation. This is to stabilize the variance and meet the assumption of normality. To transform back to the original scale, we use a bias-corrected back transformation. For this model, the design variance is smoothed using Generalized Variance Functions as in (Pratesi, 2016, Chapter 11). The mean squared error is estimated using a bootstrap procedure. In summary, this thesis contributes to the vast literature on small area estimation from an applied perspective by;

(a) Presenting for the first time regional disaggregated SAE results for selected indicators for Kenya.

(b) Combining data sources to improve the estimation of the selected disaggregated socio-economic indicators.

(c) Exploring data-driven transformations to mitigate the assumption of normality in linear and linear mixed-effects models.

(d) Presenting a robust approach to small area estimation based on the M-quantile model.

(e) Estimating the mean squared error to access uncertainty using bootstrap procedures.

Chapter 1

Estimation of disaggregated poverty and inequality indicators with application to the Kenya Integrated Household Budget Survey

1.1 Introduction

Poverty and inequality are among the world's persistent problems. Alleviating them remains an important issue in political and economic discussions. About 9.2% of the world lives in extreme poverty or on less than $1.90 a day, according to the World Bank. In 2020, the World Bank, estimated that between 88 million and 115 million additional people were pushed into extreme poverty, bringing the total to between 703 and 729 million living on less than $1.90 a day. The COVID-19 pandemic has reversed the gains in fighting global poverty for the first time in a generation (World Bank, 2020).

Sub-Saharan African (SSA) countries harbor a huge share of poverty and inequality in the world. In 2018, the World Bank's estimates show that SSA accounts for two-thirds of the global extreme poor population World Bank (2018). There has been slow progress in poverty reduction in SSA. Whereas the poverty rate decreased from 54% in 1990 to 41% in 2015, the number of poor continues to rise (Beegle and Christiaensen, 2019). This implies that the poverty rate in SSA has not fallen fast enough to keep up with population growth in the region and 433 million

Africans were estimated to live in extreme poverty in 2018, rising from 284 in 1990.

In Kenya, the proportion of people living on less than the international poverty line (US $1.90 per day) has declined from 46.8% in 2005/06 to 36.1% in 2015/16, according to the 17th edition of the Kenya economic update (World Bank, 2018). The Kenyan government through the Vision 2030, aims at transforming Kenya into an industrialized middle income country. This is by providing high quality of life to all citizens in a clean and secure environment. Through the social pillar of the Vision 2030, the government is committed to eradicate poverty through enhanced equity and wealth creation opportunities for the poor (Government of Kenya, 2007). The Sustainable Development Goals (SDGs) and the decentralized system of Government further reinforces these goals. To achieve these objectives, reliable statistics at disaggregated levels is required for planning, monitoring and evaluation and policy-making.

Poverty is a complex phenomenon. The first question is how to define it and the second, how it can be measured. The UN defines poverty as a denial of choices and opportunities, a violation of human dignity. A lack of basic capacity to participate effectively in society. It implies not having enough to feed and clothe a family, not having a school or a clinic to go to, not having the land on which to grow one's food or a job to earn one's living, not having access to credit. It means insecurity, powerlessness and exclusion of individuals, households and communities. It involves susceptibility to violence, and it often implies living on marginal and fragile environments, without access to clean water and sanitation. Atkinson (1987) notes two challenges in measuring poverty — how to summarize a multidimensional problem in a one-dimensional indicator and how to distinguish between the poor and the non-poor.

Measuring poverty and inequality is an important step towards eradication. Common measures of poverty are based on household per capita income, expenditure or consumption level (Greeley, 1994). Although poverty is a multidimensional concept, we concentrate on one-dimensional poverty measurement. This entails poverty based on expenditure in absolute and relative terms. Absolute poverty depends on a certain monetary value set for the whole world, while relative poverty depends on the country where people live. The latter is measured by use of a poverty line or a threshold obtained from an adequate minimum income in a given country by national governments (Betti and Lemmi, 2013; Atkinson, 1987).

Measuring poverty and income inequality indicators as one measure has been done in the literature. Notable are the Sen Index (Sen, 1976), the Monetary and Supplementary Fuzzy measures (Salvucci et al., 2012) and the Human Poverty Index (McGillivray and White, 1993).

A commonly used indicator is the Foster-Greer-Thorbecke (FGT) family. It gives information on incidence, intensity and severity (Boltvinik, 1999). The FGT was developed by Foster et al. (1984). The Laeken indicators is also another family by the European Council (Atkinson et al., 2002). They measure the Income Quintile Share Ratio (IQS) and the Gini coefficient.

The United Nation's SDG number one is no poverty (General Assembly, 2015). The UN member states pledged to leave no one behind. To identify the poorest and most unequal in the world requires sufficiently detailed data at lower levels. The General Assembly resolution 68/261 states that SDG indicators should be disaggregated, where relevant, by income, sex, age, race, ethnicity, migratory status, disability and geographic location, or other characteristics, in accordance with the fundamental principles of official statistics (Zhongming et al., 2021). National surveys, for example the Kenya Integrated Household Budget Survey (KIHBS), are normally used to obtain a number of indicators including ones for poverty and inequality. Using the survey data only, direct estimators such as the Horvitz-Thompson (HT) estimator (Horvitz and Thompson, 1952) for means or Totals can be computed. Practically, these surveys are not sufficient to provide detailed information at lower geographical levels due to zero or small sample sizes and the estimates are highly variable. This results in the need for SAE methods.

SAE involves the development of statistical methods and procedures for producing more reliable estimates for so called small areas or domains, i.e. with zero or small sample sizes. A domain or area is regarded as small if the domain-specific sample is not large enough to support direct estimates of adequate precision (Rao and Molina, 2015; Schaible, 2013). Domains may refer to age-sex-race group in a large geographical area (Rao and Molina, 2015). Unlike direct estimators (which rely on domain-specific data), more sophisticated SAE methods involve indirect estimation, by borrowing strength from related areas or longitudinal data. An important requirement for SAE models is availability of good auxiliary variables related to a target variable. These covariates have to be available in the survey data as well as the additional data used to estimate model-based indicators. The most common additional data sources are census data, administrative data or alternative sources such as mobile phones (Schmid et al., 2017; Hadam et al., 2020), big data sources (Marchetti et al., 2015) and social media data (Marchetti et al., 2016).

SAE models can be broadly divided into two categories: Unit-level models and area-level models. Unit-level models are used when there is access to unit values of a target variable to unit-specific explanatory variables. The best known model is the Battese-Harter-Fuller model

(Battese et al., 1988). On the other hand, area-level models are used if data is only available aggregated within areas. The most widely used area-level estimator is the Fay-Herriot estimator (Fay and Herriot, 1979). Due to data aggregation, there is usually some loss of information. The most commonly used SAE methods for estimation of especially non-linear poverty and inequality indicators are: the Empirical Best Predictor (EBP) approach (Molina and Rao, 2010), the M-Quantile approach (MQ) (Chambers and Tzavidis, 2006) and the World Bank method (ELL) (Elbers et al., 2003). In this paper, we concentrate mainly on the EBP and MQ approaches. For more about SAE, the reader is referred to (Morales et al., 2021; Rao and Molina, 2015; Pfeffermann, 2013, 2002; Jiang and Lahiri, 2006; Ghosh and Rao, 1994).

This paper is organized as follows. We describe the data sources in section 2.2. In section 1.3, we outline the small area methodology applied in this paper. In particular, the EBP and MQ approaches. In section 2.4, we present the results of the application to estimate poverty and inequality in Kenya. Lastly, in section 2.5, we give the concluding remarks, possibilities for further research and limitation of this study.

1.2 Data sources

We use two data sets in this paper. The KIHBS 2015 and the Kenya Population and Housing Census (KPHC) 2009. These data sources are public use files available on `https://statistics.knbs.or.ke/nada/index.php/catalog` upon signing up.

1.2.1 Kenya Integrated Household Budget Survey data

The main goal of the KIHBS 2015 was to obtain integrated household level data on various well-being measures such as poverty and inequality, health, education, sanitation and labor force. This was to evaluate the progress in improving well-being on the national and county level. In terms of policy, the survey was conducted to inform and provide benchmark indicators to monitor the third Medium Term Plan (MTP III) and Kenya's progress towards achievement of the SDGs. Particularly, the survey was meant to provide data for computing updated poverty and inequality indicators (Kenya National Bureau of Statistics, 2018). It was financed by the World Bank through the Kenya Statistics Programme for Results project. For the survey, 24,000 households divided into urban and rural strata were sampled. The sample was drawn from the fifth National Sample Survey and Evaluation Programme (NASSEP V) household sampling

frame, which is the frame that the KNBS currently uses to conduct household-based surveys in Kenya. It consists of 5,360 clusters split into four equal sub-samples. The frame is stratified into urban and rural areas within each of the 47 counties resulting in 92 sampling strata with Nairobi and Mombasa counties being wholly urban (Kenya National Bureau of Statistics, 2018). The creation of the sample was a three stage procedure: (i) A total of 2,400 clusters (988 in urban and 1,412 in rural areas) were sampled from NASSEP V sampling frame. (ii) Selection of 16 households from each of the clusters. (iii) The sub-sampling of 10 households (from the 16 households) for the main KIHBS.

A total of seven questionnaires was used in the survey: (i) three main questionnaires, (ii) two diaries, (iii) one market questionnaire, and (iv) one community questionnaire. The three main questionnaires were administered at the household level, while the market and community questionnaires were administered at the cluster level (Kenya National Bureau of Statistics, 2018). Out of 23,852 households that were sampled for the survey, a total of 21,773 households were successfully interviewed. The response rate was 93.6% and 88.0% for rural and urban households respectively. The main reasons for non-response are that 13 clusters are unaccounted for because of insecurity (bandit raids) and households that were inaccessible due to migration in nomadic communities. To ensure the survey is representative, survey weights are provided using inclusion probabilities. The design weights were adjusted using the survey response to provide the final weights.

According to Kenya National Bureau of Statistics (2018), the sampling weights were calculated as the inverse of the inclusion probabilities. The probability p of selecting a household into the KIHBS 2015 is a product of four factors, i.e. $p = \prod_{i=1}^{4} p_i$ where p_1 is the probability of selecting the enumeration areas (EAs) for the NASSEP V master sample among all the EAs in the 2009 census, p_2 is the probability of selecting the EA segment to form a cluster among all segments in the EA, p_3 is the probability of selecting the cluster for the KIHBS, among all the clusters in the NASSEP V master sample; and p_4 is the probability of selecting the household among all the households listed in the cluster. To take into consideration the non-proportional distribution of clusters and non-response, the cluster weights are obtained by the product of sample cluster design weight, household and cluster response adjustment, i.e. $w_{kl} = D_{kl}\frac{S_{kl}C_l}{I_{kl}c_l}$ where w_{kl} is the overall final cluster weight for cluster k in stratum l, D_{kl} is the sample cluster design weight obtained from inverse of cluster selection probabilities for cluster k in stratum l; S_{kl} is the number of listed households in cluster k in stratum l; I_{kl} is the number of responding

households in cluster k in stratum l; C_{kl} is the number of clusters in stratum l; and c_l is the number of clusters selected from stratum l.

Table 1.1: Summary of sample sizes in the KIHBS 2015 at the county level.

	Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum
Survey	326	435.5	464	458.4	485.5	545

Table 1.1 is a summary of sample sizes at the county level in Kenya. For this survey, all areas were sampled with a minimum sample size of 326 and a maximum of 545. Table 1.2 below is a summary of household consumption expenditure for the three enumeration area types in the survey as defined in the data section above. The variable equivalized household consumption expenditure is only available in the survey. From the median values, we observe a small difference between the rural areas and peri-urban in household consumption but a big difference between the two areas (rural and peri-urban) and urban areas. There is also a huge difference between the minimum and maximum value in both the rural and urban expenditure values.

Table 1.2: A summary of equivalized houshold consumption expenditure for rural, peri-urban and urban areas measured in the KIHBS 2015.

EA type	Min.	1st Quartile	Mean	Median	3rd Quartile	Max.
Rural	9.4	2,741.6	5,069.8	4,016.4	6,208.1	552,741.8
Peri-urban	172.0	3,196.0	5,907.0	4,702.0	7,144.0	59,736.0
Urban	452.7	5,610.2	10,351.4	8,460.4	12,757.1	458,332.5

According to the Kenya National Bureau of Statistics (2010), three strata were created for the place of residence, namely: Rural, Core-Urban and Peri-Urban. The definitions of these strata are contained in Kenya National Bureau of Statistics (2018). The Peri-Urban was merged with Core-Urban to create Urban stratum that has been used as a definition of urban areas by the KNBS. Therefore, for this paper, we use only two strata: Rural and Urban similar to Kenya National Bureau of Statistics (2018)

1.2.2 Census data

Model-based SAE relies on availability of good auxiliary data related to the outcome variable which is measured for all areas (Rao and Molina, 2015). For this study, we use the KPHC 2009. Census have been conducted in Kenya every decade since 1969, the latest being from

2019, which is not yet available for public use. The KNBS is the sole body mandated to collect, process and disseminate census and other statistical data. Huge investment goes into the census activity, since the data are needed to track the progress of numerous development goals and worldwide initiatives, such as the SDGs. The major objective of the 2009 census was to offer important information on the population's demographic, social, and economic features, and housing. These include population size, population composition, fertility rates, mortality and migration rates, levels of education, and size of labor force. A scanning technology was used. Technical support was provided by the United States Census Bureau (USCB) (Kenya National Bureau of Statistics, 2010). Prior to 2010, Kenya was administratively divided into provinces, districts, divisions, locations, sub-locations and villages. This was for the 2009 census. After 2010, new administrative areas were created, whereby the 46 districts were converted to counties. Thus, there are 47 counties plus Nairobi which was not a district, 290 sub-counties and 1450 wards (Government of Kenya, 2013). Therefore, as a result, we can connect the survey and census data on the same geographical level without problems. For this study, we estimate results for the county level, since national government decisions and funding are made for the counties.

Table 1.3: Summary of population sizes in Kenya Population and Housing Census 2009 at the county level in Kenya.

	Min.	1st Quartile	Mean	Median	3rd Quartile	Max.
Census	2,205	10,676	18,586	15,408	20,572	98,289

Table 2.6 is a summary of population sizes at the county level in Kenya. The census is the 10% sample, i.e. every 10th household of the whole data set is released by the KNBS (Kenya National Bureau of Statistics, 2010).

1.3 Small area methodology

In SAE, the used notation denotes by U a finite population of size N. This population is partitioned into D domains $U_1, U_2, \cdots, U_D$ of sizes $N_1, \cdots, N_D$. The subscripts $i = 1, \cdots, D$ refer to the ith domain. Within each domain $j = 1, \cdots, N_i$ refers to the jth unit of the population. A sample (s) of total size n is drawn, giving $n_1, \cdots, n_D$ as sample sizes for the domains. The size of the non-sampled part of the population (r) is then $N_i - n_i$. In-sample and out-of-sample parts of the population within areas are denoted by s_i and r_i, respectively. It is possible, that domains have no observations in the sample. These domains are called unobserved domains.

The variable of interest for estimating poverty or inequality measures is denoted by y_{ij}. If a model-based approach is used, the vector of predictor variables is indicated by x_{ij}. In this paper, four indicators will be investigated as mentioned, using and comparing three different approaches of estimation. The basis for comparisons most often is a direct estimator, which uses the sample information only to derive estimates for the indicators of interest. Since sampling weights are available with this particular survey data, weighted versions of the indicators will be presented. In the application part of this paper, the following indicators will be reported for the direct estimator:

Mean:

$$\hat{\mu_i} = \frac{\sum_{j=1}^{n_i} w_{ij} y_{ij}}{\sum_{j=1}^{n_i} w_{ij}},$$

where w_{ij} denote sampling weights for every observational unit. In cases of simple random sampling $w_{ij} = 1, \forall i = 1, \cdots D, j = 1, \cdots, n_i$. y_{ij} is the variable used to estimate the indicator (e.g. expenditure). Additionally to this linear indicator three non-linear indicators are estimated. The first two belong to the family of the FGT indices (Foster et al., 1984). Both depend on a poverty threshold, dividing income or expenditure in above and below, where observational units below the threshold are being counted towards the estimation of the indicator. A person of household below the threshold is considered poor. The first FGT measure is called Head Count Ratio (HCR) and gives the fraction of households living below the chosen poverty line, whereas the second measure Poverty Gap (PG) is defined as the average amount of income or expenditure the poor are away from the poverty line, seen as a proportion of the threshold.

FGT:

$$\widehat{HCR}_i = \frac{1}{\sum_{j=1}^{n_i} w_{ij}} \sum_{j=1}^{n_i} w_{ij} I(y_{ij} \leq t)$$

$$\widehat{PG}_i = \frac{1}{\sum_{j=1}^{n_i} w_{ij}} \sum_{j=1}^{n_i} w_{ij} (\frac{t - y_{ij}}{t}) I(y_{ij} \leq t),$$

where w_{ij} again are the sampling weights, y_{ij} the monetary variable and $I(\cdot)$ is an indicator function, which is 1 if y_{ij} is below the poverty line t and 0 else.

Furthermore, the interest lies in estimating an indicator, which in contrast to the other mentioned measures is not a poverty indicator per se, but measures inequality instead. The

indicator is called Gini coefficient (Gini, 1912) and ranges between 0 and 1. The higher the value, the more inequality there is in the areas with regards to the chosen variable y_{ij} (e.g. expenditure) and vice versa.

Gini:

$$\widehat{Gini}_i = \left[\frac{2 \sum_{j=1}^{n_i} (w_{ij} y_{ij} \sum_{l=1}^{n_i} w_{il}) - \sum_{j=1}^{n_i} w_{ij}^2 y_{ij}}{\sum_{j=1}^{n_i} w_{ij} \sum_{j=1}^{n_i} w_{ij} y_{ij}} - 1 \right],$$

where all variables are defined as with the other indicators. In contrast to the FGT measures, no poverty threshold is needed to estimate the Gini coefficient.

The R-package `emdi` (Kreutzmann et al., 2019a) is used for the direct estimators. All indicators of interest can be estimated using the direct-command.

1.3.1 The Empirical Best Predictor

In contrast to the direct estimator, the model-based Empirical Best Predictor (EBP) makes not only use of the survey data, but combines it with additional census or other register data in order to improve upon direct estimation by borrowing strength across areas. This approach was introduced by Molina and Rao (2010) and is implemented in the R-package `emdi` (Kreutzmann et al., 2019a). The idea behind this approach is to use a unit-level mixed model (Battese et al., 1988), which allows the use of covariates. The design matrix $X = (x_0, \cdots, x_p)^T$ is comprised of p auxiliary variables. The model is a mixed model, since it does not only contain a unit-level error term, but also a random effect, specific to the areas in the sample. The model is defined as follows:

$$\begin{aligned} T(y_{ij}) = x_{ij}^T \beta + u_i + \epsilon_{ij}, \text{ where } j = 1, \cdots, \ n_i, \ \ i = 1, \cdots, D, \\ u_i \sim N(0, \sigma_u^2), \ \ \epsilon_{ij} \sim N(0, \sigma_\epsilon^2)), \end{aligned} \tag{1.1}$$

where x_{ij}^T is a $(p+1) \times 1$ vector of explanatory variables, β is a $(p+1) \times 1$ vector of regression coefficients and u_i and ϵ_{ij} are random area-level and unit-level error terms, respectively. The dependent variable y_{ij} is only available in the survey data, but not in the second data source. The auxiliary variables on the other hand have to be available in both data sets. $T(\cdot)$ is a possible transformation of the dependent variable. The reasoning behind transformations is the assumption of normality for both error terms in equation 1.1. In practice, variables like income

and expenditure are very rarely symmetrically distributed, which can lead to non-normal errors for the mixed model. In the past, mostly deterministic transformations have been used to achieve normality. The best known deterministic transformation is the Log transformation:

$$T(y_{ij}) = log(y_{ij} + s),$$

where s is a shift parameter making the sum of y_{ij} and s strictly positive. Only then, is it possible to apply the logarithmic function. The results of using deterministic transformations like the Log can be improved upon, using so called data-driven transformations. The transformations described in Rojas-Perilla et al. (2020) have transformation parameters, which are estimated using the survey data in the model fitting step of the EBP. In this paper, the Log-shift (Feng et al., 2016) and Box-Box transformation (Box and Cox, 1964) will be compared to a deterministic Log-transformation and a model without transformation, to see which version works best with regards to the normality assumptions of the mixed model. The Log-shift transformation is defined as:

$$T(y_{ij}) = log(y_{ij} + \lambda),$$

where λ is an estimated optimal shift. The Box-Cox transformation is defined as:

$$T(y_{ij}) = \begin{cases} \frac{(y_{ij}+s)^{\lambda}-1}{\lambda} & \text{if } \lambda \neq 0; \\ log(y_{ij} + s) & \text{if } \lambda = 0, \end{cases}$$

where again λ is a parameter estimated using the survey data and s is a deterministic shift, chosen as with the deterministic logarithmic transformation. Both special cases of no transformation ($\lambda = 1$, only a shift occurs) and the deterministic Log transformation ($\lambda = 0$) are enclosed in the Box-Cox transformation.

Following Rojas-Perilla et al. (2020), the point estimation of poverty and inequality indicators based on the EBP under transformation using a Monte Carlo approach works as follows:

1. Under a selected transformation obtain $T(y_{ij}) = y^*_{ij}$ for in-sample observations.
2. Using model 1.1 and y^*_{ij} estimate $\hat{\beta}$, $\hat{\sigma}^2_u$ and $\hat{\sigma}^2_\epsilon$ and calculate $\hat{\gamma}_i = \hat{\sigma}^2_u/(\hat{\sigma}^2_u + \hat{\sigma}^2_\epsilon/n_i)$.
3. For $l = 1, \cdots, L$:

3.1 Generate $\nu_i^{(l)} \sim N(0, \hat{\sigma}_u^2(1-\hat{\gamma}_i))$, as well as $\epsilon_{ij}^{(l)} \sim N(0, \hat{\sigma}_\epsilon^2)$ to obtain a pseudo-population of the dependent variable for the out-of-sample observations.

$$y_{ij}^{*(l)} = x_{ij}^T \hat{\beta} + \hat{u}_i + \nu_i^{(l)} + \epsilon_{ij}^{(l)}$$

3.2 Back-transform the values of pseudo-population to original scale using the inverse of the transformation function $y_{ij}^{(l)} = T^{-1}(y_{ij}^{*(l)})$.

3.3 Calculate the indicator of interest for each area $I_i^{(l)} = N_i^{-1}(\sum_{j \in s_i} I_{ij} + \sum_{j \in r_i} \hat{I}_{ij})$.

4. Take the average value over all L replications to obtain the final point estimate for the chosen indicator

$$\hat{I}_i^{EBP} = \frac{1}{L} \sum_{l=1}^{L} I_i^{(l)}$$

The parameter λ, which is used to transform the sample target variable, is estimated by going over an interval of potential values, while maximizing the Restricted Maximum Likelihood (REML). Since linkage between sample and census is rarely possible in reality, target variable values are predicted for all observations in the census, not just the out-of-sample observations. This version of the EBP is often called census EBP (Guadarrama et al., 2016).

The uncertainty of these point estimates are quantified by using a parametric bootstrap approach following Molina and Rao (2010). When a data-driven transformation is used while estimating indicators with emdi, the additional uncertainty of estimating λ is incorporated in the bootstrap MSE (Rojas-Perilla et al., 2020).

Additionally to this version of the EBP, that does not make use of the survey weights, there is a version of the EBP, which incorporates them (Guadarrama et al., 2018). The idea of using this EBP version instead of the unweighted one was discarded after testing if the survey design was informative with the Pfeffermann-Sverchkov test (Pfeffermann and Sverchkov, 1999). The results are shown in Table 1.4

Table 1.4: Pfeffermann-Sverchkov test of sample weight ignorability

	correlation values	p-values
Residual	0.000	0.929
Squared residual	-0.001	0.947
Cubed residual	-0.004	0.638

According to (Pfeffermann and Sverchkov, 1999), a significant correlation may indicate

biased estimates in the unweighted model. First, the test checks the correlation between the residuals of the model and the weights. It then estimates the variance of the correlation using bootstrapping. Finally, a t-test is used to check whether the correlation is different from zero. This is done for the squared residuals and cubed residuals as well. Since the p-values for all types of residuals are above 5% the test does not indicate, that the weights should be used and therefore the EBP without weights was chosen for this application.

1.3.2 The M-quantile approach

One way of dealing with departures from the normality assumptions for random errors was described in section 1.3.1. Often transformations of the dependent variable, especially data-driven transformations help to produce better predictive results for model-based SAE methods like the EBP. Since transformations cannot ensure normally distributed random errors and outlier values of the target variable can cause these types of violations, a different approach is using an outlier-robust model. In this paper, point estimates and MSE estimates for all considered indicators will also be calculated using the approach outlined in Marchetti et al. (2012) as an alternative to the EBP. The idea of combining two data sources in order to estimate the indicators remains the same. The sample data is used to estimate the model and the census data will be used to predict values for the target variable for out-of-sample observations. M-quantile regression models were introduced by Breckling and Chambers (1988), where they came up with a generalization of regression techniques based on influence functions. This generalization encompasses earlier contributions to regression methodology like quantile regression (Koenker and Bassett Jr, 1978) and expectile regression (Newey and Powell, 1987). Depending on the influence function used, those two mentioned regression methods are special cases of M-quantile regression. The M-quantile $MQ_q(x;\psi)$ of order q is defined as the solution of the estimating equation $\int \psi_q(y - MQ)f(y|x)\,dy$ for the density of y given a set of covariates x. Here ψ_q is an asymmetric influence function, which is the first derivative of the asymmetric loss function ρ_q. In the case of linear M-quantile regression, the conditional M-quantile is expressed as a linear combination of regression coefficients $MQ_q(x;\psi) = x^T\beta_\psi(q)$. Minimization of $\sum_{j=1}^{n}\rho_q(y_j - x_j^T\beta_\psi(q))$ acquires estimates for the set of coefficients. Taking the first derivative and setting it to zero gives estimating equations

$$\sum_{j=1}^{n}\psi_q(r_{jq})x_j = 0,$$

where the loss function is the Huber loss function, $r_{jq} = y_i - x_i^T\beta_\psi(q)$, $\psi_q(r_{jq}) = 2\psi(\frac{r_{jq}}{s})(qI(r_{jq} > 0) + (1-q)I(r_{jq} < 0))$. The scale parameter s is chosen as $median|r_{jq}|/(0.6745)$. Furthermore the influence function resulting from the Huber loss function is the so called Huber Proposal 2 influence function $\psi(v) = vI(-c \leq v \leq c) + c \cdot sgn(v)$, with tuning constant c bounded away from 0 (Huber, 1981). This type of regression model has been used first by Chambers and Tzavidis (2006) in SAE. Building on work by Kokic et al. (1997), they took the concept of M-quantile coefficients or q-scores to derive what can be seen as area-specific pseudo-random effects. The M-quantile coefficient θ_j for observational unit j is defined as the solution to $MQ_{\theta_j}(x;\psi) = y_j$. A naive estimator for the mean indicator for an area can then be expressed as

$$\hat{\mu}_i = \frac{1}{N_i}\left[\sum_{j\in s_i} y_{ij} + \sum_{j\in r_i} x_{ij}^T\hat{\beta}_\psi(\hat{\theta}_i)\right], \quad i = 1, \cdots, D,$$

where s_i are in-sample observations, r_i are out-of-sample observations from domain i and $\hat{\theta}_i$ is the average of all M-quantile coefficients of the observations within area i. Tzavidis et al. (2010) introduced a bias-corrected version of this estimator, since Chambers and Tzavidis (2006) noticed a prevalence for biased results under heteroscedastic or asymmetric error settings. This bias corrected version will be used in this paper for the estimation of the Mean indicator:

$$\hat{\mu}_i^{CD} = \frac{1}{N_i}\left[\sum_{j\in s_i} y_j + \sum_{j\in r_i} x_{ij}^T\hat{\beta}_\psi(\hat{\theta}_i) + (1-f_i)\sum_{j\in s_i}(y_{ij} - x_{ij}^T\hat{\beta}_\psi(\hat{\theta}_i))\right], \quad i = 1, \cdots, D,$$

where $f_i = n_i N_i^{-1}$ is the sampling fraction in domain i. This bias-adjusted estimator makes use of the Chambers-Dunstan estimator of the small area distribution function (Chambers and Dunstan, 1986). For the point estimation of non-linear indicators like HCR, PG and Gini, a Monte Carlo procedure is introduced by Marchetti et al. (2012) and Marchetti and Tzavidis (2021), which is somewhat similar to to EBP approach:

1. Using the sample data y_{ij}, x_{ij} estimate $\hat{\theta}_i$ and $\hat{\beta}_\psi(\hat{\theta}_i)$ under the M-quantile regression model.

2. For $l = 1, \cdots, L$:

   2.1 Generate an out-of-sample vector of size $N_i - n_i$

$$y_{ij}^{*(l)} = x_{ij}^T\hat{\beta}(\hat{\theta}_i) + \epsilon_{ij}^{*(l)}$$

for every area. The residuals $\epsilon_{ij}^{*(l)}$ are drawn from the empirical distribution of M-quantile model residuals. This can be done either within area or from all residuals.

2.2 Using the generated data and the sample data calculate indicator of interest for each area $I_i^{(l)}$ as in the case of the EBP.

3. Take the average value over all L replications to obtain the final point estimate for the chosen indicator

$$\hat{I}_i^{MQ} = \frac{1}{L}\sum_{l=1}^{L} I_i^{(l)}$$

Since linkage between sample and census units is often not possible in practice, the prediction of the target variable is done for all observations in the census. Additionally to the estimation of the point estimates, the authors also propose a non-parametric bootstrap approach to MSE estimation for HCR, PG (Marchetti et al., 2012) and Gini (Marchetti and Tzavidis, 2021), which is based on work of Lombardıa et al. (2003). The authors present four methods of generating B bootstrap populations. In a combination of either sampling from the empirical residuals or sampling from a smoothed distribution and sampling within area i (conditional approach) or sampling from all residuals (unconditional approach) one of these combinations has to be selected. The details about these are laid out in Tzavidis et al. (2010). In this paper, the smoothed unconditional approach is followed, since Marchetti et al. (2012) themselves used it in their paper and the conditional approach can be unreliable, when area sample sizes get small. From these bootstrap populations L bootstrap samples each are drawn without replacement, so that the number of observations in each area is the same as in the original sample ($n_i^* = n_i$). After estimating the desired indicator per area, these are then used to calculate bias and variance of said indicator over B and L.

1.4 Application: poverty and inequality mapping in Kenya

This section presents our results from estimating the four poverty and inequality indicators using Kenyan data sources. As mentioned in the theory part of this paper, three methods were used to estimate the Mean expenditure, the Head Count Ratio, the Poverty Gap and the Gini coefficient. Therefore, the figures always include results for the direct estimator using only the sample data and the results of the model-based approaches using the EBP and the M-quantile estimator.
The EBP model requires auxiliary data from a census or other administrative data sources. In this

application, the second data source is the KHPC 2009. The following auxiliary variables, which are available in both the survey and census data were selected, using the Bayesian Information Criterion (BIC) similar to Rojas-Perilla et al. (2020) under a random intercept model. These variables and their definitions are shown in Table 1.5.

After selecting the auxiliary covariates, the predictive power of the model was assessed using the marginal R^2 (R^2_m) and conditional R^2 (R^2_c) following (Nakagawa and Schielzeth, 2013). In this paper, the interest lies in data-driven transformations since the EBP estimator can be biased when error terms deviate significantly from the normal distribution (Rojas-Perilla et al., 2020). The aforementioned measures were compared for models without a transformation, the Log transformation, the Box-Cox transformation and the Log-shift transformation. The idea is to show the merits of using transformations and more so data-driven scaled transformation as apposed to transformations chosen without reliance on the data. The same variables were subsequently chosen for the M-quantile model as a comparison for the EBP and direct estimator.

Table 1.5: The names and definitions of auxiliary variables available in both survey and census data selected through BIC.

Variable	Definition
Television	indicator whether an household has a television or not (binary)
Computer	indicator whether an household has a computer or not (binary)
Cooking Source	the type of cooking energy used by an household (multinomial)
Floor material	the type of floor material of an household (multinomial)
Roof material	the type of roof material of an household (multinomial)
Wall material	the type of wall material of an household (multinomial)
Habitable rooms	the number of habitable rooms (integer)
Dwelling units	the number of dwelling units in an household (integer)

Table 1.6 shows the values of the R^2_m, R^2_c, ICC values and the transformation parameter λ under no transformation, Log transformation, Box-Cox transformation and Log-shift transformation. According to Rojas-Perilla et al. (2020), transformations are used to meet the normality assumption of residuals and random effects in the EBP. The predictive power of the model is expected to increase, when these assumptions are met. As seen in the Table 1.6, under no transformation, the R^2_m and R^2_c are 0.2051 and 0.2185, under Log transformation the R^2_m and R^2_c are 0.3641 and 0.4122, under Box-Cox transformation, the R^2_m and R^2_c are 0.3695 and 0.4165. The Log-shift transformation gives R^2_m and R^2_c of 0.3747 and 0.4212, respectively. This finding is similar to that in Rojas-Perilla et al. (2020). It has to be noted, that data-driven scaled transformations perform only slightly better with regards to R^2_m and R^2_c than the adhoc chosen transformations for instance the Log in this case. This is the case with this particular data. With

other datasets, the results could favour the data-driven even more, since the parameters of the transformation are chosen to give the best fit to the data at hand.

Table 1.6: Values of R_m^2, R_c^2, λ and ICC under no transformation, logarithmic, Log-shift and Box-Cox transformations.

Transformation	R_m^2	R_c^2	ICC	λ
No	0.2051	0.2185	0.0167	-
Log	0.3641	0.4122	0.0756	-
Log-shift	0.3747	0.4212	0.0743	309.6276
Box-Cox	0.3695	0.4165	0.0745	0.047527

To further explore the validity of the normality assumption we present here probability quantile plots($Q - Q$) of the EBP only under the working model.

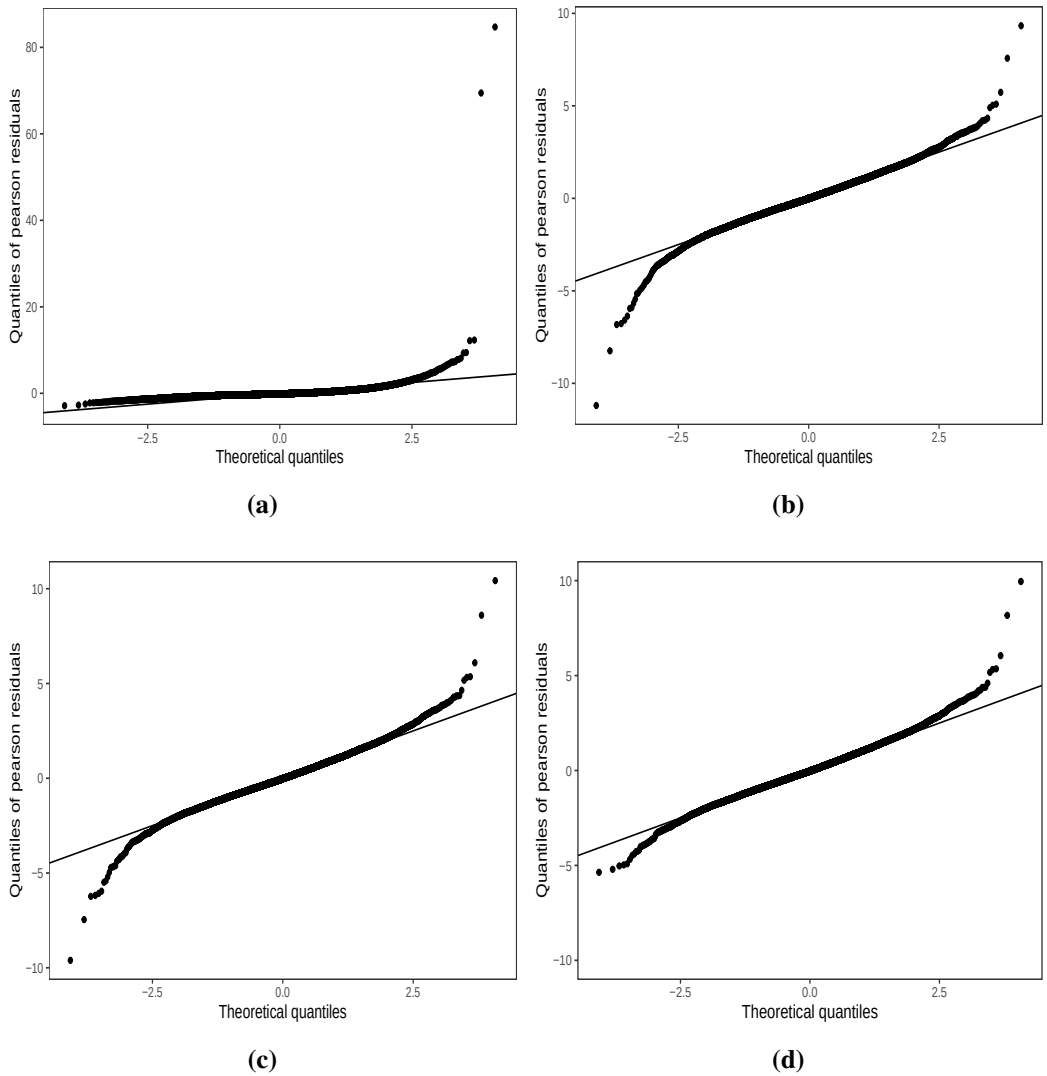


Figure 1.1: $Q - Q$ plots for Pearson residuals under: (a) No transformation (b) Log transformation (c) Box-Cox transformation and (d) Log-shift transformation

We also present density plots in Figure 1.2 of the Pearson residuals under no transformation, Log transformation, Box-Cox transformation and Log-shift transformation.

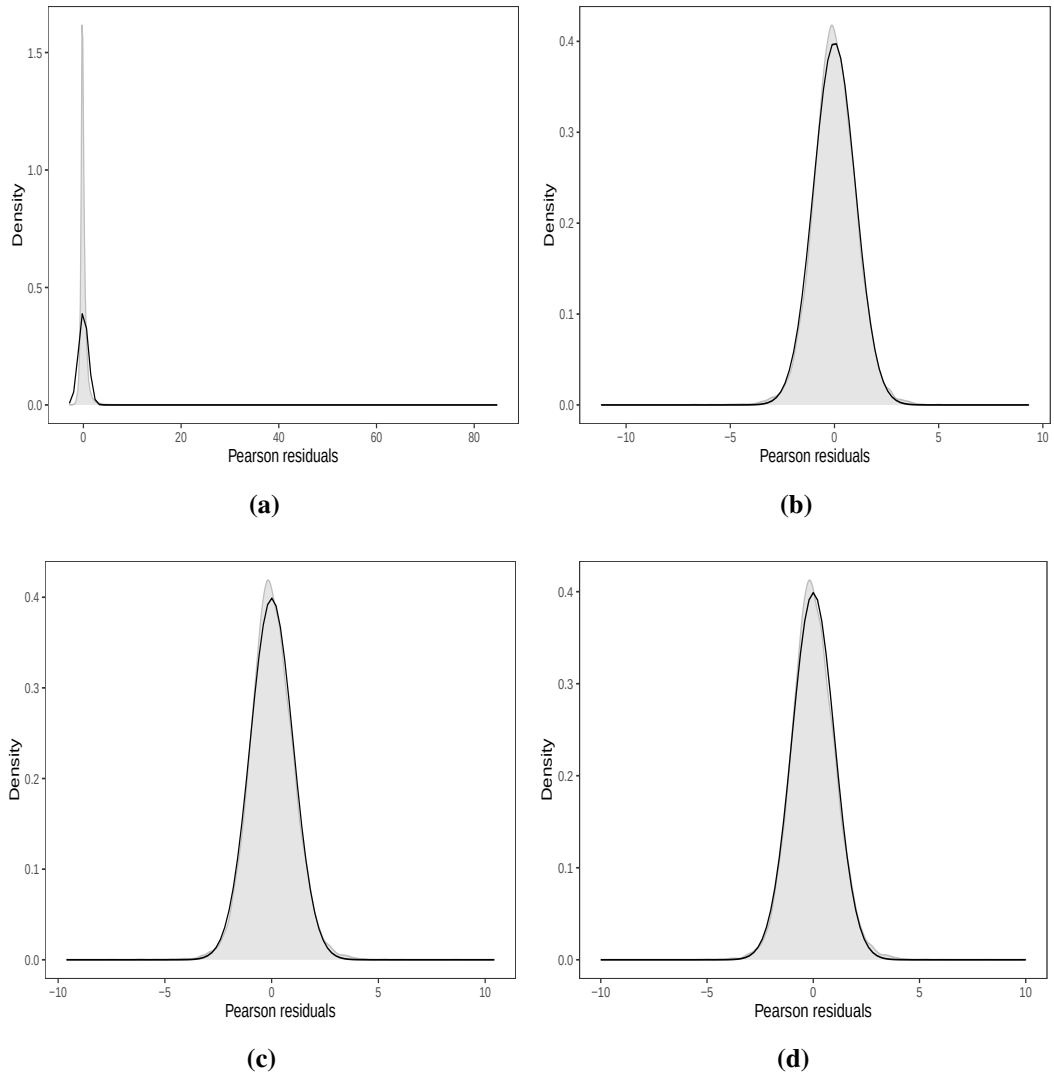


Figure 1.2: Density plots for Pearson residuals under: (a) No transformation (b) Log transformation (c) Box-Cox transformation and (d) Log-shift transformation

The $Q - Q$ plots in Figure 3.3 and density plots in Figure 1.2 further confirm the need for transformation and thus data-driven scaled transformations. Though, we note the normality assumption seems to be achieved by the three transformation (Log, Box-Cox and Log-shift) at least approximately. The skewness and kurtosis of should be close to zero and three respectively, since these are the theoretical values of a normal distribution. Corresponding values can be seen in Table 1.7 below. In terms of graphical examination, as well as skewness, all three transformations support the claim of symmetrically distributed household level errors, although the kurtosis being slightly to high. However the Kolmogorov-Smirnov (KS) test contradicts

these observations with all p-values being smaller then the usual significance level α of 0.05, which leads to the rejection of the null hypothesis of normally distributed household level errors. Since the KS-test tends to reject the null hypothesis very often in large samples, it can be concluded that all transformations lead to a large improvement in terms of the household errors being closer to a normal distribution.

Table 1.7: Skewness, kurtosis and values of the Kolmogorov-Smirnov (KS) p-values for the Pearson residuals of the working models for the EBP under the various transformations

Transformation	Skewness	Kurtosis	KS, p-values
No	43.8458	3374.7546	0.0032
Log	-0.1019	5.3859	0.0010
Box-cox	0.0826	5.1429	0.0032
Log-shift	0.2061	4.2962	0.0008

In the EBP, the random effects are assumed to be independent and identically normally distributed with mean zero and a constant variance σ^2. Again the skewness and kurtosis therefore should be close to zero and three respectively. Corresponding values can be seen in Table 1.8 below. In terms of graphical examination, as well as skewness and kurtosis all three transformations support the claim of normally distributed random effects, although the KS-test once more rejects normality in all cases.

Table 1.8: Skewness, Kurtosis and Kolmogorov-Smirnov p-values for the household level random effects for the EBP under the various transformations.

Transformation	Skewness	Kurtosis	KS, p-value
No	0.1429	3.2147	0.0000
Log	-0.2672	2.7001	0.0000
Box-Cox	-0.1959	2.6487	0.0001
Log-shift	-0.1681	2.6458	0.0000

For this paper, the Box-Cox transformation was chosen as transformation, since there is not a lot of difference between the transformations with regards to R^2 and the normality of residuals. The Box-Cox transformation is one of the most well known and very often used data-driven transformations in applications. Furthermore the Box-Cox transformation has an easy interpretation because two special cases exist. If the optimal λ is estimated to be zero, the Box-Cox transformation is the same as to take the logarithm of the data in question shifted by a shift parameter s and if the optimal value is one, then only a shift by s is executed. In terms of numbers of Monte-Carlo runs, 100 was chosen for the model-based approaches. For the

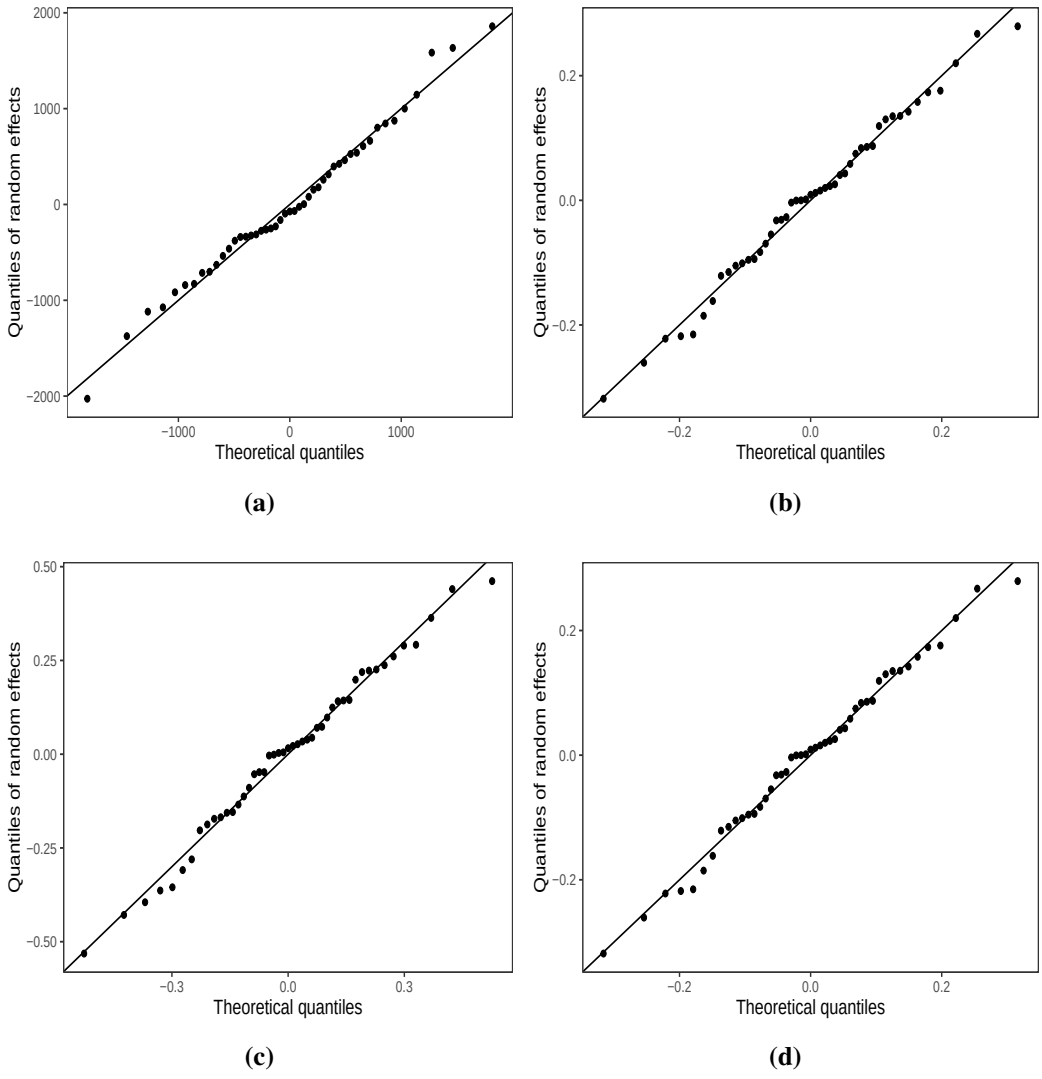


Figure 1.3: $Q - Q$ plots for quantiles of random effects under: (a) No transformation (b) Log transformation (c) Box-Cox transformation and (d) Log-shift transformation

MSE estimation, all methods use a bootstrap approximation, except for the m-quantile-model estimation of the mean indicator. There the MSE is estimated analytically as in Marchetti et al. (2012). The number of bootstrap runs was set to 200 for the direct estimator (non-parametric) and the EBP (parametric). For the M-quantile estimator, 50 were chosen in accordance with Marchetti et al. (2012), together with a number of 100 bootstrap samples. As a stability check for the MSE estimator of the M-quantile model the procedure was also run with 100 bootstrap populations. The differences were negligible and therefore the MSE estimates can be regarded as stable already with 50 bootstrap populations.

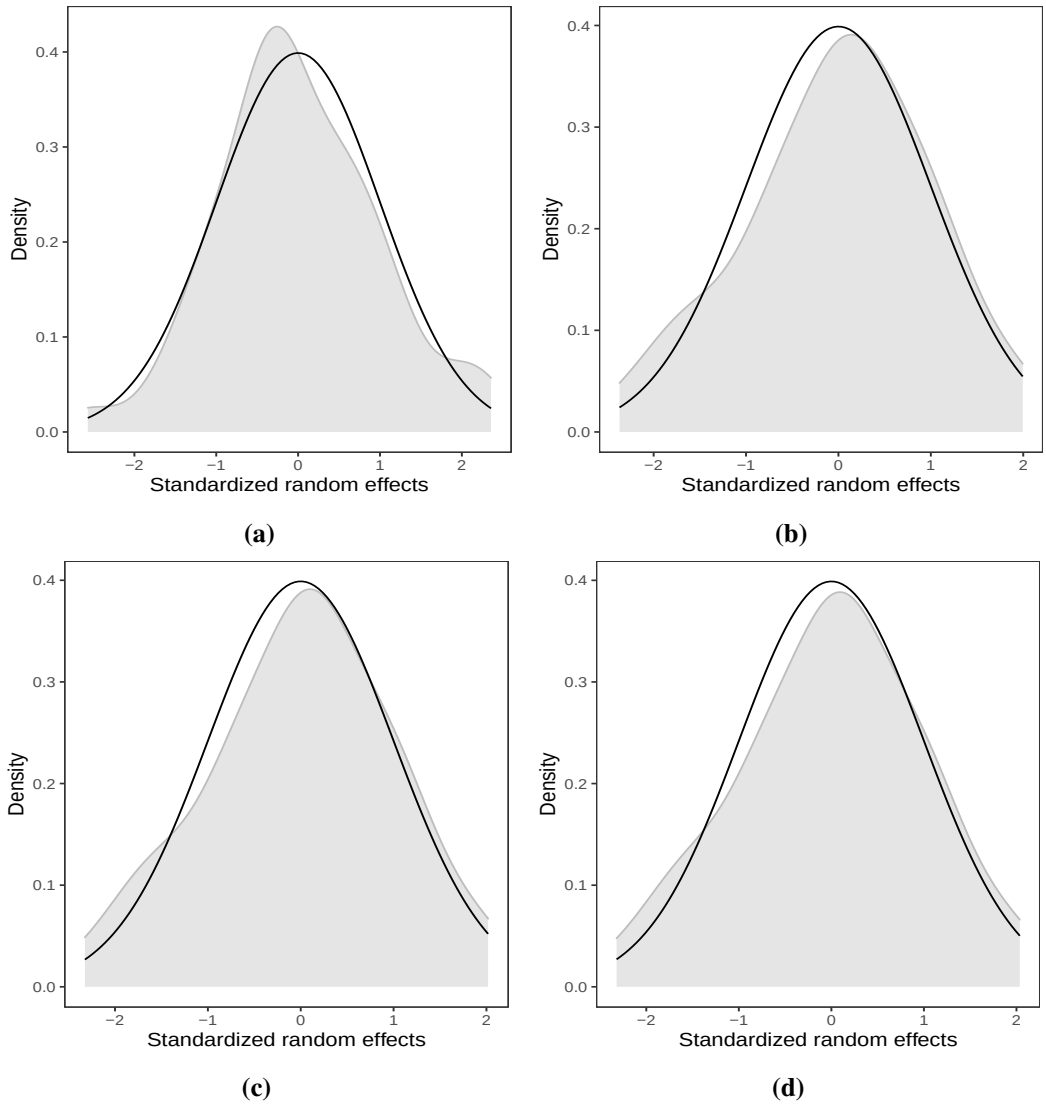


Figure 1.4: Density plots for random effects under: (a) No transformation (b) Log transformation (c) Box-Cox transformation and (d) Log-shift transformation

1.4.1 Mean

The first indicator of interest that is discussed in this section is the Mean expenditure. In comparison to the other indicators, the mean is quite a simple indicator in the sense that it is linear. This could be the reason why all applied methods agree to a certain extend with regards to the results. Table 1.9 and Figure 1.5 substantiate this claim. Overall, the estimated values for the direct estimator are highest ranging between 3,463 KES and 14,343 KES as county averages. The EBP under Box-Cox transformation has the smallest minimum value and the M-quantile model has the smallest maximum value for the indicator. The figure demonstrates that all three methods show a very similar pattern, when plotting the point estimates for all domains by increasing sample size. The highest and lowest estimated mean expenditures are reached by the counties of Nairobi and Mandera for all three methods. Overall there is accordance, when it comes to ordering the estimated values from high to low. The reason why the direct estimator seems to match the results of the model-based methods so well seems to be that firstly there are no missing domains and secondly there is no county with a very small sample size.

Table 1.9: Summary of point estimates for the Mean indicator in KES for all compared methods

Method	Min.	1st Quartile	Mean	Median	3rd Quartile	Max.
Direct estimator	3,463	4,563	6,440	6,242	6,211	14,343
EBP Box-Cox	2,858	4,321	5,420	5,292	6,073	13,019
M-quantile model	3,383	5,023	5,591	5,408	7,315	12,840

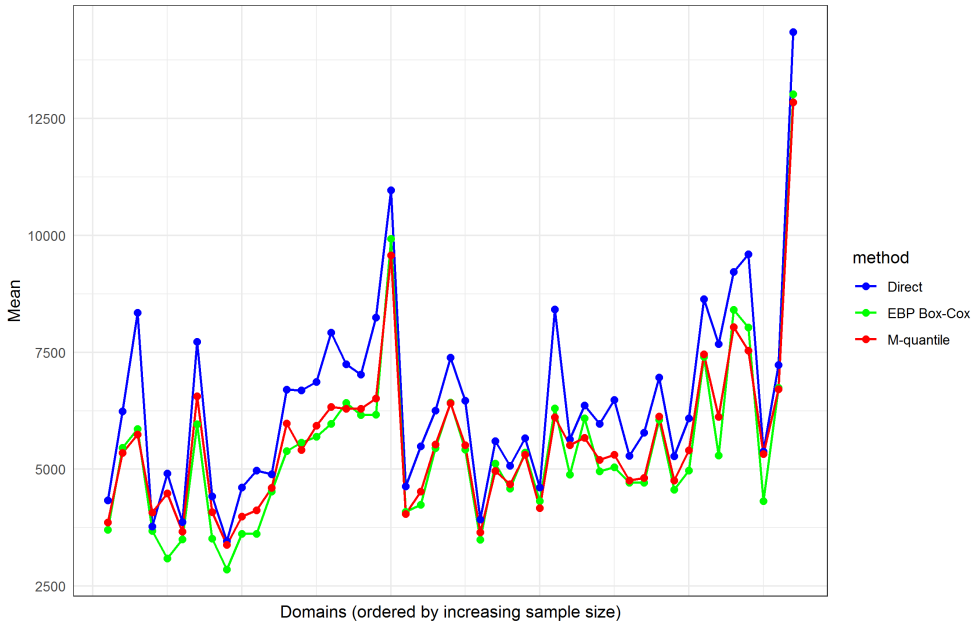


Figure 1.5: Point estimates mean comparing all estimators

After looking at the point estimates for the mean indicator, it is important to also take into account a measure of uncertainty with regards to the estimation. As for the point estimates Table 1.10 shows the minimum, 1st quartile, mean, median, 3rd quartile and maximum levels of MSE estimates. As can be seen in the Table, the EBP-method has the most stable values out of all the methods. The direct estimator as well as the M-quantile model both have substantially higher mean values than the EBP.

Table 1.10: Summary of MSE estimates for the mean indicator for all compared methods

Method	Min.	1st Quartile	Mean	Median	3rd Quartile	Max.
Direct estimator	16,505	36,281	97,721	66,970	105,603	865,182
EBP Box-Cox	14,641	23,016	29,577	25,584	30,316	101,862
M-quantile model	12,834	30,690	83,492	39,066	54,848	107,5513

The Office for National Statistics of the UK considers a Coefficient of Variation (CV) below the threshold of 20% as sufficiently good to be published (Office for National Statistics UK, 2017). In this application on Kenyan data, all compared methods stay below this cutoff value for all counties (see Table 1.11). The EBP-method has the lowest CV values ranging between 2% and 5%. The M-quantile approach mostly also stays way below 20%, although having two outlying CV's for Kisii (19.5%) and Kisumu (14.9%). Even the direct estimator without the possibility of borrowing strength from other areas via a model does not reach CV values, which might be considered too high to be reliably published. They range between 2.8% and 12.1%. For the mean indicator on county level we conclude based on these results, that even the direct estimates seem to be reliable. This result might lead one to question the use of more complex model-based estimators. Government bodies have data access for lower levels than the county. We can see, that for this level of dis-aggregation almost the same CV's can be reached with the direct estimator as with the EBP and the M-quantile model, but for lower levels with usually smaller sample sizes and often out-of-sample domains, the model-based approaches are expected to play to their strengths even more.

Table 1.11: Summary of CV estimates for the mean indicator for all compared methods

Method	Min.	1st Quartile	Mean	Median	3rd Quartile	Max.
Direct estimator	0.0282	0.0335	0.0443	0.0407	0.0463	0.1212
EBP Box-Cox	0.0207	0.0277	0.0327	0.0317	0.0376	0.0506
M-quantile model	0.0265	0.0319	0.0433	0.0371	0.0415	0.1948

Overall one can deduct, that all three methods agree more or less for the mean indicator regarding their point estimates, but the EBP method's values for the MSE are more stable than

for the other methods. There is agreement in the ordering of counties, with Nairobi having the highest and Mandera having the lowest estimated value. This maybe comes to no surprise, since Nairobi is the capital city of Kenya. In general, counties with higher mean values of expenditure are located towards the southern and south-western parts of the country in the direction of Tanzania and the Indian ocean. Counties located towards the northern parts of Kenya, especially bordering Somalia, have lower estimated mean expenditure. Therefore, one can see a clustering of counties in regards to their estimates (see maps 1.9).

1.4.2 Head Count Ratio

The next indicator of interest is the HCR, which estimates the percentage of observational units, that live in poverty, i.e. have a value of expenditure below a certain threshold. In applications, the threshold is often chosen to be 60% of the median expenditure in the sample. The 60% figure is used by Eurostat to define the head count ratio (Eurostat, 2021). In this application, two thresholds were chosen, in order to take into account different costs of living between rural and urban areas (Kenya National Bureau of Statistics, 2018). For households living in rural areas, the threshold is 3252.735 KES and for urban areas the thresholds is quite a lot higher with 5995.902 KES. It is important to state, that the rural and urban areas do not have to coincide with the small areas under investigation in this paper. Most counties are made up of both types and therefore both thresholds are being used. The two exceptions are Nairobi and Mombasa, where only the higher cut-off value for urban areas is applied. The idea of using two thresholds was used for all competing estimators, to make them comparable.

Table 1.12 in conjunction with Figure 1.6 show, first of all, that there are stark differences in proportions of households living in poverty, when comparing the counties. The maximum of 72.90% for the direct estimate can be observed in the county of Mandera, where the mean expenditure already was estimated to be the lowest. Consistently the smallest value is estimated for Nairobi with 11.08%, which had the highest mean expenditure. For the EBP estimator under Box-Cox transformation, we observe similar results. With an estimate of 73.55%, Mandera is estimated to have the highest percentage of households living under the poverty threshold. Other than Nairobi having the smallest value, here Nyeri only has 13.37% estimated, whereas Nairobi follows second with 16.50%. For the M-quantile model, firstly, we observe a minimum of 11.98%, which is estimated for the county of Nairobi, very much aligning with the other estimators. Deviating from the others the highest value for the M-quantile approach is only

56.89%, estimated for the county of Samburu. Two things stand out: Samburu does not rank lowest or second lowest among the other estimators, although being on the poorest counties for the direct and the EBP estimator. Secondly, the maximum value for the M-quantile model is around 15 to 16 percentage points smaller than for the comparing methods. The model produces more stable point estimates around the mean than its competitors (see 1.6).

A poverty and equity brief by the World Bank indicates that nationally the proportion of persons living below the national poverty line reduced from 46.8% in 2005/2006 to 36.1% in 2015/2016. Rural areas experienced more reduction in poverty from 50% in 2005/2006 to 38.8% in 2015/2016 (World Bank, 2021). In this period, urban areas outside Nairobi experienced stagnating poverty incidences. It is also noted that counties to the north and north-eastern Kenya lag behind than the rest of the counties. Particularly, Turkana, Marsabit, Mandera, Samburu and Wajir saw little progress in the 10 years period. These counties are food insecure as they are majorly nomadic pastoral who move from one place to another. Due to prolonged droughts and famine, it poses a significant threat to livestock, the main source of food and income for nearly all of the people who live in this area. When it rains the area experience flash floods and livestock and lives are lost. These counties are also prone to bandit and terrorist attack (Schilling et al., 2012; Haider, 2020). The northern and north eastern parts of the country lack behind in terms of infrastructure, lack of access roads, electricity, water, illiteracy, job opportunities and other social amenities. Health care and basic security are almost completely non-existent (Friedrich Ebert Stiftung, 2012; Fitzgibbon, 2012). The Kenya Institute for Public Policy Research and Analysis (KIPPRA) showed in an economic report of 2020, that overall poverty incidence varies highly among counties, starting with a low of 16.7% in Nairobi to a high of 79.0% in Turkana (Kenya Institute for Public Policy Research and Analysis, 2020). It also depict that counties with the lowest Gross County Product (GCP) per capita have the highest poverty rates. GCP is a geographical breakdown of Kenya's Gross Domestic Product (GDP) that gives an estimate of the size and structure of county economies (Kenya National Bureau of Statistics, 2019a). Again these are counties mostly in arid and semi-arid lands. It is also noted that these counties have the largest household sizes, i.e. Mandera (6.9), Wajir (6.1) and Garissa (5.9), where poverty rates are 77.6%, 62.6% and 65.5%, respectively (Kenya Institute for Public Policy Research and Analysis, 2020).

The maps 1.11 in the Appendix complete the picture of the point estimates for the HCR. In comparison to the maps for the mean expenditure in section 1.4.1, one can observe, that counties

towards the northern part of the country tend to have higher estimated values for the indicator, whereas the southern part of the country has lower values in comparison. This means, that for the counties, where the mean expenditure tends to be higher, the proportion of people living under the poverty threshold is lower and vice versa.

Table 1.12: Summary of point estimates for the Head Count Ratio indicator for all compared methods

Method	Min.	1st Quartile	Mean	Median	3rd Quartile	Max.
Direct estimator	0.1108	0.2181	0.3370	0.3065	0.3898	0.7290
EBP Box-Cox	0.1337	0.3024	0.4064	0.3926	0.4747	0.7355
M-quantile model	0.1198	0.3200	0.3851	0.3891	0.4418	0.5689

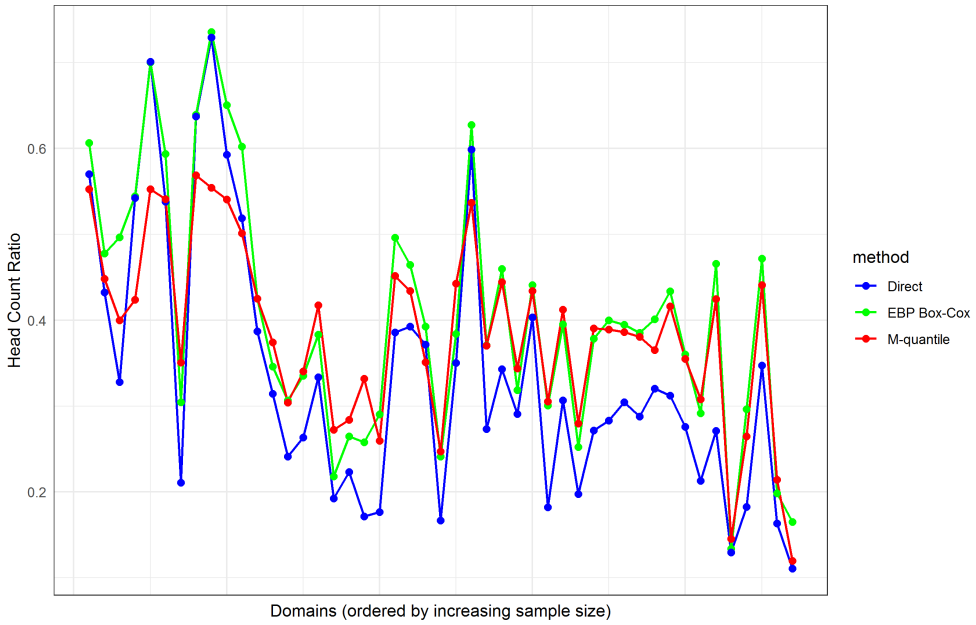


Figure 1.6: Point estimates Head Count Ratio comparing all estimators

With regards to the uncertainty of the point estimates, Table 1.13 displays the results. The MSE estimates for the M-quantile model are overall the lowest, which can be deducted from the lowest minimum value among the methods, as well as the lowest mean. On the other hand, the estimates also spike the most. Therefore, this method also has the highest MSE value. Estimates for EBP and direct estimator are more stable, with the EBP having less spikes than the direct estimator.

Table 1.13: Summary of MSE estimates for the Head Count Ratio indicator for all compared methods (all values $*10^{-4}$)

Method	Min.	1st Quartile	Mean	Median	3rd Quartile	Max.
Direct estimator	2.334	4.705	6.115	5.616	6.963	14.796
EBP Box-Cox	2.222	3.597	4.891	4.755	6.037	8.608
M-quantile model	0.955	1.285	3.634	2.494	4.260	21.136

The CVs for the HCR mostly stay below the reliability threshold of 20%, the exception being the county Nyeri, where the CV value for the M-quantile approach is 20.4%. The EBP ranges between 2.9% and 12% having lower CV values than the direct estimator for almost all counties. The CVs for the direct estimates themselves vary between 3.7% and 14.7% showing acceptable levels for all investigated areas (see Table 1.14).

Table 1.14: Summary of CV estimates for the Head Count Ratio indicator for all compared methods

Method	Min	1st Quartile	Mean	Median	3rd Quartile	Max
Direct estimator	0.0379	0.0676	0.0827	0.0761	0.1003	0.1467
EBP Box-Cox	0.0289	0.0483	0.0591	0.0565	0.0644	0.1197
M-quantile model	0.0213	0.0291	0.0512	0.0397	0.0560	0.2045

By and large the results for the HCR indicator agree with the results of the mean expenditure investigated in the last section. Where the mean expenditure is overall higher, the HCR is lower and vice versa. Therefore, we again see some north-south divide in the country. The poorer counties are located towards the border of Somalia and Ethiopia, whereas the southern counties towards Tanzania and the southern part of Uganda seem to fare better with regards to the proportion of households living under the poverty threshold.

1.4.3 Poverty Gap

The third indicator of interest in this paper is the Poverty Gap, which expresses the ratio by which the mean expenditure of poor households falls below a chosen poverty threshold in relation to the threshold. In this case, again two different thresholds are used for rural and urban areas as in subsection 1.4.2. Table 1.15 and Figure 1.7 sum up the results. From the figure, it can be deducted, that all methods results show a similar pattern with regards to the point estimate,

although the M-quantile model has mostly the highest estimates, the EBP's estimates lie in between and the estimated values using the direct method are the lowest for most counties. One remarkable exception is the maximum value for the direct estimator with a value of 37.98%, which constitutes the highest estimate among all estimates. For this method, the county with the highest PG is Turkana, followed by Mandera. Nyeri and Nairobi again have the lowest values compared to the rest of the counties. The EBP results agree with that, just having the order reversed for the poorest counties. Mandera has the highest estimate, followed by Turkana. Nyeri has the lowest estimate and Nairobi the second lowest. As with the HCR, the M-quantile method has Samburu as the area with the highest estimated value in contrast to the other estimators. Turkana follows with the second hightest PG though. As for the counties with the lowest Poverty Gap, the M-quantile method lists Nairobi with 0.029, followed by Nyeri with 0.051. Taking into account the information of maps 1.13 one again can confirm the diagonal divide through the country going from north-west to south-east. Counties north of the line overall have higher estimated PGs, than counties south of the line. The areas with the lowest Poverty Gaps are located in the center around Nairobi.

These findings concur with the other studies. According to Kenya National Bureau of Statistics and Society for International Development (2013), the PG is lowest in counties around the center of the country and parts of the southeast. This is the same picture depicted by the HCR and the mean estimator. The authors established that Tana River (46.1%), Kwale (41.8%) and Mandera (32.2%) have the highest PG. These findings clearly illustrate how much each county needs in order to pull the population out of poverty since the PG basically estimates how much on average is needed to bring each every household above the poverty line. They are helpful to the national government as it can aid in disbursement of finances to the counties. Counties which lack behind should be considered for more allocations. The objective of establishing county governments, was to promote economic development and make services more accessible to the citizens (Government of Kenya, 2013). In terms of governance, since the county governments have been in existence for over 10 years now, these results is a reflection of county government performances in terms of creating enabling environment for improved lives. Transparency International Kenya (TIK), conducted a survey on the state of governance based on transparency and accountability (Transparency International Kenya, 2016). The findings indicate that the biggest success of devolution was ease of access to services. The biggest failure was increased corruption and embezzlement of funds.

Table 1.15: Summary of point estimates for the Poverty Gap indicator for all compared methods

Method	Min	1st Quartile	Mean	Median	3rd Quartile	Max
Direct estimator	0.0160	0.0508	0.0988	0.0782	0.1094	0.3798
EBP Box-Cox	0.0319	0.0880	0.1348	0.1224	0.1601	0.3104
M-quantile model	0.0292	0.1287	0.1671	0.1657	0.1976	0.2989

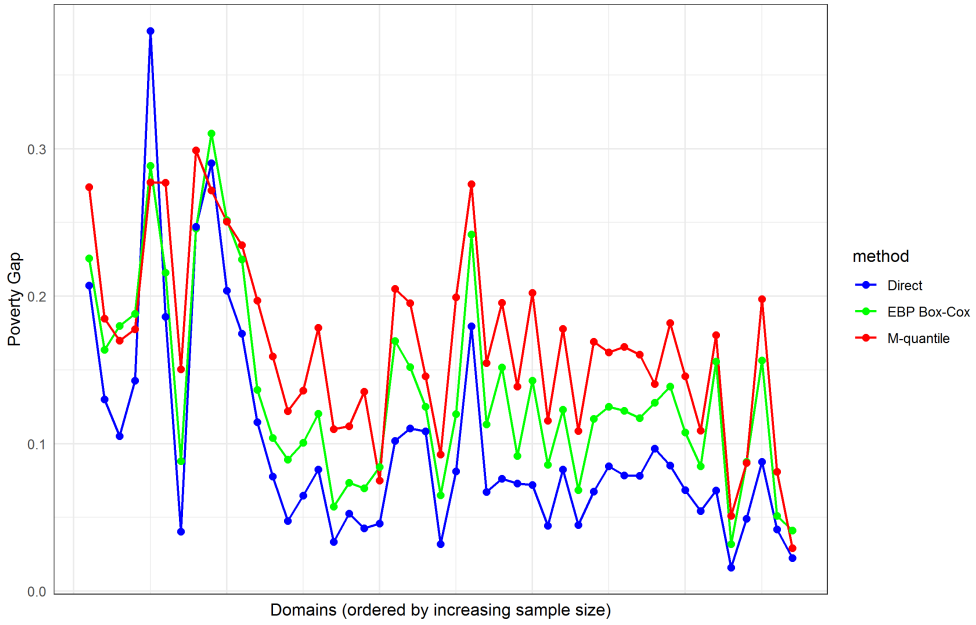


Figure 1.7: Point estimates Poverty Gap comparing all estimators

When looking at the MSE estimates for this indicator (Table 1.16), one can see, that again the estimates are more stable for the direct estimator and the EBP method. The robust approach has a maximum value for the MSE of 0.00094860, whereas the direct estimator (0.00042901) and the EBP (0.00024274) stay well below that. Overall the counties with smaller sample sizes tend to have higher estimated uncertainty than the counties with larger sample sizes. This holds true especially well for the direct estimator.

Table 1.16: Summary of MSE estimates for the Poverty Gap indicator for all compared methods (all values $*10^{-4}$)

Method	Min	1st Quartile	Mean	Median	3rd Quartile	Max
Direct estimator	0.1352	0.5980	1.0955	0.8275	1.2141	4.2901
EBP Box-Cox	0.3288	0.6831	1.0959	1.1994	1.4238	2.4274
M-quantile model	0.3848	0.9637	2.3044	1.5899	3.1229	9.4860

For the Poverty Gap, one can observe (see Table 1.17) that for the direct estimator as well as the M-quantile approach there are counties with CV values above 20%. The direct estimator has five counties that are deemed unreliable (Nyeri (25.2%), Kirinyaga (24.2%), Lamu (21%), Narok (20.4%) and Meru (20.3%)), whereas for the M-quantile model there are three counties (Nyeri (27%), Narok (22.8%) and Nairobi (22.1%)) above the threshold. The largest CV for the EBP method is 18.8% in Nyeri. Overall CVs are higher for the Poverty Gap indicator, than they are for the Head Count Ratio across all three methods.

Table 1.17: Summary of CV estimates for the Poverty Gap indicator for all compared methods

Method	Min	1st Quartile	Mean	Median	3rd Quartile	Max
Direct estimator	0.0530	0.0973	0.1242	0.1178	0.1424	0.2524
EBP Box-Cox	0.0383	0.0673	0.0883	0.0853	0.0987	0.1880
M-quantile model	0.0488	0.0625	0.0937	0.0762	0.1034	0.2709

1.4.4 Gini coefficient

The fourth indicator, that will be inspected in this paper is the Gini coefficient, measuring inequality of household expenditure. As with the mean of the household expenditure indicator a poverty line is not needed.

A study done in 2013 by the Kenya National Bureau of Statistics and Society for International Development (2013) indicate that household consumption expenditure varies nationally and in rural and urban areas in Kenya. In the report, the population is divided into quintiles, where each represents 20% of all households in Kenya in ascending order. Consumption expenditure varies more in urban areas than rural areas and rural areas show small differences. The ratio of the top quintile to the bottom quintile for rural and urban areas is 6.4 and is 6.6, respectively. At county level, the inequalities in consumption expenditure is more pronounced. The counties Nairobi, Mombasa and Kiambu show significant differences with the 5th quintile spending more than the 1st by 691 times in Nairobi, 75 in Mombasa and 20 Kiambu. It should be noted that Nairobi and Mombasa are classified as urban areas and Kiambu has both rural and urban. In the bottom 10 counties, 8 of them have at least 50% of their population in the bottom 1st quintile spending 1,440 KES or lower as compared to only 0.6% of the population in Nairobi.

It stands out immediately from Table 1.18 and Figure 1.8 the three methods do not agree as

much with each other, than as with the other indicators. The M-quantile method overall tends to have the highest estimates with 0.26 being the minimum (Nairobi) and 0.49 being the maximum (Samburu). The direct estimator on the other hand estimates the lowest Gini coefficient for Wajir (0.27) and the highest for Turkana with 0.56 (followed by Samburu), which is the highest estimate overall. Nairobi has an estimated Gini of 0.34, which is very close to the average for this method. The estimates of the EBP method are way more stable than the others, ranging from 0.31 (Wajir) to 0.4 for Kajiado. Again Nairobi ranks in the middle of all counties with 0.33. Since the country wide expenditure Gini has been estimated as 0.445 (Kenya National Bureau of Statistics and Society for International Development, 2013), these EBP results show, that counties themselves seem to be more homogeneous than the whole country is. This explanation, why the EBP results make a lot of sense even when the direct estimates do not agree as much stems from the fact, that it can also be observed for the Gini's of rural and urban areas overall with estimates of 0.361 and 0.368 respectively from the same report. The country divide shown for the other indicators cannot be confirmed for any of the used methods. The regional clustering of counties with regards to indicator estimates is by far not as strong as in the other sections of the application in this paper (Figure 1.15).

From these results, there is wide variation in household expenditure across counties in Kenya. With the new constitution in 2010, the county governments are important regions for policy-making. The report shows there has been a reduction in inequality between 1994 and 2015/16 (Kenya National Bureau of Statistics, 2020). Out of the 47 counties, 35 have experienced a reduction in inequality while 12 increased in the same period. Nairobi county experienced the highest reduction, while the highest increase is Turkana. Overall, there is a decline in inequality between 1994 and 2015/16. Several factors can be attributed to this decline. First, the increase in the share of expenditure going to the middle 50% and lower 40% and the fall in the share of expenditure going to the top 10% for the entire population and for almost all population groups. Secondly, with formation of counties, more job opportunities have been created. More funds have also been disbursed to the counties. This increased income levels. With every county managing it's own funds, economic growth has been experienced especially for counties that were historically marginalized (Kenya National Bureau of Statistics, 2020).

Table 1.18: Summary of point estimates for the Gini indicator for all compared methods

Method	Min	1st Quartile	Mean	Median	3rd Quartile	Max
Direct estimator	0.2701	0.3294	0.3559	0.3515	0.3806	0.5623
EBP Box-Cox	0.3127	0.3292	0.3398	0.3365	0.3465	0.3995
M-quantile model	0.2608	0.3797	0.4010	0.4035	0.4290	0.4894

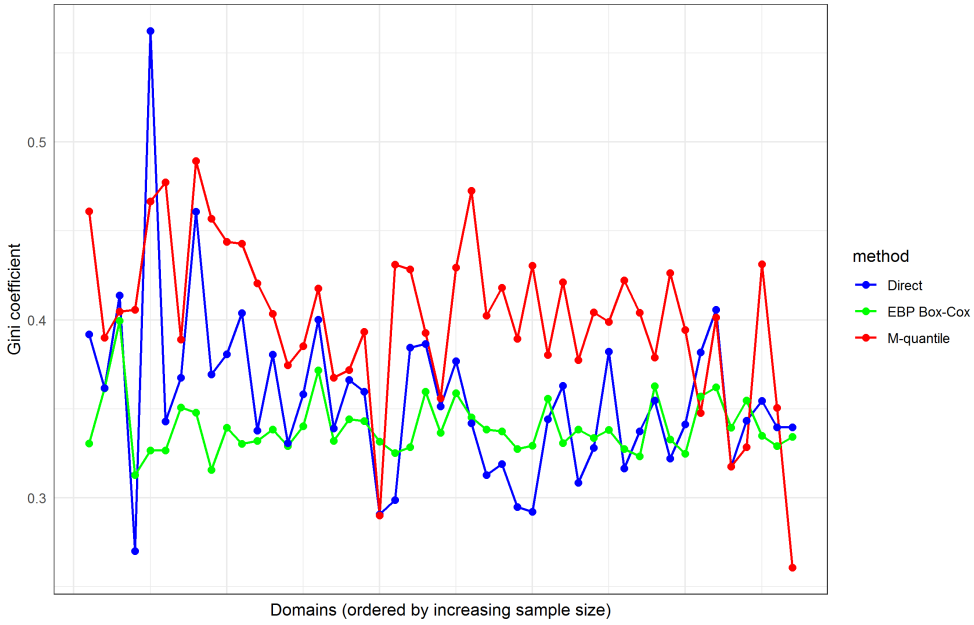


Figure 1.8: Point estimates Gini comparing all estimators

Table 1.19 summarizes the results of the MSE estimates of the Gini inequality indicator. The direct estimator exhibits the most variation in its MSE estimates. There are even two bigger spikes for Kisumu and Kisii, which are counties with one of the biggest sample sizes. For this indicator, the EBP method shows very stable uncertainty having the lowest estimates among the compared methods. The M-quantile model estimates values in between the competitors for the most part.

Table 1.19: Summary of MSE estimates for the Gini indicator for all compared methods (all values $*10^{-4}$)

Method	Min	1st Quartile	Mean	Median	3rd Quartile	Max
Direct estimator	0.8829	1.5724	3.3691	2.0515	3.3145	37.3667
EBP Box-Cox	0.0704	0.1796	0.2382	0.2280	0.2747	0.5193
M-quantile model	0.1730	0.7627	1.3094	0.9647	1.8097	3.6886

With the Gini coefficient, we observe low CV values (see Table 1.20) pretty much across all methods. No estimate reaches the 20% threshold for this indicator. The direct estimator ranges between 3% and 15.1% conveying the widest range. For this indicator, the EBP consistently has the lowest CVs varying between 0.8% and 2%. The outlier-robust approach has values between the direct and the EBP.

Table 1.20: Summary of CV estimates for the Gini indicator for all compared methods

Method	Min	1st Quartile	Mean	Median	3rd Quartile	Max
Direct estimator	0.0303	0.0365	0.0460	0.0402	0.0457	0.1506
EBP Box-Cox	0.00841	0.01257	0.0131	0.01429	0.01536	0.02055
M-quantile model	0.0160	0.0225	0.0270	0.0247	0.0317	0.0460

1.5 Conclusion

This paper shows, that combining data sources like the Kenya Integrated Household Budget Survey from 2015 and the Kenya Population and Housing Census from 2009 can improve upon county level estimates of poverty and inequality indicators done with a direct estimator. Although the direct estimates in a lot of cases are reliable in the sense, that they lie below the 20% threshold for CV's, which is used to determine if the results can be usefully published or not. The reason for this is, that Kenyan counties in the survey data have more observations to use in the estimation process, that one would normally deem too few. Nonetheless, the CV's for the model-based methods are mostly lower. This is the case especially for the EBP, which in comparison to both other estimators has the most stable MSE estimates.

This analysis of Kenyan data on county level can serve as a case study, that Small Area although more complicated in application than a direct estimator have to potential help produce statistical offices to provide estimations of poverty and inequality indicators with a high level of reliability. Statistical offices can use their data access to provide results on a lower regional level than the authors of this paper. A key point is that the desired indicators can also be estimated for out-of-sample domains, which is just not possible with direct estimators. Since often policy decisions are based on such estimates policies can in turn be targeted more specifically on low regional levels. One important take-away from this analysis is that Kenya's counties are more homogeneous themselves than the country as a whole. This can be seen on the one hand by the north-east to south-west divide for the chosen poverty indicators. Counties in the

north-eastern part of the country are considerably poorer than their south-western counterparts as our estimates show. On the other hand, it is very noticeable with the estimation of the Gini coefficient. The EBP estimates range only between 0.31 and 0.40 approximately, while the country wide expenditure Gini estimate has a value of 0.445. The same mechanism can be seen in rural and urban areas overall with estimates of 0.361 and 0.368, respectively (Kenya National Bureau of Statistics and Society for International Development, 2013). This paper shows, that employing model-based SAE methods, especially the EBP under data-driven transformations can help to improve results from direct estimation techniques. The gain in reliability by reducing MSE estimates is tremendous. One drawback is that one relies on good data sources for the second dataset. In this case, only the 2009 census was available. Between survey and census lie six years, which could mean, that relationships estimated in the employed models assume relations between covariates and target variable that might have changed. Once the new census data from 2019 is available, it is important to check the results from this paper against the ones reached with the newer data source.

Acknowledgments

The authors are grateful to Patrick Krennmair and Marina Runge for providing useful comments that improved this paper.

Appendix

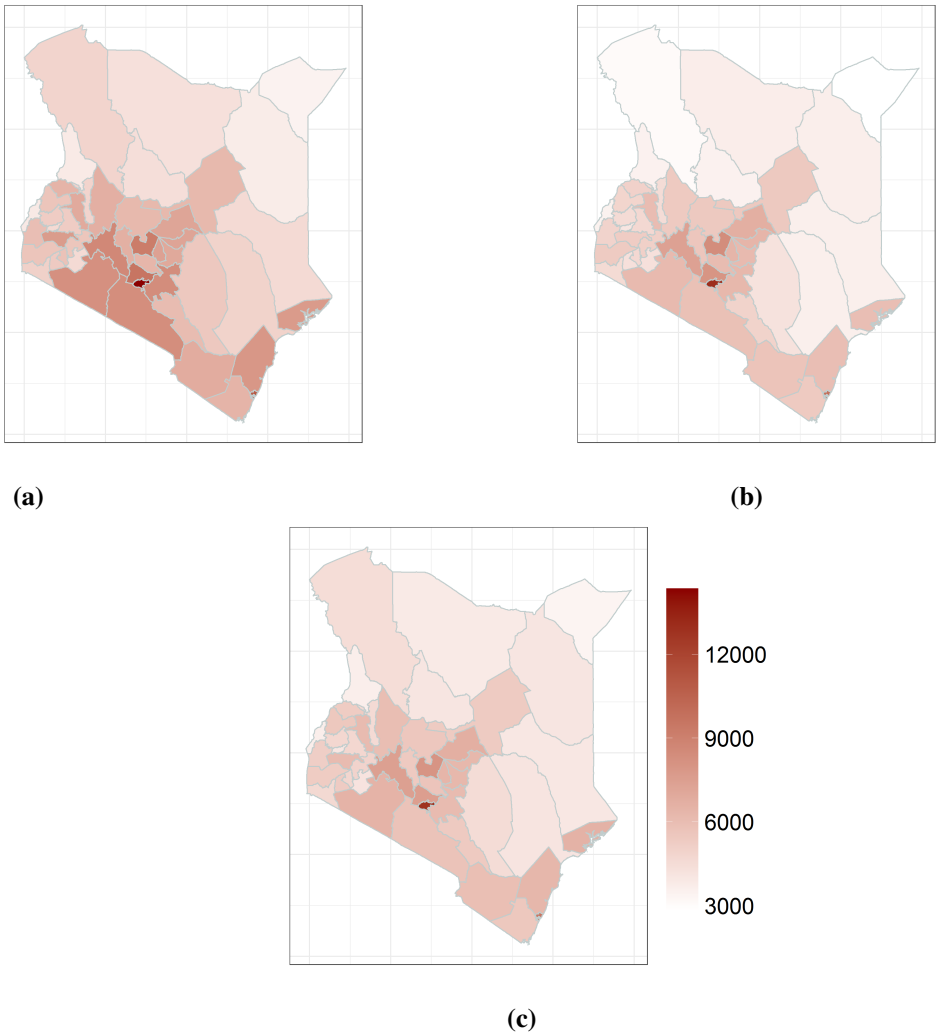


Figure 1.9: Point estimates of the mean indicator for a) Direct, b) EBP under Box-Cox transformation and c) M-quantile model estimator

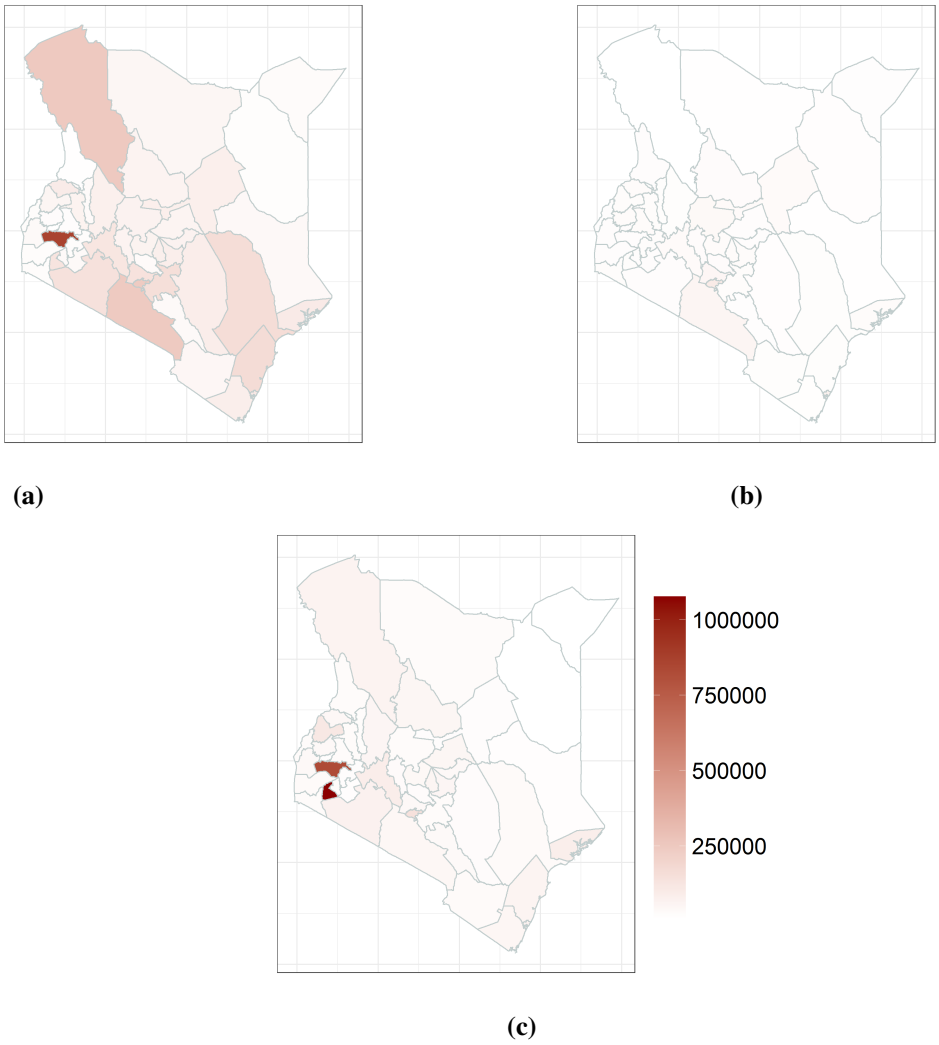


Figure 1.10: MSE estimates mean indicator for a) Direct, b) EBP under Box-Cox transformation and c) M-quantile model estimator

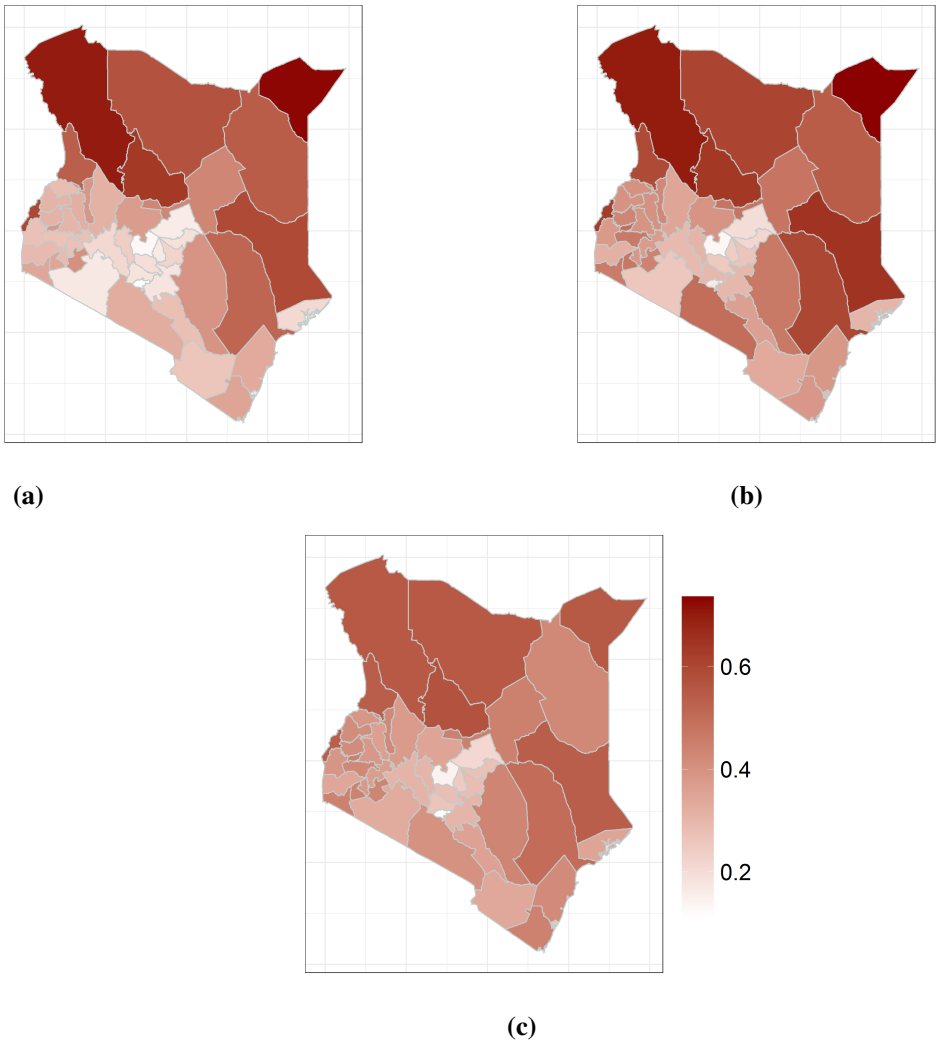


Figure 1.11: Point estimates Head Count Ratio indicator for a) Direct, b) EBP under Box-Cox transformation and c) M-quantile model estimator

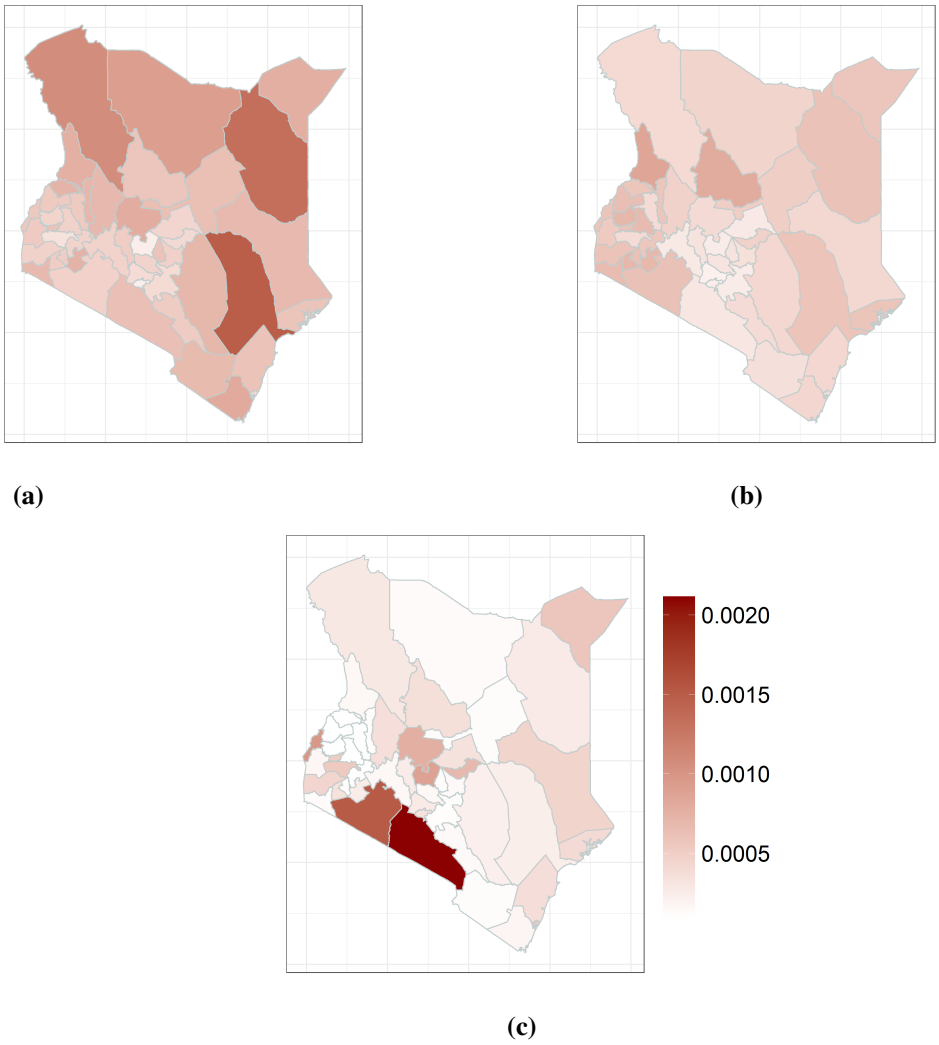


Figure 1.12: MSE estimates Head Count Ratio indicator for a) Direct, b) EBP under Box-Cox transformation and c) M-quantile model estimator

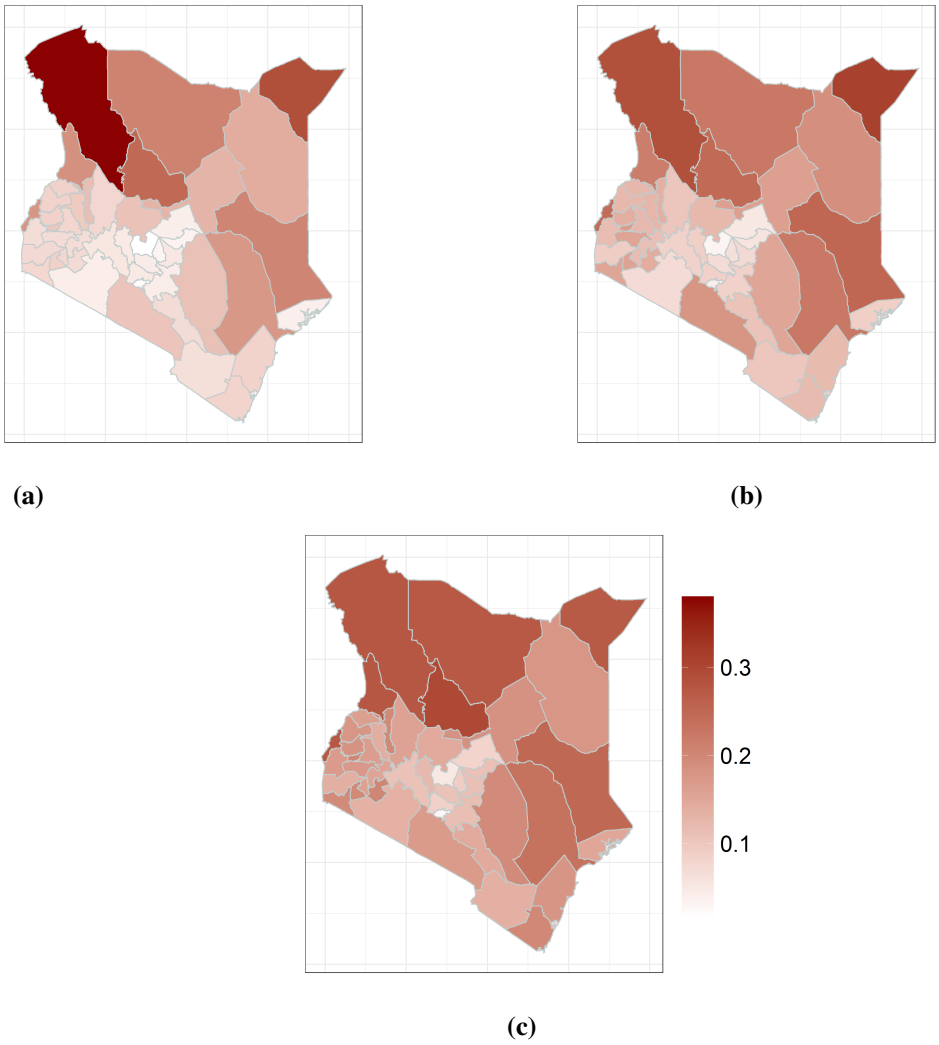


Figure 1.13: Point estimates Poverty Gap indicator for a) Direct, b) EBP under Box-Cox transformation and c) M-quantile model estimator

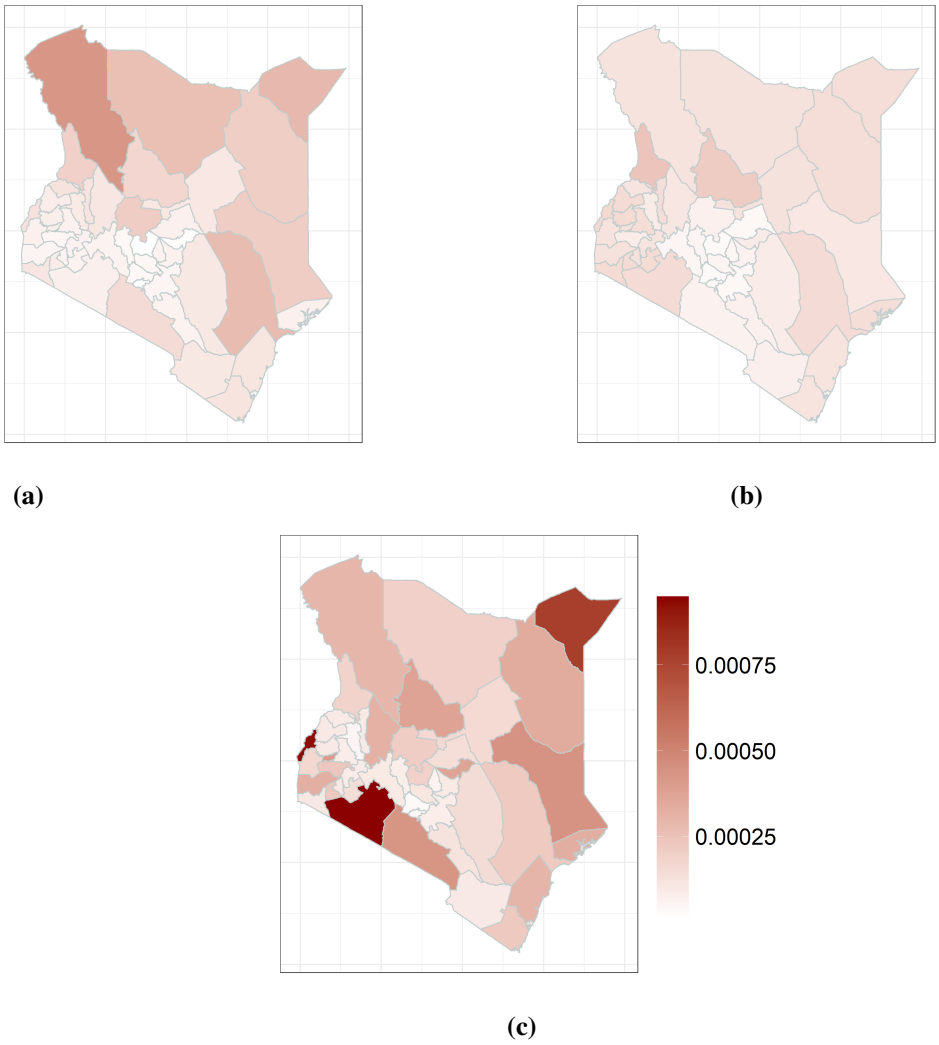


Figure 1.14: MSE estimates Poverty Gap indicator for a) Direct, b) EBP under Box-Cox transformation and c) M-quantile model estimator

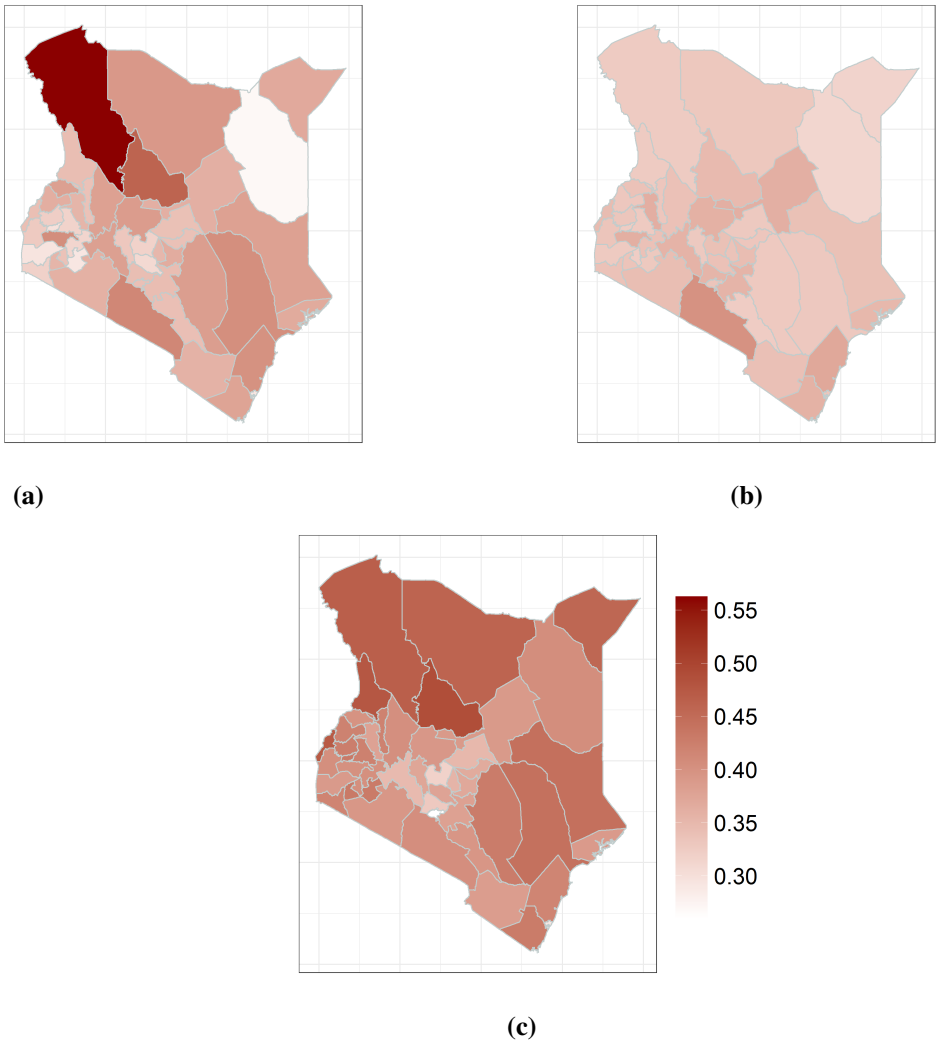


Figure 1.15: Point estimates Gini indicator for a) Direct, b) EBP under Box-Cox transformation and c) M-quantile model estimator

Figure 1.16: MSE estimates Gini indicator for a) Direct, b) EBP under Box-Cox transformation and c) M-quantile model estimator

Chapter 2

Small area estimation of health insurance coverage for Kenyan counties

2.1 Introduction

Health insurance reduces extreme health costs and out-of-pocket spending by pooling resources. It is an important component towards the attainment of Universal Health Care (UHC) (Dye et al., 2015). The goal of UHC was set by the World Health Organization (WHO) member states in 2005 (World Health Organization, 2005a). The goal is to assist member countries to achieve UHC through health system financing. UHC has been defined as the provision of the needed quality health services to the whole population with less cost (World Health Organization, 2013). In 2015, the General Assembly adopted the 2030 agenda that includes 17 Sustainable Development Goals (SDGs). The SDG goal 3.8 seeks: To attain UHC, with financial risk security, access to quality vital health care services and inexpensive key medicines and vaccines for everyone (General Assembly, 2015).

Countries in Sub-Saharan Africa face many health challenges. These include; low investment in health care, slow economic growth, extensive out-of-pocket expenditure and reduced access to health services (Sambo et al., 2013). To achieve the health related SDG's and UHC, the regional committee for Africa suggested strategies including; more investment, efficient use of health resources and expand coverage. The objective is to foster efficient and sustainable health financing and achieve these goals. Over the years these countries have prioritized investments towards achieving UHC (Lagomarsino et al., 2012; Cotlear and Rosemberg, 2018). This also follows the "Abuja Declaration" of World Health Organization et al. (2010) that set a minimum

of 15% of the total government expenditure.

To mention a few countries; Ghana's National Health Insurance Scheme (NHIS) has been in existence since 2003. The goal is to guarantee fairness and access to health care services by reducing financial barriers to access at the point of use (Kusi et al., 2015). By 2014, over 10.5 million Ghanaians (an estimated 40% of the population) were covered by the NHIS, with inpatient and outpatient visits to health facilities increasing from 0.5 to around 3 per capita between 2005 and 2014 (Wang et al., 2017). A study by Aikins et al. (2021) established that the scheme will likely achieve UHC if protected from political interference and improved accountability. In 2017 Zambia developed the National Health Strategic Plan 2017-2021. The plan outlined strategies to establish a social health insurance scheme (Ministry of Health Zambia, 2017). This was passed into the National Health Insurance Act 2018 whose goal is to provide reliable health system financing and universal access to health services. Under the NHI, all eligible citizens contribute to the pool of resources in addition to external funding. Households classified as poor by measuring absolute poverty based on monthly consumption expenditures are exempted from contributing (Ministry of Health Zambia, 2017).

Kenya is a lower-middle-income country with a population of 47.5 million, 12.2 million households and an average household size of 3.9. In Kenya, 75.1% are below 35 years and 32.73 million (68.9%) live in rural areas (Kenya National Bureau of Statistics, 2019b). Approximately 83% of Kenyans do not have financial protection against extreme health care costs. Around 1.5 million become poor due to health care costs (Ministry of Health, 2014; Okungu et al., 2017). As outlined in Vision 2030 Kenya seeks to achieve UHC by 2030. Towards this goal, several strategies have been implemented. To start with, the government piloted UHC in four out of 47 counties in Kenya (Isiolo, Kisumu, Machakos, and Nyeri). These counties were selected because they have a high incidence of both communicable and non-communicable diseases, maternal mortality, and road traffic injuries. Results from this pilot showed great success, however, not sustainable by the government capitation (Ministry of Health, Kenya, 2018). Secondly, the government abolished charges in public hospitals and health care. It also introduced free maternity services in all health care facilities (Maina and Kirigia, 2015). Further, it expanded the National Health Insurance Fund (NHIF) package from an inpatient-only package to outpatient services (Mwaura et al., 2015).

Some studies on health insurance in Kenya include; Kazungu and Barasa (2017), where they examined the levels, inequalities (where households are categorized into five socio-economic

quintiles) and factors associated with health insurance coverage in Kenya. They analyzed data from the Kenya Demographic and Health Survey (KDHS) 2009 and 2014. Results show health insurance coverage in Kenya remains low and show high inequalities. Otieno et al. (2019) carried out a study to determine the prevalence of health insurance and associated factors among households in urban slum settings in Nairobi, Kenya. They used cross-sectional data of adults aged 18 years or older from randomly selected households in Viwandani slums (Nairobi, Kenya). The study was conducted between June and July 2018. Their findings show that the prevalence of health insurance in the sample was 43%. Health insurance coverage in Viwandani slums in Nairobi, Kenya, is low.

The KDHS 2014 was a national survey. It was designed to provide reliable direct estimates at the national level only and county estimates for some selected indicators (health insurance not included). The direct estimators (they rely only on the survey data) are approximately designed unbiased and consistent (Pfeffermann, 2013). However, direct estimators generally have large variances and estimates are unreliable when the sample sizes are small (Rao and Molina, 2015) — for example at the county level in Kenya. In contrast, model-based small area methods produce more reliable estimates in terms of smaller mean squared error and coefficient of variation (Tzavidis et al., 2018). This is because they combine survey and census/administrative data through a model and therefore increase the effective sample size. For more theory on small area estimation, we refer the reader to Ghosh and Rao (1994), Pfeffermann (2002), Jiang and Lahiri (2006), Pfeffermann (2013), Rao and Molina (2015), Pratesi (2016), Tzavidis et al. (2018) and Morales et al. (2021).

For this study, therefore, we rely on SAE to better estimate the proportion of persons with health insurance for Kenyan counties. The health insurance status of a person is binary. Some approaches have been proposed to estimating binary variables in the small area context. To mention a few Bayesian approaches; Hierarchical Bayes of Malec et al. (1997), Nandram et al. (1999), Liu et al. (2007) and Empirical Bayes of MacGibbon and Tomberlin (1987), Farrell et al. (1997) and Ghosh et al. (1998). For frequentist approaches Jiang and Lahiri (2001) proposed empirical best predictor (EBP) for a binary response. Chambers et al. (2016) outlines the use of a binary logistic generalized linear mixed model (GLMM) in SAE. However, they note that GLMM based on maximum likelihood is influenced by outliers. M-quantile SAE model provides a robust alternative to GLMM's.

The rest of this paper is organized as follows. We describe the KDHS 2014 and the Kenya

Population and Housing Census (KPHC) 2009 in section 2.2. In section 2.3, we outline the statistical methodology applied in this paper. In particular, the direct estimation and the binary M-quantile small area model estimation, hereafter called the MQ model, are examined by means of the point estimation and the mean squared error. In section 2.4, we present the results of the application to estimate health insurance coverage for Kenyan counties. Lastly, in section 2.5, we give the concluding remarks based on the findings, some possibilities for further research as well limitations.

2.2 Data description

In this section, we describe the data sources used in this paper. We had access to KDHS 2014 and KPHC 2009. The links to the data sources are provided at the end of this paper. We assume that the functional relation between having health insurance and auxiliary data remains constant between the survey and census time.

2.2.1 The Kenya Demographic and Health Survey

The Demographic and Health Survey (DHS) collects, analyzes and disseminates data on population, health, HIV and nutrition in over 90 countries (Croft et al., 2018). In this study, we had access to the Kenya Demographic and Health Survey (KDHS) done in 2014. The KDHS has been conducted in Kenya after every 5 years i.e. 1989, 1993, 1998, 2003, 2008-2009 and 2014. The 2014 KDHS collected several data on household characteristics, education and employment, and health-related indicators such as HIV and child health survival (Kenya National Bureau of Statistics, 2015).

The 2014 KDHS sample was drawn from a sample master called the Fifth National Sample Survey and Evaluation Program (NASSEP V). The Kenya National Bureau of Statistics (KNBS) currently uses this framework to conduct household surveys in Kenya. It includes $5,360$ clusters derived from the 2009 Kenya Population and Housing Census (Kenya National Bureau of Statistics, 2015). The framework has a total of $96,251$ enumeration areas (EA's). The KDHS 2014 sought to create representative estimates for the majority of survey variables at the national level, for individual urban and rural regions, for regional (formerly provincial) levels, and selected indicators at the county level. To meet these objectives, the sample was designed to comprise 40,300 households from 1,612 clusters spread across the country, with 995 clusters in

rural areas and 617 clusters in urban areas. Samples were selected separately in each sampling stratum using a two-stage sample approach. In the first stage, the $1,612$ EA's were chosen with equal probability from the NASSEP V frame. The properties from the listing operations served as the sampling frame for the second round of selection, which included selecting 25 households from each cluster (Kenya National Bureau of Statistics, 2015).

Three main questionnaires were used in the KDHS; (i) A household questionnaire, (ii) A questionnaire for women aged 15 to 49, and (iii) A questionnaire for men aged 15 to 54. They were based on model questionnaires designed for the DHS program, as well as questionnaires used in earlier KDHS surveys and Kenya's current information needs. During the questionnaire development process, input was sought from relevant stakeholders and data users. Producing county-level estimates necessitated gathering data from a large number of families within each county, resulting in a significant rise in sample size from around 10,000 homes in the 2008-09 KDHS to 40,300 households in 2014 (Kenya National Bureau of Statistics, 2015).

A total of 39,679 households were selected in the sample, of which 36,812 were found occupied at the time of the fieldwork. Of these, 36,430 households were successfully interviewed, yielding an overall household response rate of 99%. The shortfall of households occupied was primarily due to structures that were found to be vacant or destroyed and households that were absent for an extended time. Among the households selected using the full questionnaires, a total of 15,317 women were identified as eligible for the full women's questionnaire, of whom 14,741 were interviewed, generating a response rate of 96% (Kenya National Bureau of Statistics, 2015). A total of 14,217 men were identified as eligible in these households, of whom 12,819 were successfully interviewed, generating a response rate of 90%. For this application, we use only complete cases for our variable of interest giving a total sample of 12,007 men and 14,730 women. County-specific samples sizes for women and men are summarized in Table 2.1. The women sample sizes are higher than for men because most indicators in the DHS (fertility, maternal mortality rate, infant mortality rate and neonatal mortality rate) relates to children and women.

Table 2.1: Summary of sample sizes for women and men over counties in the Kenya Demographic and Health Survey 2014.

Questionnaire	Min.	Q1.	Median	Mean	Q3.	Max.
Women	236.0	275.5	310.0	313.5	342.0	460.0
Men	118.0	227.0	250.0	255.5	287.0	370.0

2.2.2 Socio-demographic characteristics

Table 2.4 are socio-demographic characteristics of respondents in the survey. It included women (aged 15 to 49 years) and men (aged 15 to 54 years.) For comparison purposes, we selected men aged 15 to 49 years. The majority of the respondents are between ages 15 to 34 years. Kenya is composed majorly of a youthful population. According to the 2019 Kenyan census the median age is approximately 20 years. The survey also inquired whether respondents lived in urban or rural areas. Most women (63%) live in rural areas while most men (61%) live in urban areas. The KDHS 2014 was planned to give representative estimates for most of the survey indicators at the national levels. Other characteristics include education level (no education, primary, secondary or higher), wealth index (poorer, poorest, middle, richer and richest) and marital status (never married, married, widowed, separated, divorced). The majority of women and men are either never married or married.

2.2.3 Direct estimation and type of insurance per wealth quantile

The 2014 KDHS asked respondents if they were covered by any health insurance and, if yes, what type. We first estimated health insurance coverage for the whole country (using KDHS only). We used the R package emdi (Kreutzmann et al., 2019b) for direct estimation. Table 2.2 shows the percentage of women and men age 15-49 covered by health insurance at the national level together with the mean squared error and the coefficient of variation. A small percentage of Kenyans aged 15-49 (18% of women and 21% of men) have health insurance. The mean squared error values are very low, 0.000016 and 0.000027 for women and men respectively. Also, the coefficient of variation is 2.1% and 2.4% for women and men groups. Therefore the estimates are reliable (as expected at the design level of the survey).

Table 2.2: Estimated proportions (direct estimates) with health insurance, mean squared error and coefficient of variation for women and men at the national level in Kenya using the Kenya Demographic and Health Survey 2014.

Gender	Percent	Mean squared error	Coefficient of variation
Women	0.18	0.000016	0.021
Men	0.21	0.000027	0.024

In this paper, we are interested in estimating health insurance coverage at the county level. Table 2.3 below is a summary of the coefficient of variations(CV's) for the direct estimates. The CV's are quite high reaching values of 63% and 67% for women and men, respectively.

Model-based SAE methods that borrow strength from other counties of interest are required to increase the accuracy of the estimation.

Table 2.3: Descriptive statistics of the coefficients of variation for the distribution of health insurance coverage at the county level in Kenya for the direct estimates.

Gender	Min.	Q1.	Median	Mean	Q3.	Max
Women	0.073	0.120	0.157	0.186	0.215	0.634
Men	0.078	0.126	0.150	0.188	0.221	0.670

Table 2.5 shows the type of health insurance for each wealth quantile (built based on household asset data (Kenya National Bureau of Statistics, 2015)) category for women and men. Among those covered, the national insurance scheme is the most common type for both genders. Employer-based insurance is the next most common type of insurance. This is because employers are obliged by law to provide insurance to their employees. The trend in insurance coverage varies per wealth quintile, with the richer and richest most covered across all insurance types. Ilinca et al. (2019) also found out that there are significant levels of inequality in access to health services in Kenya across the wealth quintile. Mwenda et al. (2021) established that poor households pay more for health care especially for outpatient services. This is because poor households cannot pay or do not have health insurance hence more out of pocket spending. In the study they also noted that the rich also spent more on outpatient care owing to their financial abilities.

2.2.4 The Kenya Population and Housing Census

For model-based SAE we need supplemental data collected from all areas. We had access to the Kenya Population and Housing Census (KPHC) 2009 in this case. Kenya has consistently conducted a census every ten years, i.e. 1969, 1979, and so on, with the most recent being in 2019. Under Kenyan legislation, the KNBS is the primary government body in charge of collecting, processing, and disseminating census and other statistical data. Statistics are needed to track the progress of numerous development goals and worldwide initiatives, such as the SDG's. The main goal of the KPHC 2009 was to offer essential information on the population's demographic, social, and economic features, as well as housing. These include population size and composition, fertility, mortality and migration rates, levels of education, labour force size, and so on. The data for this census was taken using scanning technology, with technical help from the United States Census Bureau (USCB) (Kenya National Bureau of Statistics, 2010).

Table 2.4: Socio-demographic characteristics for women and men in the Kenya Demographic and Health Survey 2014.

Demographic characteristics	Women		Men	
Age	Frequency	Percent	Frequency	Percent
15-19	2859	19.4	2811	23.4
20-24	2537	17.2	1981	16.5
25-29	2859	19.4	1940	16.2
30-34	2104	14.3	1701	14.2
35-39	1876	12.7	1484	12.4
40-44	1367	9.3	1197	10.0
45-49	1128	7.7	893	7.4
Residence				
Rural	9262	62.9	4644	38.7
Urban	5468	37.1	7363	61.3
Region				
Central	1509	10.2	1246	10.4
Coast	1840	12.5	1503	12.5
Eastern	2494	16.9	2142	17.8
Nairobi	460	3.1	370	3.1
North Eastern	779	5.3	591	4.9
Nyanza	2010	13.6	1542	12.8
Rift Valley	4252	28.9	3483	29.0
Western	1386	9.4	1130	9.4
Education Level				
No education	1980	13.4	4124	34.3
Primary	7398	50.2	4570	38.1
Secondary	4103	27.9	1980	16.5
Higher	1249	8.5	1333	11.1
Wealth index				
Poorer	2864	19.4	2442	20.3
Poorest	3399	23.1	2503	20.8
Middle	2841	19.3	2465	20.5
Richer	2839	19.3	2578	21.5
Richest	2787	18.9	2019	16.8
Marital status				
Never married	9009	61.2	5742	47.8
Married	4053	27.5	5624	46.8
Widowed	580	3.9	56	0.5
Seperated	750	5.1	421	3.5
Divorced	338	2.3	164	1.4

Table 2.5: Percentages for each type of health insurance per socio-economic quintiles in Kenya from Kenya Demographic and Health Survey 2014.

Gender	Type	National	Employer	Mutual	Private	Prepayment	Other	None
Women	Wealth quintile							
	Poorest	1.54	0.24	0.15	0.12	0.00	0.00	97.96
	Poorer	4.78	0.95	0.11	0.35	0.00	0.11	93.71
	Middle	10.57	1.14	0.21	0.39	0.00	0.14	87.55
	Richer	18.41	2.28	0.57	0.39	0.00	0.21	78.14
	Richest	26.25	6.13	0.83	2.56	0.00	0.25	63.97
Men	Poorest	2.74	0.07	0.00	0.20	0.00	0.07	96.92
	Poorer	7.35	0.80	0.07	0.22	0.00	0.22	91.34
	Middle	11.54	1.04	0.30	0.52	0.00	0.15	86.45
	Richer	20.05	2.96	0.08	1.52	0.00	0.72	74.68
	Richest	30.53	10.96	0.10	4.20	0.00	0.61	53.59

This census was conducted based on old administrative areas, such as villages, sub-locations, locations, divisions, districts, and provinces. Kenya had 46 legal districts, minus Nairobi, the capital city, which was the 47th district. After 2010, these districts were changed to the present 47 counties with no changes in borders (Government of Kenya, 2013). As a result, we can connect the survey and census data. The census data serve as potential covariates in the small area model described in section 2.3.4.

Table 2.6: Summary of population sizes in Kenya Population and Housing Census 2009 at the county level in Kenya.

	Min.	1st Quartile	Mean	Median	3rd Quartile	Max.
Census	2,205	10,676	18,586	15,408	20,572	98,289

Table 2.6 is a summary of population sizes at the county level in Kenya. The census is the 10% sample, i.e. every 10th household of the whole data set is released by the KNBS (Kenya National Bureau of Statistics, 2010).

2.3 Statistical Methodology

In this section, we outline the methodology applied in this paper. To begin with we describe the direct estimation using the survey data only. We then introduce M-quantile regression differentiating it from standard mean regression. Next, we give the general small area estimation setting. Thereafter, we discuss the M-quantile small area model together with the point and mean squared error estimation.

2.3.1 Direct estimation

The Horvitz-Thompson (HT) estimator of Horvitz and Thompson (1952) is used to estimate the population proportion $\bar{Y}_i$ for area i, $i = 1, 2, ..., m$, where m is the total number of areas of the whole population, from a complex sampling design. Using this estimator, the direct estimator of the target proportion for area i based on sample data is defined as $\hat{\bar{y}}_i^{\text{dir}} = \frac{1}{N_i} \sum_{j=1}^{n_i} w_{ij}\, y_{ij}$, $i = 1, 2, ..., m$, where $\hat{\bar{y}}_i^{\text{dir}}$ is the direct proportion estimator for area i, N_i is the population size in area i, y_{ij} is the response of individual/household j in area i and w_{ij} are sampling weights — inverse of first order inclusion probabilities. The weights compensate for unequal probabilities of sampling and unit non-response. The HT-estimator for population proportions is design-unbiased (Särndal et al., 2003). However, the variance reaches high values for areas with small sample sizes. For KDHS 2014, all counties were sampled, although sample sizes in some counties are not sufficient to provide reliable direct estimates as seen from the high values of coefficient of variation (beyond 20% using guidelines set by the UK Office for National Statistics [ONS]).

2.3.2 Small area estimation

In SAE we assume the following idealized setting: There is a finite population U of size N which is divided into m disjoint areas of sizes $N_1, N_2, ..., N_m$ where $i = 1, 2, ..., m$ is the ith small area. A sample of size n is taken from this population using a complex sampling design with sample sizes $n_1, n_2, ..., n_m$ for each area i. The sampled and non-sampled units will be denoted by s and r respectively. Let y_{ij} be the response variable of interest of individual/household j in area i and has been observed for sampled units only; $\mathbf{x}_{ij}$ denote a $p \times 1$ vector of unit level covariates with intercept. In general it is assumed that the values of $\mathbf{x}_{ij}$ are known for all units in the population, as are the values $\mathbf{z}_i$ of a $q \times 1$ vector of area level covariates. We are interested in using sample values of y_{ij} and the population values of x_{ij} and z_i to estimate the small area i proportion of health insurance coverage given by $\bar{Y}_i = N_i^{-1} \sum_{j \in U_i} y_{ij}$.

2.3.3 M-quantile regression

The standard linear regression summaries the average relationship between a continuous response y_i given explanatory variables x_i i.e. $E[y_i|x_i]$ where $i = 1, 2, ..., n$ is the number of observations. This does not give a complete picture of the conditional distribution of the response variable given the explanatory variables, and we might be interested in other parts of this distribution

for example the $10th$ percentile. In the same manner, a relationship between the response and the explanatory variables can also be established using the conditional median function instead. The quantile $q \in (0, 1)$ is that y which splits the data into proportions q below and $(1 - q)$ above such that $F(y_q) = q$ and $y_q = F^{-1}(q)$. The median has $q = 0.5$. Whereas the mean regression minimizes the squared error, a regression model based on the median (or median regression) minimizes the least absolute deviation (LAD). Median regression is also more robust to outliers than mean regression and no parametric assumption is required. To generalize the mean and median regression, we discuss the expectile and quantile regression. Expectile regression (Newey and Powell, 1987) generalizes the mean regression to estimate the expectiles while quantile regression generalizes the median regression to estimate other parts of the conditional distribution (quantiles) of y given x (Koenker and Bassett Jr, 1978; Koenker and Hallock, 2001). M-quantile regression (Breckling and Chambers, 1988) estimates the conditional distribution lying between the quantiles and expectiles. It is an extension of M-estimation of Huber (1992). The M-quantile of order q of a continuous random variable y with distribution function $F(y)$ is the value Q_q that satisfies

$$\int \psi_q \left(\frac{y - Q_q}{\sigma_q} \right) dF(y) = 0, \tag{2.1}$$

where $\psi_q(t) = 2\psi(t)\{qI(t > 0) + (1 - q)I(t \leqslant 0)\}$, ψ is an influence function defined by the user and σ_q is an appropriate scale measure for the random variable $Y - Q_q$. According to Chambers and Tzavidis (2006) when the response variable y is binary, there is no obvious definition of a quantile function as in the continuous case in 2.1. But, given that the influence function ψ is continuous and monotone non-decreasing, the M-quantiles of a binary variable exist and are unique. In that case we are interested in predicting $P(y = 1) = p$ which means that 2.1 becomes

$$pq\psi \left(\frac{1 - Q_q}{\sigma_q} \right) = (1 - p)(1 - q)\psi \left(\frac{Q_q}{\sigma_q} \right). \tag{2.2}$$

Since y is binary, following Chambers and Tzavidis (2006), we impose a linear logistic function as

$$Q_q(\mathbf{x}_j; \psi) = \exp\left(\mathbf{x}_j^{\mathrm{T}} \boldsymbol{\beta}_q\right) \left\{1 + \exp\left(\mathbf{x}_j^{\mathrm{T}} \boldsymbol{\beta}_q\right)\right\}^{-1}, \tag{2.3}$$

where $\boldsymbol{\beta}_q$ are regression coefficients estimated using a robust maximum likelihood estimating

equations following (Cantoni and Ronchetti, 2001).

2.3.4 M-quantile small area model

Small area estimation mostly uses random-area effects to characterize between area variations beyond that explained by auxiliary variables in the model (Rao and Molina, 2015). However, mixed effect models depend on distributional assumptions (for example, the assumption of normally distributed residuals). Further, it requires the specification of the random part of the model. An alternative approach to mixed effect modeling is the use of M-quantile models in SAE. The M-quantile model for SAE was proposed by (Chambers and Tzavidis, 2006). They model the between-area heterogeneity using M-quantile coefficients. In this case, the population model is specified and fitted at unit level without specifying any small area geography. First define q_{ij} such that $y_{ij} = Q_{q_{ij}}(\mathbf{x}_{ij}; \psi)$, i.e. q_{ij} is a random index that varies between 0 and 1. Since the response variable is binary, we specify a linear logistic function, where the population M-quantile model for q_{ij} (and hence y_{ij}) is then defined by

$$Q_{q_{ij}}(\mathbf{x}_{ij}; \psi) = \exp\left(\mathbf{x}_{ij}^{\mathrm{T}} \boldsymbol{\beta}_{q_{ij}}\right) \left\{1 + \exp\left(\mathbf{x}_{ij}^{\mathrm{T}} \boldsymbol{\beta}_{q_{ij}}\right)\right\}^{-1}. \tag{2.4}$$

Chambers and Tzavidis (2006) called q_{ij} the M-quantile coefficients. The M-quantile coefficient for area i is given by; $\theta_i = E[q_{ij} \mid i]$, where the expectation is conditional on the distribution of the random indices q_{ij} within area i.

2.3.5 Point estimation

To estimate the population proportion we proceed as follows. We first note that the empirical value $\hat{q}_{ij}$ of the random index q_{ij} is the solution to $y_{ij} = \hat{Q}_{\hat{q}_{ij}}(\mathbf{x}_{ij}; \psi)$ and this value is referred to as the estimated M-quantile coefficient of y_{ij} (Chambers and Tzavidis, 2006).

1. Obtain sample observations in area i using a non-informative sampling method (for example a two-stage cluster sampling design).

2. Derive the Estimate $\hat{\theta}_i$ of the area i specific M-quantile coefficient θ_i as the sample average of the estimated M-quantile coefficients for that area; otherwise it is set $\hat{\theta}_i = 0.5$.

3. Compute the corresponding M-quantile predictor of the average $\bar{y}_i$ in small area i as $\hat{\bar{y}}_i^{MQ} = N_i^{-1}\left\{\sum_{j \in s_i} y_{ij} + \sum_{j \in r_i} \hat{Q}_{\hat{\theta}_i}(\mathbf{x}_{ij}; \psi)\right\}$. If y is binary, model the regression

M-quantiles of order q by 2.4.

2.3.6 Estimating the mean squared error

In this study, we estimate the mean squared error (MSE) of the proposed estimator using the approach by Chambers et al. (2014) based on the linearisation approach. In essence, the assumption is that the working model used in concluding influences the obtained values from the area under study. Therefore, the MSE of interest is relied upon it and is equal to a conditional prediction variance plus a squared conditional prediction bias. Further, the estimated area level M-quantile coefficient values are assumed as having some slight variations and can be considered as fixed. According to Chambers et al. (2016), a first order approximation to the conditional prediction variance of $\hat{\bar{y}}_i^{MQ}$ is then

$$\begin{aligned}\operatorname{Var}\left(\hat{\bar{y}}_i^{MQ} - \bar{y}_i \mid \theta_i\right) =& N_i^{-2}\left\{\operatorname{Var}\left[\sum_{j\in r_i}\hat{Q}_{\theta_i}\left(\mathrm{x}_j;\psi\right)\right] + \sum_{j\in r_i}\operatorname{Var}\left(y_j\right)\right\} \\ \approx& N_i^{-2}\left\{\left[\sum_{j\in r_i} Q_{\theta_i}\left(\mathrm{x}_j;\psi\right)\mathbf{x}_j^T\right]\operatorname{Var}\left(\hat{\boldsymbol{\beta}}_{\theta_i}\right)\left[\sum_{j\in r_i} Q_{\theta_i}\left(\mathrm{x}_j;\psi\right)\mathbf{x}_j^T\right]^T\right. \\ &\left. + \sum_{j\in r_i}\operatorname{Var}\left(y_j\right)\right\},\end{aligned} \tag{2.5}$$

which can be estimated by

$$\begin{aligned}\widehat{\operatorname{Var}}\left(\hat{\bar{y}}_i^{MQ}\right) =& N_i^{-2}\left\{\left[\sum_{j\in r_i}\hat{Q}_{\hat{\theta}_i}\left(\mathrm{x}_j;\psi\right)\mathbf{x}_j^T\right]\widehat{\operatorname{Var}}\left(\hat{\boldsymbol{\beta}}_{\hat{\theta}_i}\right)\left[\sum_{j\in r_i}\hat{Q}_{\hat{\theta}_i}\left(\mathrm{x}_j;\psi\right)\mathbf{x}_j^T\right]^T\right. \\ &\left. + \sum_{j\in r_i}\widehat{\operatorname{Var}}\left(y_j\right)\right\}.\end{aligned} \tag{2.6}$$

In this case $\widehat{\operatorname{Var}}\left(\hat{\boldsymbol{\beta}}_{\hat{\theta}_i}\right)$ is a sandwich-type estimator. The $\widehat{\operatorname{Var}}\left(y_j\right)$ can be calculated either by using the sample data from area i, $\widehat{\operatorname{Var}}\left(y_j\right) = \hat{\bar{y}}_i\left(1-\hat{\bar{y}}_i\right)$, or by pooling data from the entire sample, in which case $\widehat{\operatorname{Var}}\left(y_j\right) = \hat{\bar{y}}(1-\hat{\bar{y}})$. According to Chambers et al. (2016) the pooled estimator should lead to more stable prediction variance estimates when area sample sizes are very small and the conditional prediction bias can be approximated using the results of Copas

(1988) as

$$E\left(\hat{\bar{y}}_i^{MQ} - \bar{y}_i \mid \theta_i\right) \approx -\frac{1}{2N}\left\{\frac{\partial}{\partial \boldsymbol{\beta}_{\theta_i}}\Psi\left(\boldsymbol{\beta}_{\theta_i}\right)\right\}^{-1}\left\{\operatorname{tr}\left[\left\{\frac{\partial}{\partial \boldsymbol{\beta}_{\theta_i}\partial \boldsymbol{\beta}_{\theta_i}^T}\Psi\left(\boldsymbol{\beta}_{\theta_i}\right)\right\}\operatorname{Var}\left(\hat{\boldsymbol{\beta}}_{\theta_i}\right)\right]\right\} \left\{\frac{\partial}{\partial \boldsymbol{\beta}_{\theta}}\sum_{j\in r_i} Q_{\theta_i}\left(\mathrm{x}_j;\psi\right)\right\}, \tag{2.7}$$

with corresponding plug-in estimator

$$\widehat{\operatorname{Bias}}\left(\hat{\bar{y}}_i^{MQ}\right) = -\frac{1}{2N}\left\{\frac{\partial}{\partial \boldsymbol{\beta}_{\theta_i}}\Psi\left(\boldsymbol{\beta}_{\theta_i}\right)\Big|_{\boldsymbol{\beta}_i=\hat{\boldsymbol{\beta}}_{\hat{\theta}_i}}\right\}^{-1}\left\{\operatorname{tr}\left[\left\{\frac{\partial}{\partial \boldsymbol{\beta}_{\theta_i}\partial \boldsymbol{\beta}_{\theta_i}^T}\Psi\left(\boldsymbol{\beta}_{\theta_i}\right)\Big|_{\boldsymbol{\beta}_{\theta_i}=\hat{\boldsymbol{\beta}}_{\hat{\theta}}}\right\}\widehat{\operatorname{Var}}\left(\hat{\boldsymbol{\beta}}_{\hat{\theta}_i}\right)\right]\right\} \left\{\frac{\partial}{\partial \boldsymbol{\beta}_{\theta_i}}\sum_{j\in r_i} Q_{\theta_i}\left(\mathbf{x}_j;\psi\right)\Big|_{\boldsymbol{\beta}_{\theta_i}=\hat{\boldsymbol{\beta}}_{\hat{\theta}_i}}\right\}. \tag{2.8}$$

The estimator of the conditional mean squared error of $\hat{\bar{y}}_i^{MQ}$ is then

$$\operatorname{Mse}\left(\hat{\bar{y}}_i^{MQ}\right) = \widehat{\operatorname{Var}}\left(\hat{\bar{y}}_i^{MQ}\right) + \left\{\widehat{\operatorname{Bias}}\left(\hat{\bar{y}}_i^{MQ}\right)\right\}^2. \tag{2.9}$$

2.4 Results

In this section we present the results of estimating health insurance coverage in Kenya at the county level. The respondents were asked the question; Are you covered by any health insurance? Therefore the response is a binary variable coded as 0 (No) and 1 (Yes). We first fitted a binary logistic generalized linear mixed model (GLMM) and show that the assumption of normality of random effects is not met.

2.4.1 Initial analysis using binary logistic GLMM

To begin with we first fitted a binary logistic GLMM with normally distributed random effects using the function glmer in R package lme4 (Bates et al., 2015). Plots (a) and (b) from Figure 2.1 represent the QQ plots for Pearson residuals obtained from fitting a logistic GLMM for women and men respectively. Within the same fitted model, the random effects for women and men were obtained are also displayed in plots (c) and (d) respectively. The Pearson residuals

are not normally distributed. Although the random effects show normality especially for the men data, there is a slight departure from the tails. The Shapiro-Wilk normality test using significance level of 0.05 does not reject the null hypothesis that the random effects are normally distributed for men (p-value = 0.591) but it does for women (p-value=0.00473).

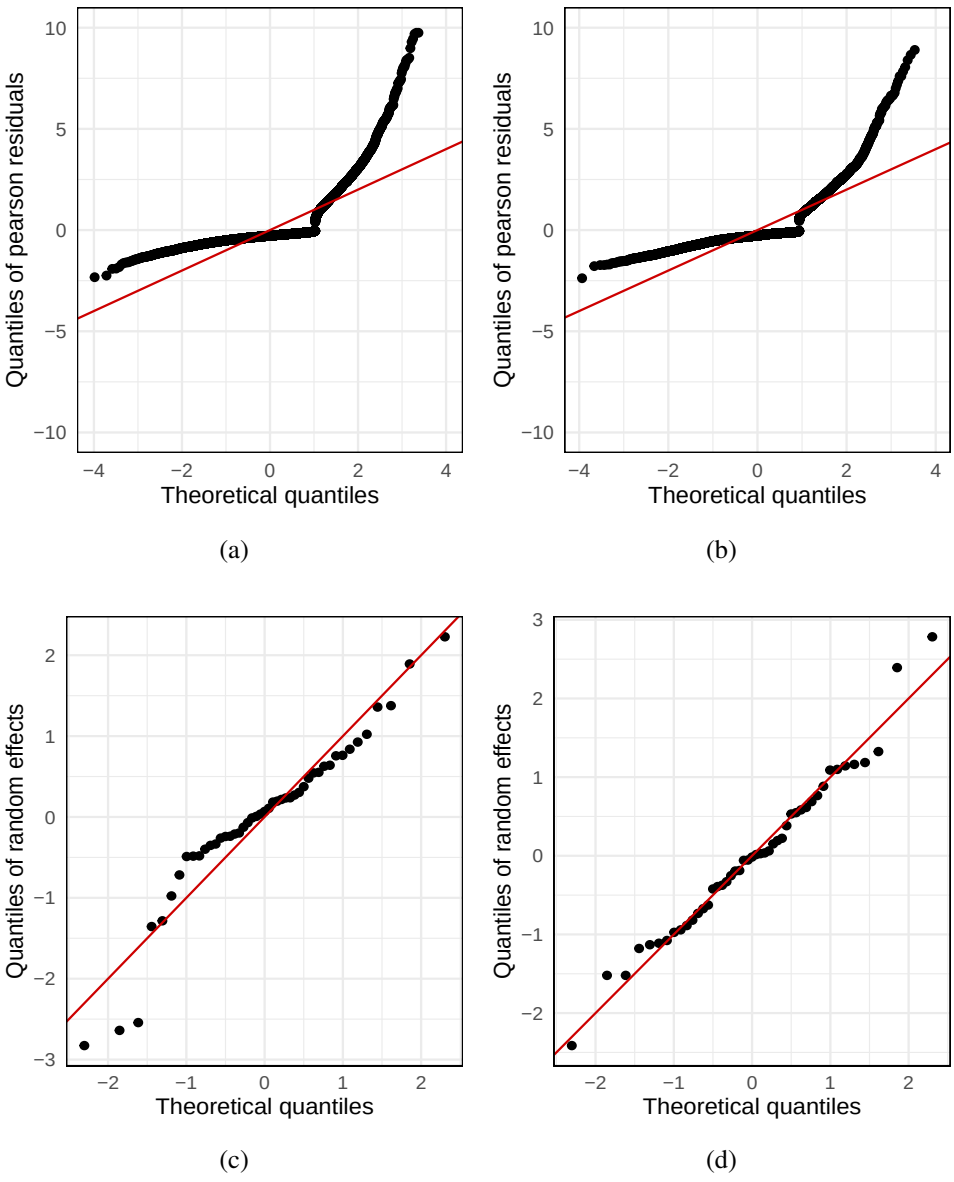


Figure 2.1: QQ-plots for Pearson residuals for (a) women (b) men and random effects for (c) women (d) men of health insurance coverage at the county level in Kenya.

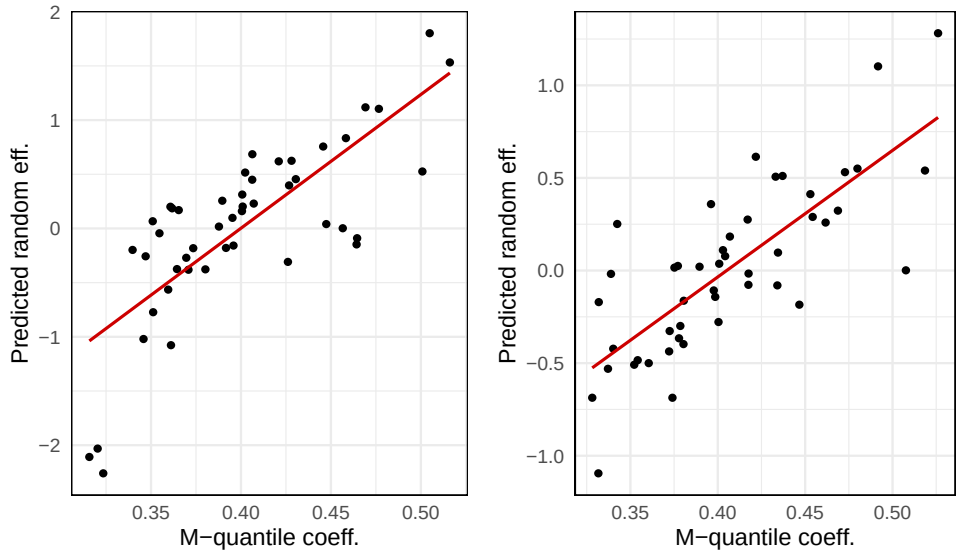
Table 2.7: Estimated fixed effects coefficients, variance components for the random effect and likelihood ratio test from fitting a generalized linear mixed model to health insurance data from Kenya.

	Women				Men			
	Estimate	Std. Error	z value	Pr(>|z|)	Estimate	Std. Error	z value	Pr(>|z|)
Intercept	-3.87664	0.37817	-10.25100	0.00000	-3.36832	0.29500	-11.41800	0.00000
Age	0.03095	0.00338	9.14900	0.00000	0.03077	0.00408	7.54800	0.00000
RelationToHead=2	-0.42258	0.06806	-6.20900	0.00000	0.47509	0.23885	1.98900	0.04669
RelationToHead=3	-0.57801	0.09281	-6.22800	0.00000	-0.47476	0.09738	-4.87500	0.00000
EmploymentStatus=1	-0.12982	0.05279	-2.45900	0.01393	0.17550	0.10854	1.61700	0.10588
EducationLevel=2	0.96049	0.06132	15.66400	0.00000	1.47206	0.05942	24.77400	0.00000
EducationLevel=3	1.63818	0.07047	23.24700	0.00000	-1.12156	0.26877	-4.17300	0.00003
Residence=1	0.73861	0.05487	13.46000	0.00000	-0.51425	0.05928	-8.67500	0.00000
MaritalStatus=2	-0.41803	0.09746	-4.28900	0.00002	-0.52183	0.09138	-5.71100	0.00000
MaritalStatus=3	-1.09058	0.15780	-6.91100	0.00000	-1.37671	0.54140	-2.54300	0.01099
MaritalStatus=4	-0.66043	0.19122	-3.45400	0.00055	-1.28689	0.31528	-4.08200	0.00004
MaritalStatus=5	-0.90623	0.13709	-6.61000	0.00000	-1.32604	0.20376	-6.50800	0.00000
Region=2	-1.08478	0.61113	-1.77500	0.07589	-0.37307	0.41679	-0.89500	0.37073
Region=3	0.60538	0.46117	1.31300	0.18928	0.55371	0.29280	1.89100	0.05861
Region=4	1.47144	0.51299	2.86800	0.00413	0.85915	0.32349	2.65600	0.00791
Region=5	0.54445	0.41713	1.30500	0.19182	0.66483	0.26522	2.50700	0.01219
Region=6	0.24644	0.55106	0.44700	0.65473	0.47969	0.34717	1.38200	0.16706
Region=7	0.98432	0.49054	2.00700	0.04479	0.56304	0.30997	1.81600	0.06930
Region=8	1.09669	0.90439	1.21300	0.22527	0.93010	0.55562	1.67400	0.09413
Variance component	0.6829	LRT = 1087.248	Pr(>χ^2)	0.00000	0.2448	LRT= 1111.015	Pr(>χ^2)	0.00000

Table 2.7 shows the estimated model parameters, standard errors and corresponding *p-values* for women and men. The fixed effects are age $(15 - 49)$ years, relationship to household head (1=head, 2=spouse, 3=others), employment status (0=unemployed, 1=employed), education level (1=completed primary, 2=secondary school and above, 3=no formal schooling), residence (0=rural, 1=urban), marital status (1=never married, 2=married, 3=widowed, 4=divorced, 5=seperated), region (1=Coast, 2=North Eastern, 3=Eastern, 4=Central, 5=Rift Valley, 6=Western, 7=Nyanza, 8=Nairobi). The regression coefficient of age has a positive sign (for both women and men), implying age increases the probability of access to health insurance. The table also shows the variance component for the random part of the model. To test whether the variance components are significant to measure unobserved heterogeneity we use the Likelihood Ratio Test (LRT). For women, the test statistic is 1,087.248, with a *p-value* of 0.0000, and for men equals 1,111.015, *p-value* = 0.0000. Therefore we reject the null hypothesis of no significance and conclude there is evidence of significant unobserved heterogeneity.

2.4.2 Binary M-quantile modeling

The diagnostic plots in section 2.4.1 show that the model assumptions of GLMM are not met. According to Chambers et al. (2016), GLMM's have attractive properties that can be used to model binary response variables. However, using GLMM's in SAE is not straightforward since the estimation of model parameters can be numerically demanding. Apart from computational complexity for using GLMM in small area estimation, if model are not met, invalid conclusions could be obtained. To reduce the adverse effects from deviations from distributional assumptions (provide robust inference), while at the same time borrowing strength from domains, we explore the use of M-quantile small area estimation model. Robust in this case means the estimator is reasonably efficient and unbiased, small deviations from model assumptions do not substantially affect the model performance and large deviations will not totally invalidate the model entirely. Since SAE via M-quantile uses M-quantile coefficients as opposed to random effects in GLMM, we find the correlation between the predicted area effects and area-level M-quantile coefficients. This was suggested by Chambers et al. (2016). The correlation equals 0.769 for women model and 0.77 for men. Figure 2.2 visualizes the scatter plots between the predicted random effects from mixed model and M-quantile model.



(a) (b)

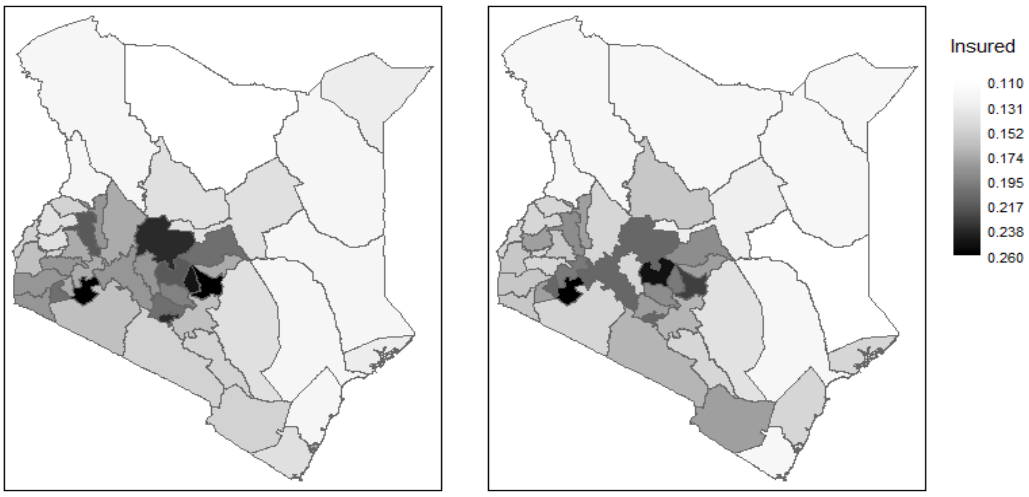
Figure 2.2: Scatter plots for the predicted random effects (estimated with GLMM) and the M-quantile coefficients (estimated with the M-quantile model) for (a) women (b) men at the county level in Kenya.

There is a high correlation between the area level M-quantile coefficients and the predicted area effects from GLMM to capture area variability. However, according to Chambers et al. (2016) M-quantiles provides an alternative SAE method when GLMM assumptions are not met. Table 2.8 shows quantiles of the point estimates of the proportion of persons with health insurance at the county level in Kenya. On average the mean and median for both women and men for direct and MQ estimates are comparable. A higher proportion of men are covered with health insurance compared to women. Overall, these proportions are quite low despite the efforts put by the government. This finding implies the government should explore other better options to increase coverage.

Table 2.8: Distribution of health insurance coverage proportions over counties in Kenya for women and men aged 15 - 49 years.

Gender	Estimator	Min.	Q1.	Median	Mean	Q3.	Max.
Women	Direct	0.0127	0.0891	0.1467	0.1518	0.1987	0.3654
	MQ	0.1033	0.1278	0.1478	0.1559	0.1724	0.2403
Men	Direct	0.0291	0.1001	0.1673	0.1731	0.2414	0.4193
	MQ	0.1042	0.1311	0.1548	0.1611	0.1804	0.2715

Figure 2.3 shows smooth maps of health insurance coverage for the 47 counties in Kenya using M-quantile estimation. The observed distribution is similar for women and men. Counties with highest coverage for women are Bomet (24%), Embu (23.6%), Kirinyaga (22.8%) and



(a) (b)

Figure 2.3: Maps showing the proportion covered with health insurance for (a) women and (b) men for the 47 counties in Kenya.

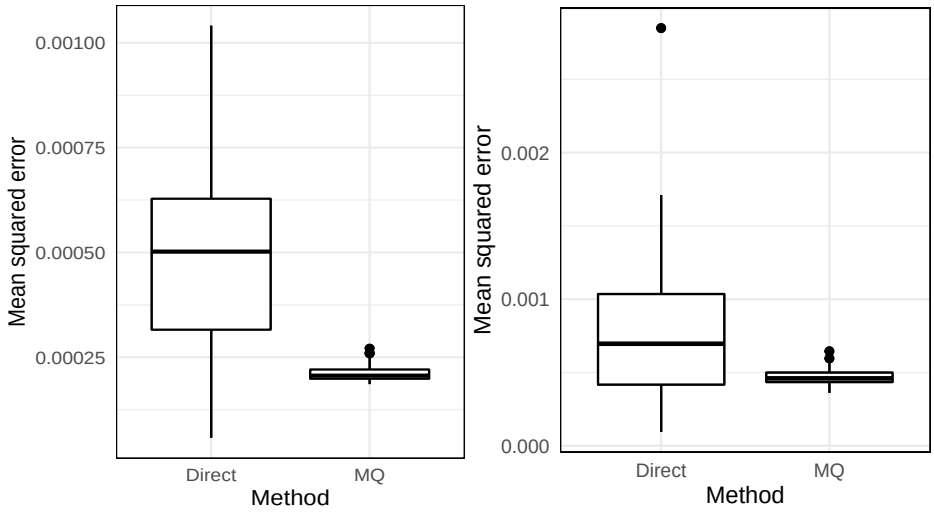
Nairobi (22%). The five least covered counties are West Pokot (10.9%), Turkana (10.8%), Garissa (10.7%) and Marsabit (10.3%). For men the leading counties are; Nairobi (27.2%), Bomet (24.8%), Nyeri (23.4%), Nyamira (22.2%) while the least covered counties are; Mandera (11%), Tana River (10.6%), Garissa (10.6%) and Kwale (10.4%). From the findings, we note that counties neighboring Nairobi have higher coverage rates. Since these counties are close to Nairobi which is the capital city of Kenya, with more employment opportunities, people living here are able to afford health insurance premiums. This is in contrast to counties further away like Turkana and Garissa. These results have been possible with the use of SAE methodology.

2.4.3 Evaluation of the M-quantile SAE model estimates

We evaluate the model-based results based on three criteria: (i) smaller MSE and CV for MQ compared to direct estimates, where the MSE is the sum of the variance and bias squared of the estimator, while the CV measures the dispersion of the estimates around the mean. (ii) consistency and (iii) usefulness to users. This has been proposed by Brown et al. (2001). The same approach has been used by Chandra et al. (2018) when estimating poverty incidence in the state of Bihar in India.

For MSE, Figure 2.4 (a) and (b) are the box plots of estimated MSEs of the estimated health insurance coverage for women and men. The MSE for both women and men are smaller for MQ compared to direct estimates. For CV's, Table 2.9 shows quantiles of the coefficient of

variation for the estimated health insurance percentages at county level in Kenya. For direct estimates especially for counties with small samples, the CV's reach values greater than 60% for both women and men. For MQ estimates all the CV's were less than 20. Since Kenya has not set a guideline for publishing official statistics, we use other statistical offices like the UK's Office for National Statistics (ONS). They set a CV of 20% for publishable official statistics. Therefore, our estimates meet this cut-off.



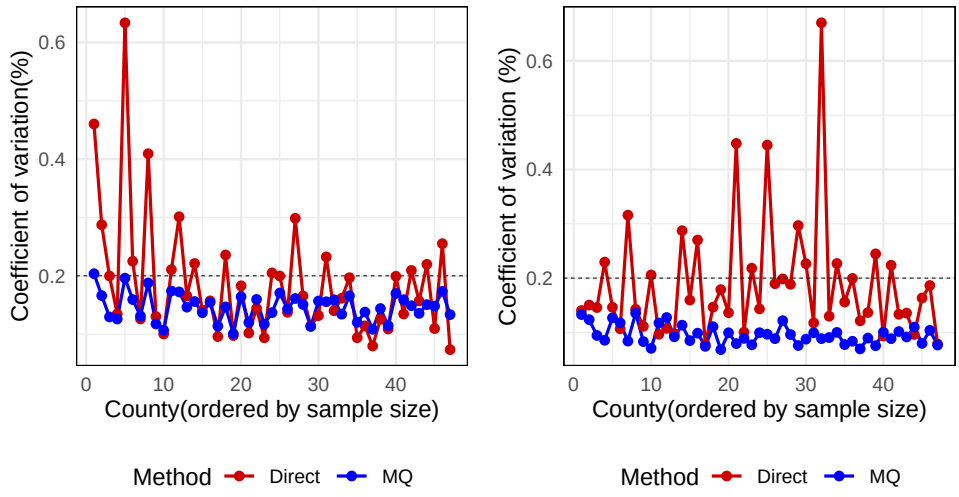
(a) (b)

Figure 2.4: Box-plots showing the mean squared error for the distribution of health insurance coverage percentages at the county in Kenya for (a) women (b) men

Table 2.9: Quantiles of the coefficient of variation for the estimated health insurance percentages at county level in Kenya.

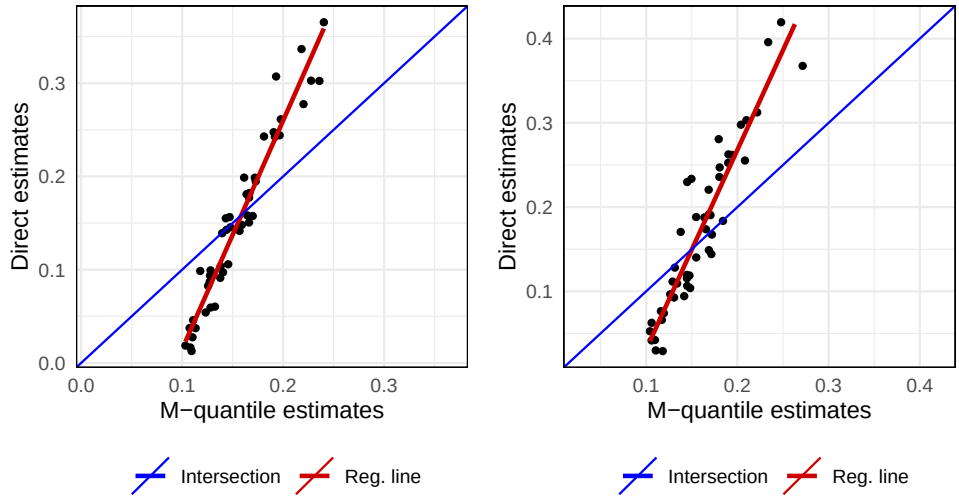
Gender	Estimator	Min.	Q1.	Median	Mean	Q3.	Max.
Female	Direct	0.073	0.120	0.157	0.186	0.215	0.634
	MQ	0.101	0.130	0.148	0.146	0.160	0.204
Male	Direct	0.078	0.126	0.150	0.188	0.221	0.670
	MQ	0.068	0.083	0.092	0.095	0.103	0.136

Figure 2.5 are line graphs with counties ordered by increasing sample sizes. To start with as expected the CV's for MQ estimates are smaller than direct estimates for all counties. As the sample sizes per county increase, the CV's for direct estimates reduce, especially for the women sample. By contrast, however, the CV's for direct estimates do not reduce with increasing sample size for men. For this analysis, the samples ranged from 236 to 460 for women and from 118 to 370 for men.



(a) (b)

Figure 2.5: Line graphs showing coefficient of variation of health insurance coverage with counties ordered by increasing sample sizes for: (a) women (b) men.



(a) (b)

Figure 2.6: Scatter plots with regression and intersection line ($y = x$) comparing direct and M-quantile health insurance coverage estimates for (a) women (b) men at the county level in Kenya.

For bias-diagnostics, Figure 2.6 is a scatter plot comparing direct estimates of the proportion of persons covered with health insurance and corresponding M-quantile estimates. According to Brown et al. (2001) the plots is based on the idea that, if the model-based estimates are "close" to the model-based SAE values of interest, then unbiased direct estimators should behave like random variables whose expected values correspond to the values of the model-based estimates.

That is, the model-based estimates should be unbiased predictors of the direct estimates. To check this, we plot appropriately scaled values of these estimates (x-axis) against similarly scaled direct estimates (y-axis) and then test whether the OLS (ordinary least squares) regression line fitted to these points is significantly different from the identity line. To check homoscedasticity assumption required for OLS fitting, we ran the Goldfeld-Quandt test. Under the null hypothesis, the Goldfeld-Quandt test statistic follows an F distribution with degrees of freedom as specified in the parameter. For men $(GQ = 0.39191, df1 = 22, df2 = 21, p - value = 0.983)$, and women $(GQ = 0.44733, df1 = 22, df2 = 21, p - value = 0.9662)$ we do not reject the null hypothesis and conclude heteroscedasticity is not present. Therefore, the homoscedasticity assumption is satisfied. We note that the M-quantile estimates are generally consistent with the direct estimates. Even though the model-based results depict some bias, the aggregated results in Table 2.10 are close.

Table 2.10: Aggregated direct and MQ estimates of health insurance coverage at the national level in Kenya.

Gender	Estimator	Proportion
Women	Direct	0.1546
	MQ	0.1573
Men	Direct	0.1796
	MQ	0.1628

For usefulness to users, we still adopt the criteria proposed by Brown et al. (2001). Accordingly, in SAE applications aggregation or bench-marking of small area estimates at higher level is always desirable by the users. In small area applications, National statistical offices involved in generating the small area estimates always expect that the small area estimates are aggregated/ bench-marked to higher level estimate. At higher level of aggregation, the direct estimates are considered to be reliable and therefore the model-based small area estimates are expected to be near to the direct estimates when they are aggregated. We checked the aggregation of model-based small area estimates at the county level. We computed national level insurance coverage by aggregating the direct estimates and MQ small area estimates, as $\sum_d (n_d.Direct_d)/\sum_d n_d$ and $\sum_d (n_d.MQ_d)/\sum_d n_d$, respectively. Table 2.10 shows the aggregated estimates. The MQ estimates aggregate well to national level direct estimate.

2.5 Concluding remarks

In conclusion, health insurance reduces health care costs by pooling resources. It is an important component towards achieving UHC. Assessing health insurance coverage for policy-making requires reliable data especially at disaggregated levels. In this study, we have combined survey and census data through an M-quantile model to get better estimates when sample sizes are small. This has the advantage that we avoid specifying random effects while providing robust inference against deviations from model assumptions. Findings show model-based estimates have smaller MSE's and CV's than direct estimates. Health insurance coverage remains low overall in Kenyan counties. Among those covered, our findings show inequality in health insurance coverage across the wealth quintiles with the highest coverage being the richer and richest, especially for private insurance which requires monthly contributions. Health insurance in Kenya is mostly voluntary except for public and civil servants. The majority of Kenyans also work in the informal sector where health insurance is not compulsory. With the current voluntary health insurance scheme, health insurance coverage remains low. Kenya should establish a mechanism mainly funded by taxation to extend prepaid coverage to its population. Despite financial constraints, Kenya should provide total subsidy to the poor through NHIF. Further, Kenya should give a partial insurance subsidy, through the NHIF to people within the informal sector. Two possible directions for further research are a) to allow for more disaggreaged domains like the sub-county level and b) to incorporate additional predictors of health insurance coverage like geospatial data. A limitation of this study is the time difference between the survey and census data. Data collected around the same time might yield to more accurate results. Despite the limitation, this study has estimated the health insurance coverage at the county level in Kenya with better precision compared to direct estimates. It has been possible to establish the variation in health insurance coverage between counties, noting that counties neighboring Nairobi have more proportions of persons with health insurance.

Funding

Financial support for this research has been provided by the Kenyan-German postgraduate programme in form of a scholarship.

Conflict of interest

The author declares no conflict of interest.

Chapter 3

Estimating county level overweight prevalence in Kenya using small area methodology

3.1 Introduction

Globally, the prevalence of overweight and obesity has increased more than three times between 1975 and 2016 (World Health Organization, 2021). Prevalence is the proportion of subjects with a specific characteristic in a population — in this case, the proportion of persons who are overweight. In 2016 the World Health Organization (WHO) estimated that 1.9 billion and 650 million adults were overweight and obese respectively (World Health Organization, 2021). The WHO defines overweight and obesity as abnormal or excess fat accumulation that present risk to human health. The Body Mass Index (BMI) is the basic and commonly used measure. It is a simple index used to classify overweight (BMI>25) and obesity (BMI>30) for adults. The BMI is a ratio of a person's weight in kilograms to the square of the height in meters (kg/m^2). According to the World Health Organization (2021), overweight is associated with increased risk for other non-communicable diseases (NCD's) such as type-2 diabetes and hypertension. Worldwide, the prevalence of overweight and obesity is higher for women than men (World Health Organization, 2021). A number of studies on maternal overweight such as Sebire et al. (2001), Kulie et al. (2011), Chowdhury et al. (2016) and Mkuu et al. (2018) have found that maternal overweight can affect both the mother and the unborn child. It can lead to higher rates of miscarriage, still-births and congenital anomalies. During pregnancy, overweight can later

affect the health for the mother and child including increased risk of heart disease, hypertension and diabetes.

Over the last decades, health challenges in low income and middle-income countries have revolved mainly on communicable diseases and under-nutrition (Pawloski et al., 2012). Sub-Saharan Africa harbour a large proportion of communicable diseases such as Malaria, HIV/AIDS and Tuberculosis (TB). However, due to urbanization and better incomes, a nutritional transition from health patterns associated with communicable diseases to health patterns associated with over-nutrition has occurred (Pawloski et al., 2012; Jones-Smith et al., 2012; Awuah et al., 2014; Steyn and Mchiza, 2014; Agyemang et al., 2014). Much attention and funding have gone into combating communicable diseases. With emerging NCDs coexisting with communicable diseases, this presents more challenges. For countries in sub-Saharan Africa, overweight and obesity present a tough challenge because persons who grow with under-nutrition, are prone to adding up more weight as they grow up. This is defined by WHO as malnutrition and is characterized by the coexistence of under-nutrition with overweight and obesity within individuals and households for a lifetime (World Health Organization, 2021).

In Kenya, more people move to towns and urban areas in search of jobs. This has the potential of improving their living standards from better income earned. However, a lot of time is spent working and reduced physical activity. They also have access to high-calorie fast foods within urban settings. Due to this among other factors, problems of increased body weight is on the rise (Kenya National Bureau of Statistics et al., 2015). The Ministry of Health Kenya (MOHK) notes that in addition to existing communicable diseases, this causes a double burden of disease in morbidity, mortality and medical expenses (Kenya National Bureau of Statistics et al., 2015). NCDs are a major public health concern with significant social and economic effects in terms of health care needs, loss in productivity, and premature death. They are a great setback to attaining the Sustainable Development Goals (SDGs) of the United Nations (UN) (General Assembly, 2015) if appropriate interventions are not implemented. Mkuu et al. (2018) using Kenya Demographic and Health Survey (KDHS) of 2014 found that 20.5% of the Kenyan women are overweight and 9.1% are obese. A study by Muthuri et al. (2014) on Kenyan school-going children established that out of 563 children, aged 9 to 11 years 3.7% were underweight, 14.4% were overweight, and 6.4% were obese. While Mbochi et al. (2012) on a cross-sectional study with 365 women aged 25 to 54 years in Nairobi, Kenya showed that BMI increased with age, greater socio-economic group, increased expenditure, increased parity

and more number of living rooms.

The Kenyan government is committed to improving the overall health of its citizens. To do this, data at disaggregated levels is required. However, this is lacking. Especially to ascertain the extent of the problem and identify the most affected groups and regions. Reliable data will also help to inform policy-making. The government of Kenya has come up with some policies and strategic plans such as the Kenya Health Policy (KHP), 2014-2030. The KHP outlines how Kenya seeks to improve the public health status in line with the Kenyan Constitution, Vision 2030 and SDGs (Kenya National Bureau of Statistics et al., 2015). Specifically, this policy was developed to respond to local and global development efforts to attain MDGs. It also targets NCDs, social determinants of health and the management of emerging and re-emerging health threats. Another strategy is the Kenya National Strategy for the Prevention and Control of Non-communicable Diseases, 2015-2020. The main objective is to reduce the preventable burden, avoidable death, sickness, risk factors and cost due to NCDs. To fulfill these goals, the Kenya National Bureau of Statistics (KNBS) carried out the inaugural survey on NCDs in 2015. This was a national cross-sectional household survey. It was designed to estimate indicators on risk factors for NCDs for persons aged 18 to 69 years at the national level. According to the survey, the common and important risk factors for NCDs are daily smoking, overweight or obesity, elevated blood pressure, low physical activity and a minimum of 400g of fruit and vegetables per day. Additionally, 28% of those sampled were either overweight or obese. Women (38%) were either overweight or obese as compared to 18% of men (Ministry of Health Kenya, 2014).

The KSSNDRF 2015 was a national survey. It was designed to provide reliable (design-based) estimates at the national level only. The design-based estimators (they rely only on the survey data) are approximately designed unbiased and consistent. However, direct estimators generally have large variances and estimates are unreliable when the sample sizes are small — for example at the county level in Kenya. In contrast, model-based small area methods produce more reliable estimates in terms of smaller MSE and coefficient of variation (CV) (Tzavidis et al., 2018). This is because they combine survey and census/administrative data through a model and therefore, increase the effective sample size. For more overviews on small area estimation(SAE) we refer the reader to (Rao and Molina, 2015; Pfeffermann, 2013).

For this study, therefore, we rely on SAE to better estimate the prevalence of overweight at the county level. To the best of our knowledge, this is the first study to use SAE and estimate the prevalence of overweight in Kenya. Our main data source is KSSNDRF 2015. The prevalence

estimates of overweight obtained from survey data only are called direct estimates hereafter in this paper. Initial analysis shows the coefficient of variation for the direct estimates reaches high values given the small sample sizes at the county level. We use an area level model proposed by Fay and Herriot (1979). Since the proportion of persons who are overweight in a particular county must lie between [0,1], we transform the dependent variable with the arcsine square root transformation. This non-linear transformation has been previously applied by Casas-Cordero et al. (2016) to estimate poverty in Chile, Schmid et al. (2017) to estimate literacy in Senegal and Hadam et al. (2020) to estimate regional unemployment in Germany. The estimates obtained are on a transformed scale. To make valid inferences we need to transform back to the original scale. Since bias is introduced due to transformation we use a bias-corrected back transformation. This is similar to the one used by Hadam et al. (2020). To assess the accuracy of our estimates we compute the (MSE) based on a parametric bootstrap that incorporates the additional uncertainty due to the bias-correction.

The rest of this paper is organized as follows. We describe the KSSNDRF 2015 and the Kenya Population and Housing Census (KPHC) 2009 in section 3.2. In section 3.3, we outline the small area methodology applied in this paper. In particular, the Fay-Herriot model, hereafter called the FH model, transformation, back transformation and MSE estimation. In section 3.4, we present the results of the application to estimate the prevalence of overweight in Kenya including model selection. Lastly, in section 3.5, we give the concluding remarks, possibilities for further research and limitation of this study.

3.2 Data sources: survey and census data

In this section, we describe the data sources used in this paper. We used the KSSNDRF 2015 and the KPHC 2009. The two datasets were provided by the KNBS under the Kenya National Data Archive (KeNADA) as public use files. Since the survey and census data were collected at different years, we assume the functional relation between overweight and auxiliary data remains constant.

3.2.1 Kenya STEPwise Survey for Non-communicable Diseases Risk Factors (KSSNDRF) 2015

The KSSNDRF 2015 adopted the WHO STEPwise approach to Surveillance (STEPS). This approach is a simple, flexible and standardized method for collecting, analyzing and disseminating data in countries that are members of WHO. Until 2016, 122 WHO member countries had completed data collection on STEPs survey (Riley et al., 2016). The WHO uses a tool called the STEPS Instrument to collect and measure NCDs risk factors. The tool covers three different NCDs risk factor assessment i.e. (i) A questionnaire (ii) Physical measurements and (iii) Biochemical measurements. The questionnaire gathers data on socio-demographic information, aspects of an individual's medical history related to the main NCDs, and risk behaviours. Physical measurements assess overweight and obesity and increased blood pressure while the biochemical measurements include blood and urine sampling to measure raised blood glucose, cholesterol and lipids (World Health Organization, 2005b).

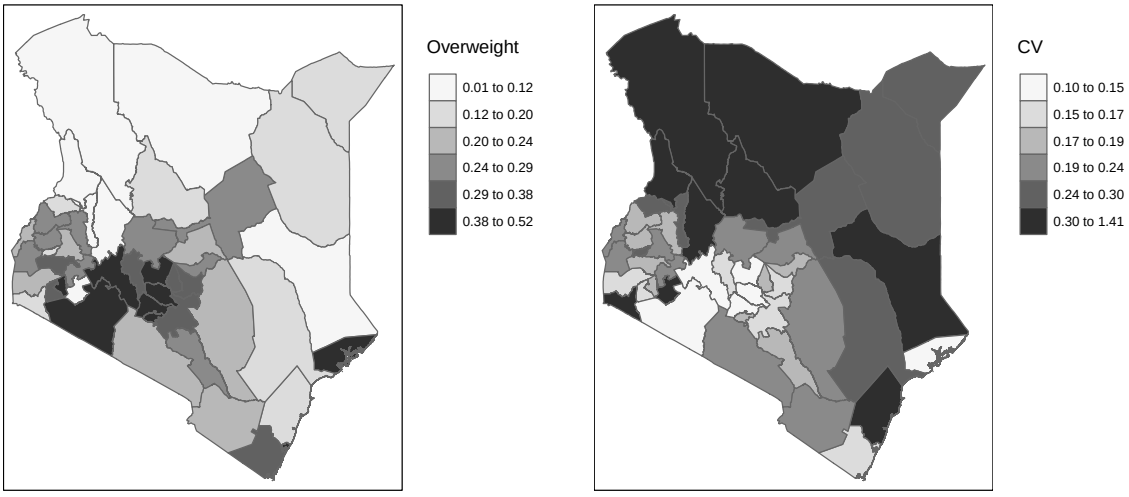
The STEPS Instrument allows each country to adapt and expand on the main variables and risk factors. Kenya adopted the STEPS approach in a sequential process consisting of three steps of information gathering. First, data on demographic and behaviour were collected. Demographic data included questions on age, sex, marital status, education and occupation. It also included questions on housing and social amenities. Questions on behaviour included tobacco use, alcohol consumption, diet, physical activity, history of blood pressure and diabetes, history of cardiovascular diseases, injury and oral health. The second step involved physical measurements on blood pressure, heart rate, height, weight, waist and hip circumference. This is to assess overweight and obesity. The last step collected data on biochemical measurements on blood glucose and blood lipids (Kenya National Bureau of Statistics et al., 2015).

The KSSNDRF 2015 was a national cross-sectional household survey designed to estimate indicators on risk factors for NCDs for persons aged 18 to 69 years. A sample size of 6,000 individuals was designed to give reliable estimates on the national level by sex (male and female) and residence (rural and urban). The survey used the fifth National Sample Surveys and Evaluation Programme (NASSEP V) maintained by the KNBS. The NASSEP V is a sample frame used for household surveys in Kenya and contains 5,360 clusters split into four subsamples. The KSSNDRF 2015 adopted a three-stage cluster sampling design which involves the selection of clusters, households and eligible individuals. First, 200 clusters (100 urban and 100 rural) were selected from one subsample of NASSEP V sample frame. Secondly, a

uniform sample of 30 households from the listed households in each cluster was selected. The last step involved randomly selecting one individual from all eligible household members using a programmed Kish selection method of sampling Kish (1949). iPAQ personal computer and personal digital assistants (PDAs) were used at this stage. Each interviewer was provided with an iPAQ together with its accessories and an extra battery. The PDAs automatically saved the data in their internal memory and also in a Secure Digital Card (SD card)(Kenya National Bureau of Statistics et al., 2015).

Currently, the KNBS officially reports the prevalence of overweight only on a national level where the survey is reliable. Apart from the KSSNDRF 2015, the other survey that collects and reports data on health related characteristics is the Kenya Demographic and Health Survey (KDHS). Kenya has conducted the KDHS in 1989, 1993, 1998, 2003, and 2008-09. Up to 2014, the previous KDHS has collected data on health characteristics in Kenya except for data on NCDs. The KDHS 2014, was also the first national survey to provide estimates for demographic and health indicators at the county level. However, the KDHS 2014 collected BMI for only women aged 15 to 49 years old. We selected the KSSNDRF 2015 since it collected BMI data on men and women — unrestricted to a particular age group.

Figure 3.1 presents direct estimates of overweight prevalence and coefficient of variation based on KSSNDRF 2015. Kenya has 47 counties which is the second administrative level after the national. For this survey, all the 47 counties were sampled. A total of 4,500 individuals were successfully interviewed at the primary stage sampling giving a response rate of 95%. We had access to a total sample size of 4,288 of which 4,014 are complete cases.



(a) (b)

Figure 3.1: Maps showing: (a) Direct point estimates of overweight prevalence and (b) the corresponding coefficient of variation based on KSSNDRF only.

Table 3.1 shows summary statistics of sample sizes, direct estimates and the corresponding CVs of overweight over counties. The minimum and maximum CV are 10% and 141% respectively. Currently, there is no internationally accepted cutoff point for CV's to report official statistics. Further, Kenya hasn't set a threshold based on CVs for reporting official statistics. Therefore, we follow the guidelines of other statistics offices, for instance, the Office for National Statistics (ONS) in the UK uses a CV of 20% as a threshold for publishing official results. Based on this 23 domains out of 47 have CV's greater than this threshold.

Table 3.1: Summary statistics of sample sizes, overweight point estimates and respective coefficient of variation over the 47 counties in Kenya.

	Min.	Q1.	Median	Mean	Q3.	Max.
Sample size	53.000	75.000	84.000	85.000	95.000	152.00
Direct estimates	0.0090	0.1601	0.2378	0.2447	0.3224	0.5199
CV	0.1003	0.1611	0.1918	0.2409	0.2818	1.4121

3.2.2 The Kenya Population and Housing Census 2009

The first comprehensive census in Kenya was done in 1948. The next was in 1962 with 8.6 million people. The census helped in setting up political and administrative structures. After, independence in 1969, a third census was conducted with 10.6 million people. Since then, Kenya has continuously conducted a census after every 10 years i.e. 1969, 1979, and so on, the most recent being 2019. The meta-data for 2019 has not been released for public use. The KNBS

under the Statistics Act 2006 of the Kenyan laws is the main government agency responsible for collecting, analyzing and disseminating census and other statistical data.

Census is a large statistical undertaking and requires huge finances, planning and personnel. The implementation of the 2009 KPHC for the Republic of Kenya (RoK) used an estimated 8.4 billion Kenyan shillings (appr. 9 million US dollars) (Kenya National Bureau of Statistics, 2010). This huge investment is justified as it is a key exercise for the government of Kenya and interested stakeholders. The statistical information is required for monitoring the implementation of various development objectives and global initiatives e.g. the United Nations Millennium Development Goals (UNMDGs). It serves as a basis for adequate policy planning. Data on fertility and mortality are important in dispatching services related to births and deaths. The country's growth rate can also be accessed. To provide social amenities to different age groups, the census provides data on the composition of a country's population by age. Minority and age groups who require special amenities are identified through this data. To determine tax relief, data on the dependency ratio is needed. Persons are born in different places and move from one place to another. Data on migration is important to understand migration trends and required interventions.

For this study, we had access to the 2009 KPHC. This was the 5th census after independence. It was done from the night of 24 and 25 to 31 August 2009. The main objective was to provide key information on the demographic, social and economic characteristics of the population and housing. These include size and composition of the population, fertility, mortality and migration rates, levels of education, size of labour force, e.t.c. In this census, data was captured through scanning technology with technical assistance provided by the United States Census Bureau (USCB) (Kenya National Bureau of Statistics, 2010). This census was based on old administrative areas i.e. villages, sub-locations, locations, divisions, districts and provinces. There were 46 legal districts in Kenya excluding Nairobi — the capital city which constituted the 47th district. These districts were converted to the current 47 counties without change of boundaries after 2010 (Government of Kenya, 2013). Therefore, we can link the survey and census data. The variables in the census serve as potential covariates in the small area model introduced in section 3.3 for predicting overweight in Kenya. At this point, we state that the response variable — overweight is unreported in the census.

3.3 Small area estimation methodology

In this section, we outline the SAE method. First, in section 3.3.1 we describe the standard FH model. Since the FH model does not guarantee that the prevalence of overweight lies in the interval $[0, 1]$ we describe an arcsine square root transformed FH model in section 3.3.2. For the standard FH and arcsine square root transformed FH models, we explain the estimation of regression parameters, sampling variances, random effects and the MSE.

3.3.1 The Fay-Herriot model

We assume that a finite population of size N which is divided into m disjoint areas of sizes $N_1, N_2, ..., N_m$ where $i = 1, 2, ..., m$ is the ith small area. A sample of size n is taken from this population using a complex sampling design with sample sizes $n_1, n_2, ..., n_m$ for each area i. Further, we assume the response variable y_{ij} of individual j in area i has been measured without error in the survey. In this paper, we are interested in estimating the mean prevalence of overweight in Kenya with reduced uncertainty by incorporating extra covariates from census data. We consider the area level FH model (Fay and Herriot, 1979) where the direct estimator of the population mean is given as

$$\hat{\bar{y}}_i^{\text{dir}} = \frac{1}{n_i} \sum_{j=1}^{n_i} w_{ij} \cdot y_{ij},\ i = 1, 2, ..., m, \tag{3.1}$$

where $\hat{\bar{y}}_i^{\text{dir}}$ is the direct mean estimator for area i, w_{ij} are sampling weights. The weights compensate for unequal probabilities of sampling and unit non-response. This is the Horvitz-Thompson (HT) estimator of Horvitz and Thompson (1952) for estimating population means and totals. A big advantage of the FH model is the ability to take into account the sampling design using the HT estimator (Särndal et al., 2003). The first stage of the FH model is a function of the direct estimator in Equation 3.1 above and the sampling errors as

$$\hat{\bar{y}}_i^{\text{dir}} = \theta_i + \varepsilon_i, \tag{3.2}$$

where θ_i is the population mean and ε_i is the sampling error assumed to be normally distributed and independent i.e. $\varepsilon_i \sim N\left(0, \sigma^2_{\varepsilon_i}\right)$. In theory the sampling errors ε_i for the FH model, are assumed known. However, in practice this is estimated. In section 3.3.2 we outline how we estimated the sampling variance for this application. In the second stage θ_i is linked to available

area level covariates

$$\theta_i = x_i^{tr}\beta + v_i, \tag{3.3}$$

where x_i are area level auxiliary variables, β is vector of regression parameters and v_i are area level random effects. The random effects are assumed to be independently normally distributed i.e. $v_i \sim N\left(0, \sigma_v^2\right)$. Combining the sampling model 3.2 and linking model 3.3 we obtain the area level model given by

$$\hat{\bar{y}}_i^{\text{dir}} = x_i^{tr}\beta + v_i + \varepsilon_i, \tag{3.4}$$

which is a linear mixed model with $E\left[v_i\right] = E\left[\varepsilon_i\right] = 0$. According to Rao and Molina (2015), $\hat{\beta}$ can be estimated as the best linear unbiased estimator (BLUE) of β and the random effect $\hat{v}_i$ as the empirical best linear unbiased predictor (EBLUP) of v_i (Henderson, 1975). The variance σ_v^2 can be estimated by the Maximum Likelihood Method (ML) or the Residual Maximum Likelihood Method (REML) (Hartley and Rao, 1967; Patterson and Thompson, 1971; Datta and Lahiri, 2000). Under this combined model, the EBLUP is obtained as

$$\begin{aligned} \hat{\bar{y}}_i^{\text{FH}} &= x_i^{tr}\hat{\beta} + \hat{v}_i, \\ &= \hat{\gamma}_i \hat{\bar{y}}_i^{\text{dir}} + \left(1 - \hat{\gamma}_i\right) x_i^{tr}\hat{\beta}, \end{aligned} \tag{3.5}$$

where $\hat{\gamma}_i$ is the shrinkage factor for area i given by $\hat{\gamma}_i = \frac{\sigma_v^2}{\sigma_v^2 + \sigma_{\hat{\varepsilon}_i}^2}$.This EBLUP is a weighted combination of the direct estimator $\left(\hat{\bar{y}}_i^{\text{dir}}\right)$ and the synthetic estimator $\left(x_i^{tr}\hat{\beta}\right)$. In practical applications, many small areas have zero sample sizes and the direct estimator is unavailable, therefore we depend on the synthetic estimator (Rao and Molina, 2015). According to Pfeffermann (2013) and Rao and Molina (2015), when small area estimates are produced, they should be accompanied by a valid measure of precision. The mean squared error (MSE) is still the standard measure of uncertainty in official small area statistics. We, therefore, determine the accuracy of our EBLUP by calculating the MSE. As stated by Rao and Molina (2015), this MSE can be obtained based on the method used to estimate the variance of the random effect. Following Prasad and Rao (1990) and Datta and Lahiri (2000), an analytical MSE is obtained if the method chosen is ML or REML.

3.3.2 The arcsine square root transformed FH model

The prevalence of overweight is a proportion and must lie on the interval of $[0, 1]$. However, the FH model outlined above can give estimates outside this range. Secondly, the FH model is a linear mixed model(LMM) in which some assumptions are made. Specifically, normality, linearity and homoscedasticity of error variance. To meet the condition of $[0, 1]$ interval and assumptions for the LMM, we, therefore, transform the vector of direct estimators. As in Schmid et al. (2017) and Hadam et al. (2020) we use the arcsine square root transformation. This transformation also stabilizes the variance (Carter and Rolph, 1974; Efron and Morris, 1975). In this case, we transform only the response variable — vector of direct estimators. Both-sides transformation for linearity can make the error term heteroscedastic (Carroll and Ruppert, 1988). Since we have chosen our model a prior, we assume it fits the data adequately. We first define the function $g(z) = \frac{1}{sin}(\sqrt{z})$. Equation 3.4 above becomes

$$\frac{1}{sin}\left(\sqrt{\hat{\bar{y}}_i^{\text{dir}}}\right) = x_i^{tr}\beta + v_i + \varepsilon_i, \tag{3.6}$$

where $\varepsilon_i \sim N\left(0, \tilde{\sigma}^2_{\varepsilon_i}\right)$ and $v_i \sim N\left(0, \sigma_v^2\right)$. As mentioned in 3.3.1, in theory the sampling variances are assumed known, but estimated in practice (Rao and Molina, 2015). In this study, we estimate the sampling variance directly from the sample data. The sampling variances can be unstable especially for small sample sizes (Bell, 2008; Hawala and Lahiri, 2010). Generalized Variance Functions (GVF) have been used to smooth the sampling variance (Maples et al., 2009; Hawala and Lahiri, 2010; Pratesi, 2016; Hawala and Lahiri, 2018). In this paper, we adopt a similar approach as used in Pratesi (2016). The variance smoothing model is given by $\frac{\hat{p_i}(1-\hat{p_i})}{var(p_i)} = \beta \cdot n_i + e_i$, where $e_i \sim D(0, \tau^2)$, $\hat{p_i}$ is the prevalence in area i, n_i is the sample size in area i and β is a linear regression coefficient.

For the arcsine square root transformed model, we estimate the sampling variance as in Jiang et al. (2001), Schmid et al. (2017) and Hadam et al. (2020) by $\tilde{\sigma}^2_{\varepsilon_i} = (4n_i^*)^{-1}$ where n_i^* is the effective sample size in area i. The effective sample size is estimated by $n_i^* = \frac{n}{deff}$ where n is sample size and $deff$ is the design effect. The design effect is the ratio of the variance of the direct estimator under simple random sampling to the variance under the complex sampling design of the survey (Särndal et al., 2003). Figure 3.2 show a plot of the direct variances against GVF smoothed variances.The graph show that the direct variances follow a similar pattern as the GVF smoothed variances. However, the later show a smooth behavior. From Equation 3.6

the parameters β and the random effects v_i are estimated as in the standard FH model described in section 3.3.1. The arcsine square root transformed FH model is obtained by replacing these parameters with their respective estimates yielding

$$\hat{\bar{y}}_i^{\text{FH, trans}} = \hat{\gamma}_i \left(\frac{1}{sin} \sqrt{\hat{\bar{y}}_i^{\text{dir}}} \right) + (1 - \hat{\gamma}_i)\, x_i^{tr} \hat{\beta}. \tag{3.7}$$

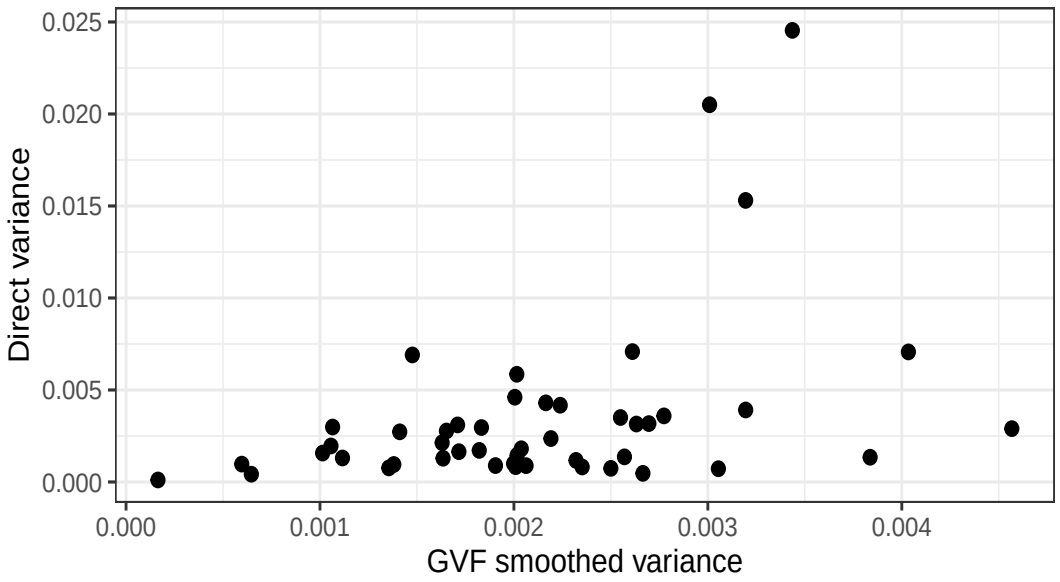


Figure 3.2: A scatter plot of direct and GVF smoothed sampling variance.

With Equation 3.7 we obtain the prevalence of overweight on a transformed scale. To obtain estimates on the original scale, we transform them back to the original scale. Mathematically, one would use $z = sin^2(z)$. This kind of back transformation is naive as it introduces bias due to the non-linear transformation. Some studies that have used this transformation include Casas-Cordero et al. (2016) and Schmid et al. (2017). In this paper, we adopt a bias-corrected back transformation as proposed in Hadam et al. (2020). In their paper, if for the transformed FH-model the assumptions are fulfilled, the transformed FH estimator is normally distributed with $a \sim N(a,\ b)$, where $a = \hat{\bar{y}}_i^{\text{FH, trans}}$ and $b = \frac{\hat{\sigma}_v^2 \tilde{\sigma}_{\varepsilon_i}^2}{\hat{\sigma}_v^2 + \tilde{\sigma}_{\varepsilon_i}^2}$. The bias-corrected back transformed FH estimator $\hat{\bar{y}}_i^{\text{FH, back}}$ is obtained by computing the expected value of the naive back transformation under the assumed normal distribution. This integral can be solved using numerical integration techniques. Through simulation studies, they show a reduction in bias due to this correction. To estimate the MSE of the bias-corrected back transformed FH estimates, we also adopt a parametric bootstrap method used in Hadam et al. (2020). They present a procedure for

estimating confidence intervals and MSE for an arcsine square root transformed bias-corrected FH estimator. Through simulation studies, they show this bootstrap shows a good performance for MSE estimation and confidence intervals.

3.4 Application: estimating the prevalence of overweight in Kenya

In this section, we apply the small area method presented in section 3.2 to estimate the prevalence of overweight in Kenya. We implement this in the R package emdi Kreutzmann et al. (2019a). From this section hereafter, the estimates obtained from this methodology will be referred to as FH_trans or FH estimates or simply model-based estimates.

3.4.1 Model selection and diagnostics

The model in section 3.2 requires aggregated auxiliary data. For this study, we had access to census data. According to Rao and Molina (2015), a key requirement for the success of small area methods is the availability of good useful auxiliary data from census or administrative records. Some studies such as Schmid et al. (2017) and Hadam et al. (2020) have used auxiliary data from mobile phone data as an alternative. Based on Mkuu et al. (2018), Mbochi et al. (2012), Asiki et al. (2018), we first selected likely predictors of overweight from census data. We then fit a full model with all the covariates. Lastly, we select predictive covariates using the Akaike Information Criterion (AIC) for the FH model. The STEPwise procedure we used involved forward and backward selection. The final model had an adjusted R^2 of 59%. The selected covariates are: age, gender with categories male and female, education with categories; no formal schooling, completed primary school, secondary school and above;marital status with categories; never married, married, widowed and divorced; employment with categories government employed, self-employed, unemployed, employed by NGO and others.

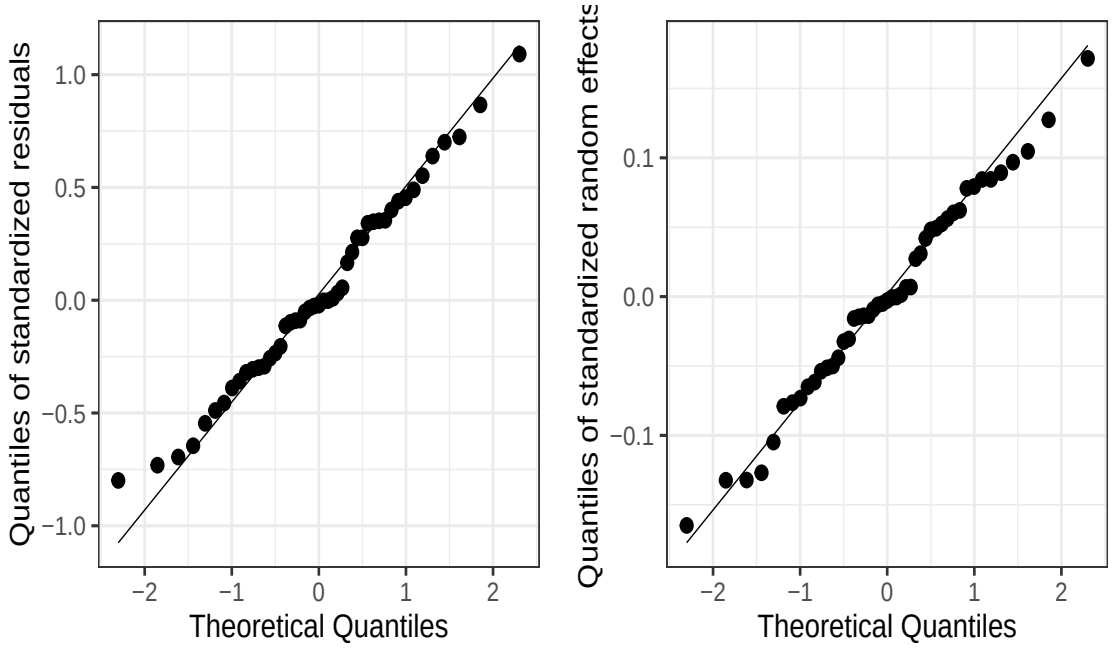
Table 3.2 are significant predictors of overweight at $\alpha = 0.05$. It also shows the corresponding standard errors, t-values and p-values. We note that age, marital status and employment have positive coefficients while gender, level of education and household size have negative coefficients. Our results are in agreement with other studies on predictors of overweight. To mention a few, Groenveld-van Dijk (2013) found that gender, age, education, wealth and ethnicity are highly correlated with the prevalence of overweight and obesity in Kenya. A study by Mbochi (2010), showed that age, parity, socio-economic status and physical activity are all

Table 3.2: Significant predictors of overweight in Kenya and corresponding coefficients, standard errors, t-values and p-values.

Parameter	Estimate	SE	t-value	p-value
Intercept	-0.2763	0.6716	-0.4114	0.6808
Age	0.0372	0.0149	2.4981	0.0125
Female	-2.3412	0.8679	-2.6974	0.0070
Married	1.6591	0.5294	3.1339	0.0017
Secondary	-0.2512	0.0967	-2.5969	0.0094
Unemployed	6.9896	2.4552	2.8469	0.0044
Household size	-0.0508	0.0230	-2.2049	0.0275

The estimated random effects variance $\hat{\sigma}_v^2 = 0.00817$.

significant predictors of overweight and obesity in Kenya. Muthuri et al. (2014) on a study of Kenyan schoolchildren found that parent's education level, income, and type of school attended — either private or public were positively associated with overweight/obesity. Mkuu et al. (2018) and Christensen et al. (2008) found urbanization to be significant in predicting the prevalence of overweight and obesity in Kenya. Since physical inactivity has been shown to significantly predict overweight, Gichu et al. (2018) established that gender, age, education level and wealth index significantly predict physical inactivity.

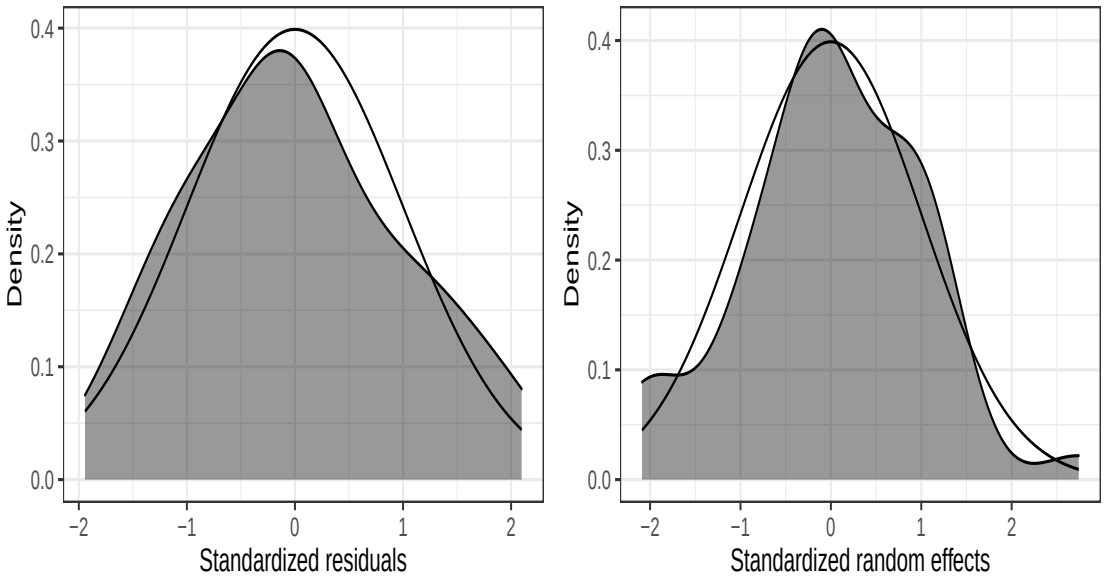


(a) (b)

Figure 3.3: (a) Quantiles of standardized residuals and (b) Standardized random effects for the arcsine square root transformed model with GVF smoothed variance.

Figure 3.3, shows QQ-plots of standardized residuals (a) and standardized random effects (b) for assessing normality assumptions in the sampling and linking models. The residuals and

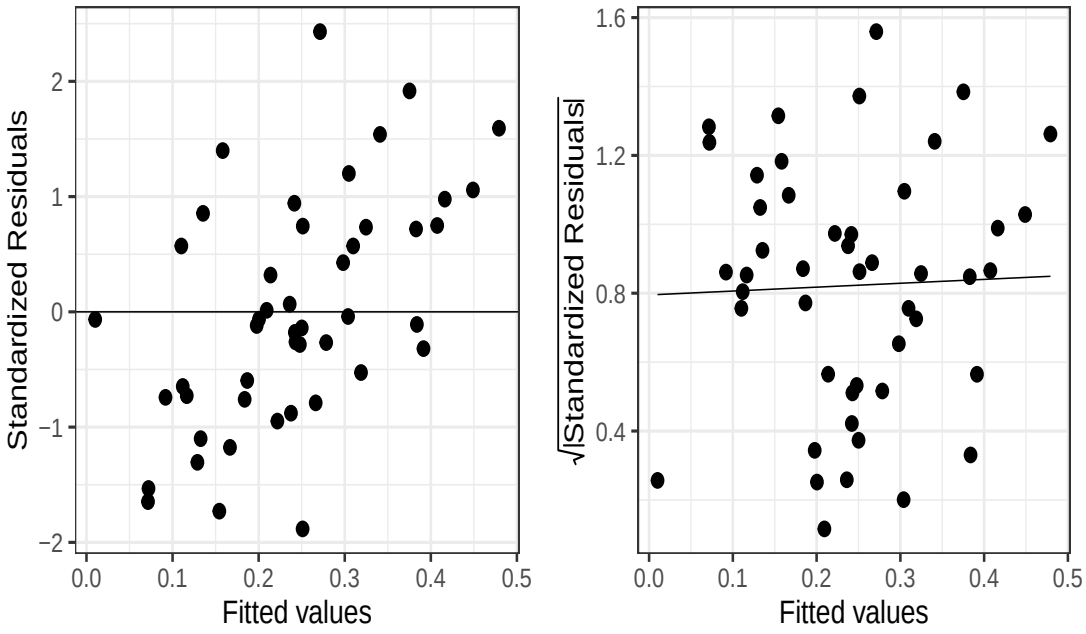
random effects lie in the QQ-line with only a few deviating, especially at the tails. Therefore, it is reasonable to assume the approximate normality of error terms. To further confirm this, we present density plots in Figure 3.4. Lastly, the Shapiro-Wilk test for standardized residuals equals 0.98552 with a p-value of $0.8213 > 0.05$. Therefore, we fail to reject the null hypothesis and conclude the distribution of the data is not significantly different from the normal distribution.



(a) (b)

Figure 3.4: Density plots for: (a) standardized residuals and (b) standardized random effects for the arcsine square root transformed model with GVF smoothed variance.

Figure 3.5 is a plot of standardized residuals versus fitted values(a) and scale-location plot(b). This is to test for linearity and homoscedasticity assumption in the model. The residuals are randomly distributed around $y = 0$ with no apparent pattern. They form an approximate horizontal band around the zero line. Therefore, linearity can be assumed. The scale-location plot shows the smooth line is roughly horizontal across the plot. There is also no clear pattern among the residuals. Therefore, the homoscedasticity assumption is met. Further, to test for randomness in this pattern we use the runs test. The p-value is 0.4588 which is greater than $\alpha = 0.05$ and we fail to reject the null hypothesis of randomness. Therefore, we have sufficient evidence to say that the residuals are random.



(a) (b)

Figure 3.5: Scatter plots for: (a) standardized residuals verses fitted values and (b) scale-location plot for the transformed Fay-Herriot model.

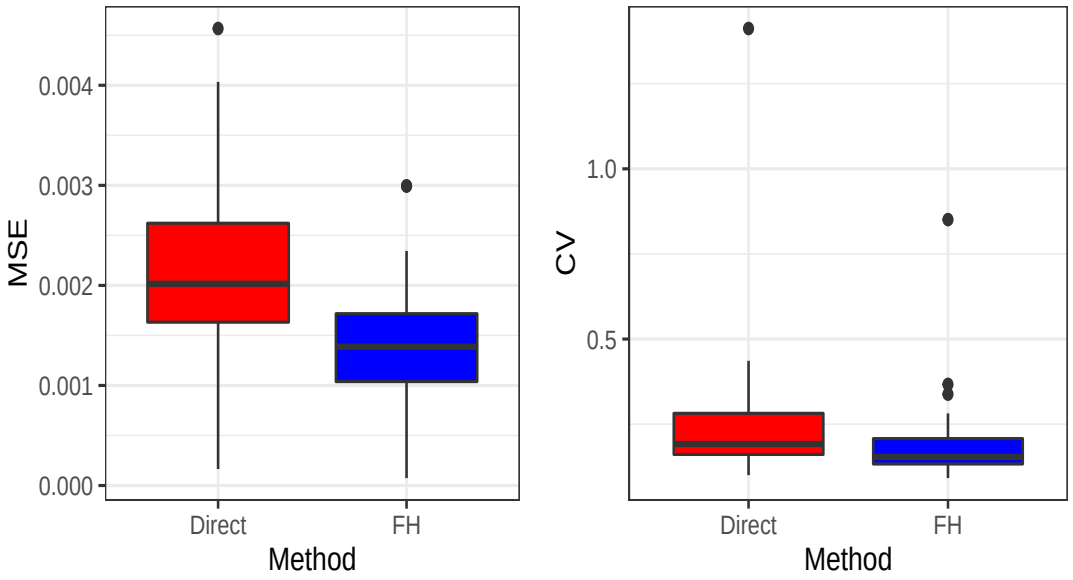
3.4.2 Diagnostics for model-based small area estimates

In addition to model diagnostics in subsection 3.4.1 we present diagnostics for small area model-based estimates. We assess the reliability and validity of the FH estimates. We follow closely the guidelines of Brown et al. (2001). As noted by Chandra et al. (2018), the FH estimates should; (i) be more precise than direct estimates. (ii) be consistent with the unbiased direct estimates. (iii) give useful results to users. To assess for precision we use the MSE and CV. Table 3.3 below shows the summary statistics for the point estimates, the MSE and CV.

Table 3.3: Summary statistics for the point estimates, mean squared error and coefficient of variation for the county level overweight prevalence in Kenya.

		Min.	Q1.	Median	Mean	Q3.	Max.
Point Est.	Direct	0.00908	0.16018	0.23785	0.24478	0.32245	0.51998
	FH	0.01013	0.16247	0.24208	0.24206	0.30730	0.47918
MSE	Direct	0.00016	0.00163	0.00201	0.00212	0.00262	0.00456
	FH	0.00004	0.00102	0.00133	0.00143	0.00177	0.00386
CV	Direct	0.10030	0.16110	0.19180	0.24090	0.28180	1.41210
	FH	0.09118	0.12692	0.15030	0.18295	0.20557	0.65955

From Table 3.3, the direct estimates range between 0.009088 and 0.51998 while the FH estimates lie between 0.01013 and 0.47918. Therefore, the FH have a smaller range compared to the direct estimates. An important advantage of SAE is the shrinkage of direct estimates towards the regression estimates from additional auxiliary data (Datta et al., 2012). The summary statistics for the MSE and CV shows the accuracy gained in using the small area method outlined in this paper. There is a reduction in MSE and CV for the FH estimates. For instance, the maximum MSE for the direct and FH are 0.00456 and 0.00386 respectively. We also note the FH estimates are shrunk such that the maximum CV reduced significantly from 141% to 65% for the direct and FH estimates respectively. The gain in accuracy is further shown in the boxplots for MSE and CV in Figure 3.6.

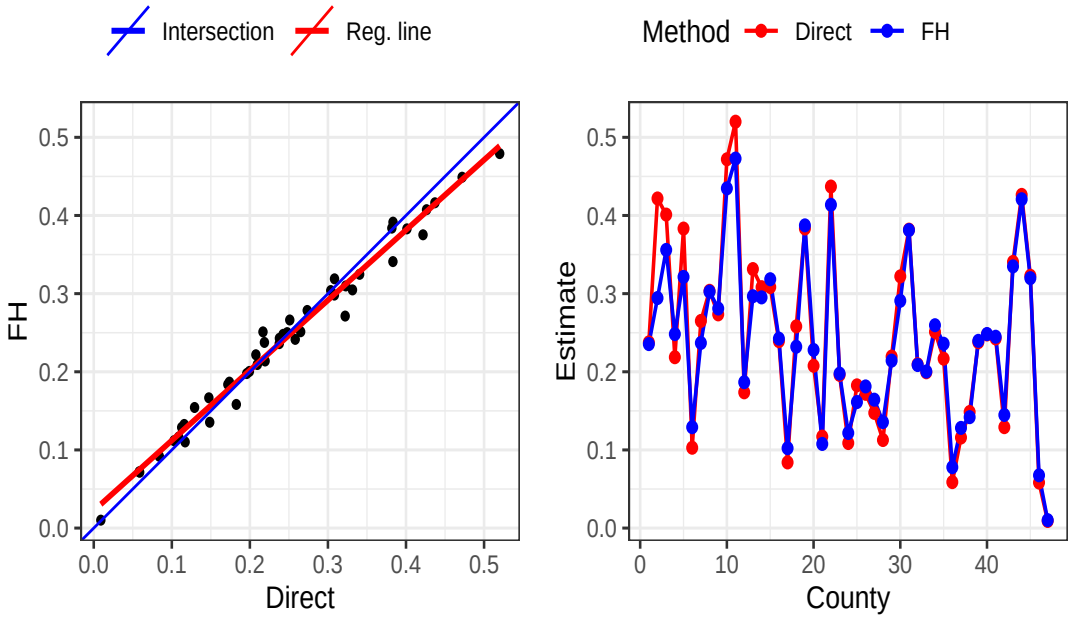


(a) (b)

Figure 3.6: Box plots showing: (a) mean squared errors and (b) coefficient of variation for direct and FH estimates of overweight prevalence.

For bias diagnostics as outlined by Brown et al. (2001) we first plot the direct $(x - axis)$ and FH estimates $(y - axis)$. Figure 3.7 (a) shows a scatter plot of fitted regression line and the identity line ($y = x$) i.e. $\beta_0 = 1$ and $\beta_1 = 0$. The regression line is fitted by least squares. As stated by Chandra et al. (2018), if the direct estimates are unbiased, then regressing them on the true values should be linear. And correspond to the line $y = x$. Therefore the scatter plot would be evenly distributed around the identity line. In our case $\beta_0 = 0.898$ and $\beta_1 = 0.02236$, implying the FH estimates are approximately design unbiased. Further, Brown et al. (2001)

provide a test to compare estimators. The test computes the correlation between the regression synthetic part of the model and direct estimates. Using the Brown test we fail to reject the null hypothesis that the FH estimates are statistically significantly different from the direct estimates at α =0.05. The line graph in Figure 3.7 (b) is a comparison of direct and FH estimates where the counties are ordered by decreasing the MSE of the direct estimator. We note as the MSE of the direct estimator reduces, the direct estimates approach the FH estimates. This is a gain as according to Rao and Molina (2015), the main reason for using model-based estimators is the reduction in MSE.



(a) (b)

Figure 3.7: A plot of direct and FH estimates with (a) regression and intersection line ($y = x$) and (b) a line graph showing overweight prevalence point estimates — direct and FH and counties where counties are ordered by decreasing MSE of the direct estimator.

Figure 3.8 is a line graph of CV's with counties ordered by decreasing sample sizes. For all counties, the CV for FH estimates is smaller than those of the direct estimator. As the sample size reduces per county, the difference in CV between the direct and FH estimates increases. This partly explains the need for using FH estimates for areas with small sample sizes.

3.4.3 Distribution of overweight prevalence in Kenya

We found the national overweight prevalence to be 27.98%. The KNBS reported 28% of Kenyans are overweight (Kenya National Bureau of Statistics et al., 2015). Figure 3.9, are two maps showing the county level distribution of overweight prevalence in Kenya (a) and corresponding

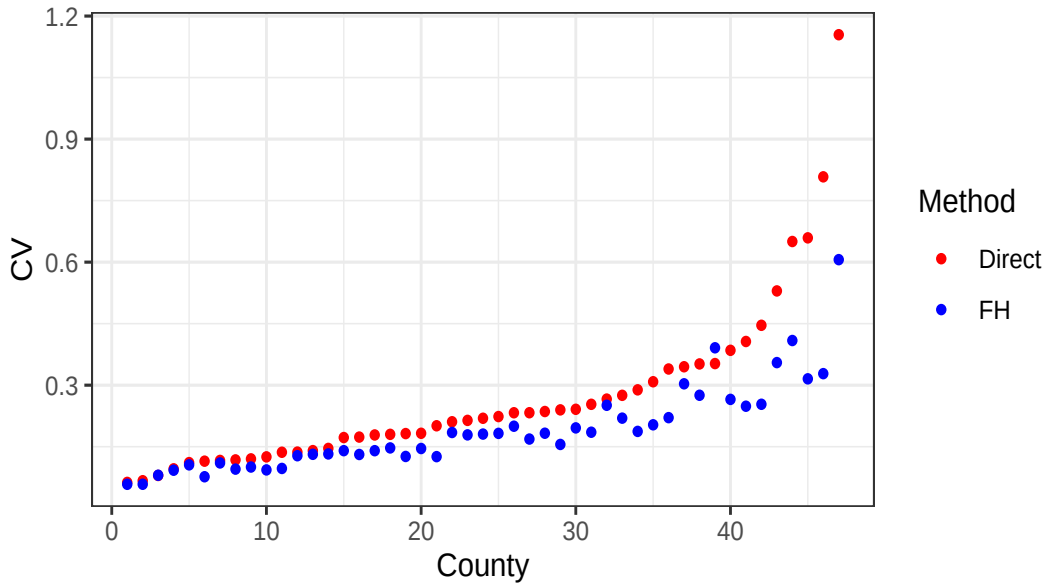
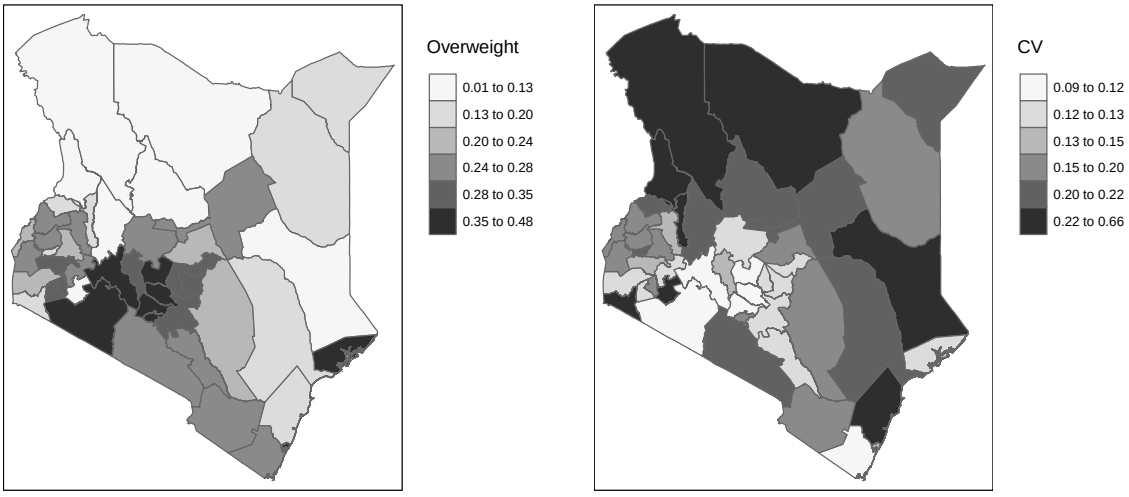


Figure 3.8: A line graph showing coefficients of variation for direct and FH estimates and counties ordered by decreasing sample size.

CV's (b). The CV's indicate the sampling variability as a percentage of FH estimates. The map is essential in identifying regions with high and low overweight prevalence. It shows that the prevalence of overweight varies geographically. Kenya is administratively divided into 47 counties, 290 sub-counties, and 1450 wards. These administrative units are clearly defined by geographical boundaries. Therefore, for this study $m = 1, 2, ..., 47$. The counties Nairobi (39%), Kiambu (45%), Nakuru (41%), Murang'a (48%), Nyeri (41%), Mombasa (38%) and Lamu (38%) have relatively high proportions of overweight. There are 15 counties with a prevalence above the national mean of 28%. It is important to note that the counties Nairobi, Nakuru, Mombasa and Nyeri are located in major towns of Kenya. The high prevalence of overweight seen in Murang'a and Lamu could be explained by socio-cultural factors. However, more research should be done to establish the underlying reasons and targeted interventions implemented.



(a) (b)

Figure 3.9: Maps showing the county level distribution of model-based overweight prevalence in Kenya: (a) Point estimates and (b) the corresponding coefficient of variation.

3.5 Conclusion

Establishing the extent of overweight prevalence is important for the public health surveillance of a country. The information acquired is vital to inform policy-making and resource allocation. This study presents a new data source of overweight prevalence at the county level relevant to the Kenya Health Policy (KHP), 2014-2030 and the Kenya Vision 2030. We have combined survey and census data through a model-based SAE methodology to better estimate the prevalence of overweight at the county level. Our model-based prevalence estimates have smaller MSE's and CV's than direct estimates. We found that counties within urban areas — including the major towns like Nairobi, Nakuru, Nyeri and Mombasa — have a higher prevalence of overweight compared to rural counties. Although we focus on overweight prevalence in Kenya, the presented method can also be applied to other indicators in developing countries with similar data sources. Health is devolved in Kenya. Therefore, counties with high prevalence should do more research and tailor interventions. We provide overweight prevalence estimates at the county level. It will be interesting to further extend this research to include more disaggregated domains like age, sex, gender and ethnicity. One limitation of this study is the time difference between the survey and census data. Data collected around the same time might yield more accurate results. Despite the limitation, this study has estimated the prevalence of overweight at the county level in Kenya with better precision.

Acknowledgments

I am grateful to the Editors and referees for comments that significantly improved the paper. I also thank Sylvia Harmening, Nora Würz, Natalia Rojas-Perilla and Timo Schmid for providing useful comments that improved this paper. Lastly, I thank the Kenya National Bureau of Statistics for providing data and the Kenyan-German postgraduate programme, in which this study is partly funded.

Bibliography

Agyemang, C., S. Boatemaa, G. A. Frempong, and A. de Graft Aikins (2014). Obesity in Sub-Saharan Africa. In: Ahima R. (eds), Chapter 4, pp. 41–53. Cham: Springer International Publishing.

Aikins, M., P. T.-N. Tabong, P. Salari, F. Tediosi, F. M. Asenso-Boadi, and P. Akweongo (2021). Positioning the national health insurance for financial sustainability and universal health coverage in Ghana: A qualitative study among key stakeholders. Plos One 16(6), e0253109.

Asiki, G., S. F. Mohamed, D. Wambui, C. Wainana, S. Muthuri, M. Ramsay, and C. Kyobutungi (2018). Sociodemographic and behavioural factors associated with body mass index among men and women in Nairobi slums: AWI-Gen project. Global Health Action 11(sup2), 1470738.

Atkinson, A. B. (1987). On the measurement of poverty. Econometrica: Journal of the Econometric Society 55(44), 749–764.

Atkinson, T., B. Cantillon, E. Marlier, and B. Nolan (2002). Social indicators: The EU and social inclusion. Oxford: Oxford University Press.

Awuah, R. B., J. K. Anarfi, C. Agyemang, G. Ogedegbe, and A. d.-G. Aikins (2014). Prevalence, awareness, treatment and control of hypertension in urban poor communities in Accra, Ghana. Journal of Hypertension 32(6), 1203–1210.

Bates, D., M. Mächler, B. Bolker, and S. Walker (2015). Fitting linear mixed-effects models using lme4. Journal of Statistical Software 67(1), 1–48.

Battese, G. E., R. M. Harter, and W. A. Fuller (1988). An error-components model for prediction of county crop areas using survey and satellite data. Journal of the American Statistical Association 83(401), 28–36.

Beegle, K. and L. Christiaensen (2019). Accelerating Poverty Reduction in Africa. Report, The World Bank.

Bell, W. R. (2008). Examining sensitivity of small area inferences to uncertainty about sampling error variances. In Proceedings of the American Statistical Association, Survey Research Methods Section.

Betti, G. and A. Lemmi (2013). Poverty and social exclusion: New methods of analysis. New York: Routledge.

Boltvinik, J. (1999). Poverty measurement methods: An overview. UNDP Social Development & Poverty Elimination Division.

Box, G. E. and D. R. Cox (1964). An analysis of transformations. Journal of the Royal Statistical Society: Series B (Methodological) 26(2), 211–243.

Breckling, J. and R. Chambers (1988). M-quantiles. Biometrika 75(4), 761–771.

Brown, G., R. Chambers, P. Heady, and D. Heasman (2001). Evaluation of small area estimation methods . An application to unemployment estimates from the UK LFS. In Proceedings of Statistics Canada Symposium.

Cantoni, E. and E. Ronchetti (2001). Robust inference for generalized linear models. Journal of the American Statistical Association 96(455), 1022–1030.

Carroll, R. J. and D. Ruppert (1988). Transformation and weighting in regression. CRC Press.

Carter, G. M. and J. E. Rolph (1974). Empirical bayes methods applied to estimating fire alarm probabilities. Journal of the American Statistical Association 69(348), 880–885.

Casas-Cordero, J. E. C., and P. Lahiri (2016). Poverty mapping for the Chilean comunas. In Analysis of Poverty Data by Small Area Estimation, Chapter 20, pp. 379–403. John Wiley & Sons.

Chambers, R., H. Chandra, N. Salvati, and N. Tzavidis (2014). Outlier robust small area estimation. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 76(1), 47–69.

Chambers, R., N. Salvati, and N. Tzavidis (2016). Semiparametric small area estimation for binary outcomes with application to unemplnoyment estimation for local authorities in the uk. Journal of the Royal Statistical Society. Series A (Statistics in Society) 179(2), 453–479.

Chambers, R. and N. Tzavidis (2006). M-quantile models for small area estimation. Biometrika 93(2), 255–268.

Chambers, R. L. and R. Dunstan (1986). Estimating distribution functions from survey data. Biometrika 73(3), 597–604.

Chandra, H., K. Aditya, and U. Sud (2018). Localised estimates and spatial mapping of poverty incidence in the state of bihar in india. an application of small area estimation techniques. Plos One 13(6), e0198502.

Chowdhury, M. A. B., M. J. Uddin, M. R. Haque, and B. Ibrahimou (2016). Hypertension among adults in bangladesh: evidence from a national cross-sectional survey. BMC Cardiovascular Disorders 16(1), 1–10.

Christensen, D. L., J. Eis, A. W. Hansen, M. W. Larsson, D. L. Mwaniki, B. Kilonzo, I. Tetens, M. K. Boit, L. Kaduka, K. Borch-Johnsen, et al. (2008). Obesity and regional fat distribution in Kenyan populations: Impact of ethnicity and urbanization. Annals of Human Biology 35(2), 232–249.

Cochran, W. G. (2007). Sampling Techniques. John Wiley & Sons.

Copas, J. B. (1988). Binary regression models for contaminated data. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 50(2), 225–253.

Cotlear, D. and N. Rosemberg (2018). Going universal in Africa. World Bank.

Croft, T. N., A. M. Marshall, C. K. Allen, F. Arnold, S. Assaf, S. Balian, et al. (2018). Guide to DHS statistics.

Datta, G., M. Ghosh, et al. (2012). Small area shrinkage estimation. Statistical Science 27(1), 95–114.

Datta, G. S. and P. Lahiri (2000). A unified measure of uncertainty of estimated best linear unbiased predictors in small area estimation problems. Statistica Sinica 10(2), 613–627.

Dye, C., T. Boerma, D. Evans, A. Harries, C. Lienhardt, J. McManus, T. Pang, R. Terry, and R. Zachariah (2015). Research for universal health coverage: World health report 2013. Science Translational Medicine 199(5), 199ed13.

Efron, B. and C. Morris (1975). Data analysis using Stein's estimator and its generalizations. Journal of the American Statistical Association 70(350), 311–319.

Elbers, C., J. O. Lanjouw, and P. Lanjouw (2003). Micro-level estimation of poverty and inequality. Econometrica 71(1), 355–364.

Eurostat (2021). Glossary: At-risk-of-poverty rate.

Farrell, P. J., B. MacGibbon, and T. J. Tomberlin (1997). Empirical Bayes small-area estimation using logistic regression models and summary statistics. Journal of Business & Economic Statistics 15(1), 101–108.

Fay, R. and R. Herriot (1979). Estimates of income for small places: An application of James-Stein procedures to census data. Journal of the American Statistical Association 74(366a), 269–277.

Feng, Q., J. Hannig, and J. Marron (2016). A note on automatic data transformation. Stat 5(1), 82–87.

Fitzgibbon, C. (2012). Economics of resilience study—Kenya country report. Report, World Bank.

Foster, J., J. Greer, and E. Thorbecke (1984). A class of decomposable poverty measures. Econometrica: Journal of the Econometric Society 52(3), 761–766.

Friedrich Ebert Stiftung (2012). Regional Disparities and Marginalization in Kenya. Nairobi: Elite PrePress.

General Assembly (2015). Transforming our world: the 2030 Agenda for Sustainable Development. Resolution, United Nations.

Ghosh, M., K. Natarajan, T. Stroud, and B. P. Carlin (1998). Generalized linear models for small-area estimation. Journal of the American Statistical Association 93(441), 273–282.

Ghosh, M. and J. N. Rao (1994). Small area estimation: An appraisal. Statistical Science 9(1), 55–76.

Gichu, M., G. Asiki, P. Juma, J. Kibachio, C. Kyobutungi, and E. Ogola (2018). Prevalence and predictors of physical inactivity levels among Kenyan adults (18–69 years): An analysis of STEPS survey 2015. BMC Public Health 18(3), 1–7.

Gini, C. (1912). Variabilità e mutabilità. Reprinted in Memorie di metodologica statistica (Ed. Pizetti E.

Government of Kenya (2007). Kenya Vision 2030. A Globally Competitive and Prosperous Kenya. Report, Government of Kenya.

Government of Kenya (2013). The constitution of Kenya. National Council for Law Reporting.

Greeley, M. (1994). Measurement of poverty and poverty of measurement. Ids bulletin 25(2), 50–58.

Groenveld-van Dijk, E. (2013). The burden of overweight and obesity in kenya analyses of the known determinants and control. Master's thesis, Royal Tropical Institute.

Guadarrama, M., I. Molina, and J. Rao (2016). A comparison of small area estimation methods for poverty mapping. Statistics in Transition new series 1(17), 41–66.

Guadarrama, M., I. Molina, and J. Rao (2018). Small area estimation of general parameters under complex sampling designs. Computational Statistics & Data Analysis 121, 20–40.

Hadam, S., N. Würz, and A.-K. Kreutzmann (2020). Estimating regional unemployment with mobile network data for functional urban areas in Germany. Working paper, Institute for Statistics and Econometrics, Freie UniversitÃ¤t Berlin.

Haider, H. (2020). Conflict analysis of North Eastern Kenya. Technical report, K4D Emerging Issues Report 36. Brighton, UK: Institute of Development Studies.

Hartley, H. O. and J. N. K. Rao (1967). Maximum-likelihood estimation for the mixed analysis of variance model. Biometrika 54(1-2), 93–108.

Hawala, S. and P. Lahiri (2010). Variance modeling in the us small area income and poverty estimates program for the american community survey. In Proceedings of the American Statistical Association, Section on Bayesian Statistical Science, Section on Survey Research Methods, Alexandria, VA: American Statistical Association.

Hawala, S. and P. Lahiri (2018). Variance modeling for domains. Statistics and Applications 16(1), 399–409.

Henderson, C. R. (1975). Best linear unbiased estimation and prediction under a selection model. Biometrics 31(2), 423–447.

Horvitz, D. and D. Thompson (1952). A generalization of sampling without replacement from a finite universe. Journal of the American Statistical Association 47(260), 663–685.

Huber, P. (1981). Robust statistics. New York: John Wiley and Sons, Inc.

Huber, P. J. (1992). Robust estimation of a location parameter. In Breakthroughs in statistics, pp. 492–518. Springer.

Ilinca, S., L. Di Giorgio, P. Salari, and J. Chuma (2019). Socio-economic inequality and inequity in use of health care services in Kenya: evidence from the fourth Kenya household health expenditure and utilization survey. International Journal for Equity in Health 18(1), 1–13.

Jiang, J. and P. Lahiri (2001). Empirical best prediction for small area inference with binary data. Annals of the Institute of Statistical Mathematics 53(2), 217–243.

Jiang, J. and P. Lahiri (2006). Mixed model prediction and small area estimation. Test 15(1), 1–96.

Jiang, J., P. Lahiri, S.-M. Wan, and C.-H. Wu (2001). Jackknifing in the Fay-Herriot model with an example. Technical report, Department of Statistics, University of Nebraska, Lincoln.

Jones-Smith, J. C., P. Gordon-Larsen, A. Siddiqi, and B. M. Popkin (2012). Is the burden of overweight shifting to the poor across the globe? Time trends among women in 39 low-and middle-income countries (1991–2008). International Journal of Obesity 36(8), 1114–1120.

Kazungu, J. S. and E. W. Barasa (2017). Examining levels, distribution and correlates of health insurance coverage in Kenya. Tropical Medicine & International Health 22(9), 1175–1185.

Kenya Institute for Public Policy Research and Analysis (2020). The Kenya economic report 2020: Creating an Enabling Environmentfor Inclusive Growth in Kenya. Report, Kenya Institute for Public Policy Research and Analysis.

Kenya National Bureau of Statistics and Society for International Development (2013). Exploring Kenya's Inequality: Pulling Apart or Pooling Together? Report, Kenya National Bureau of Statistics and Society for International Development.

Kenya National Bureau of Statistics (2010). The 2009 Kenya Population and Housing Census: Volume 1C: Population Distribution by Age, Sex, and Administrative Units. Technical report, Kenya National Bureau of Statistics.

Kenya National Bureau of Statistics (2015). Kenya demographic and health survey 2014. Technical report, Kenya National Bureau of Statistics.

Kenya National Bureau of Statistics (2018). Basic Report on Well-Being in Kenya: Based on the 2015/16 Kenya integrated Household Budget survey (KIHBS). Report, Kenya Bureau of Statistics.

Kenya National Bureau of Statistics (2019a). Gross County Product. Technical report, Kenya National Bureau of Statistics.

Kenya National Bureau of Statistics (2019b). The 2019 Kenya population and housing census: Population by county and sub-county. Technical report, Kenya National Bureau of Statistics.

Kenya National Bureau of Statistics (2020). Inequality trends and Diagnostics in Kenya 2020. Report, Kenya National Bureau of Statistics.

Kenya National Bureau of Statistics, WHO, and Ministry of Health Kenya (2015). Kenya STEPwise survey for non-communicable diseases risk factors 2015 report. Technical report, Ministry of Health,Kenya.

Kish, L. (1949). A procedure for objective respondent selection within the household. Journal of the American Statistical Association 44(247), 380–387.

Koenker, R. and G. Bassett Jr (1978). Regression quantiles. Econometrica: Journal of the Econometric Society 46(1), 33–50.

Koenker, R. and K. F. Hallock (2001). Quantile regression. Journal of Economic Perspectives 15(4), 143–156.

Kokic, P., R. Chambers, J. Breckling, and S. Beare (1997). A measure of production performance. Journal of Business & Economic Statistics 15(4), 445–451.

Kreutzmann, A.-K., S. Pannier, N. Rojas-Perilla, T. Schmid, M. Templ, and N. Tzavidis (2019a). The R package emdi for estimating and mapping regionally disaggregated indicators. Journal of Statistical Software 91(7), 1–33.

Kreutzmann, A.-K., S. Pannier, N. Rojas-Perilla, T. Schmid, M. Templ, and N. Tzavidis (2019b). The R package emdi for estimating and mapping regionally disaggregated indicators. Journal of Statistical Software 91(7), 1–33.

Kulie, T., A. Slattengren, J. Redmer, H. Counts, A. Eglash, and S. Schrager (2011). Obesity and women's health: An evidence-based review. The Journal of the American Board of Family Medicine 24(1), 75–85.

Kusi, A., U. Enemark, K. S. Hansen, and F. A. Asante (2015). Refusal to enrol in Ghana's national health insurance scheme: is affordability the problem? International Journal for Equity in Health 14(1), 1–14.

Lagomarsino, G., A. Garabrant, A. Adyas, R. Muga, and N. Otoo (2012). Moving towards universal health coverage: Health insurance reforms in nine developing countries in Africa and Asia. The Lancet 380(9845), 933–943.

Liu, B., P. Lahiri, and G. Kalton (2007). Hierarchical Bayes modeling of survey-weighted small area proportions. In Proceedings of the American Statistical Association, Survey Research Section, pp. 3181–3186.

Lombardıa, M., W. González-Manteiga, and J. Prada-Sánchez (2003). Bootstrapping the Chambers–Dunstan estimate of a finite population distribution function. Journal of Statistical Planning and Inference 116(2), 367–388.

MacGibbon, B. and T. J. Tomberlin (1987). Small area estimates of proportions via empirical Bayes techniques. Citeseer, 341–346.

Maina, T. and D. Kirigia (2015). Annual evaluation of the abolition of user fees at primary healthcare facilities in Kenya. Technical report, Washington, DC: Futures Group, Health Policy Project.

Malec, D., J. Sedransk, C. L. Moriarity, and F. B. LeClere (1997). Small area inference for binary variables in the national health interview survey. Journal of the American Statistical Association 92(439), 815–826.

Maples, J., W. Bell, and E. T. Huang (2009). Small area variance modeling with application to county poverty estimates from the American community survey. In Proceedings of the Section on Survey Research Methods, Alexandria, VA: American Statistical Association.

Marchetti, S., C. Giusti, and M. Pratesi (2016). The use of twitter data to improve small area estimates of households' share of food consumption expenditure in Italy. AStA Wirtschafts-und Sozialstatistisches Archiv 10(2), 79–93.

Marchetti, S., C. Giusti, M. Pratesi, N. Salvati, F. Giannotti, D. Pedreschi, S. Rinzivillo, L. Pappalardo, and L. Gabrielli (2015). Small area model-based estimators using big data sources. Journal of Official Statistics 31(2), 263–281.

Marchetti, S. and N. Tzavidis (2021). Robust estimation of the Theil index and the Gini coeffient for small areas. Journal of Official Statistics 37(4), 955–979.

Marchetti, S., N. Tzavidis, and M. Pratesi (2012). Non-parametric bootstrap mean squared error estimation for M-quantile estimators of small area averages, quantiles and poverty indicators. Computational Statistics & Data Analysis 56(10), 2889–2902.

Mbochi, R. W. (2010). Overweight and obesity prevalence and associated socio-economic factors, physical activity and dietary intake among women in Kibera division, Nairobi. Ph. D. thesis, Doctoral Dissertation, Kenyatta University.

Mbochi, R. W., E. Kuria, J. Kimiywe, S. Ochola, and N. P. Steyn (2012). Predictors of overweight and obesity in adult women in Nairobi province, Kenya. BMC Public Health 12(1), 1–9.

McGillivray, M. and H. White (1993). Measuring development? The UNDP's human development index. Journal of International Development 5(2), 183–192.

Ministry of Health, G. o. K. (2014). 2013 Kenya household health expenditure and utilisation survey. Technical report, Government of Kenya.

Ministry of Health Kenya (2014). Kenya Health Policy 2014-2030. Technical report, Government of Kenya.

Ministry of Health, Kenya (2018, November). CS health launches UHC pilot registration. Online. Available from: https://www.health.go.ke/cs-health-launches-uhc-pilot-registration-machakos-kenya-november-10-2018/.

Ministry of Health Zambia (2017). Zambia national health strategic plan 2017-2021. Technical report, Ministry of Health.

Mkuu, R. S., K. Epnere, and M. A. B. Chowdhury (2018). Prevalence and predictors of overweight and obesity among Kenyan women. Preventing Chronic Disease 15(E44), 170401.

Molina, I. and J. Rao (2010). Small area estimation of poverty indicators. Canadian Journal of Statistics 38(3), 369–385.

Morales, D., M. D. Esteban, A. Pérez, and T. Hobza (2021). A course on small area estimation and mixed models: Methods, theory and applications in R. Springer Nature.

Muthuri, S. K., L.-J. M. Wachira, V. O. Onywera, and M. S. Tremblay (2014). Correlates of objectively measured overweight/obesity and physical activity in Kenyan school children: results from ISCOLE-Kenya. BMC Public Health 14(1), 1–11.

Mwaura, R. N., E. Barasa, G. N. Ramana, J. Coarasa, and K. Rogo (2015). The path to universal health coverage in Kenya: Repositioning the role of the national hospital insurance fund.

Mwenda, N., R. Nduati, M. Kosgey, and G. Kerich (2021). What drives outpatient care costs in Kenya? An analysis with generalized estimating equations. Frontiers in Public Health 8(648465), 1–18.

Nakagawa, S. and H. Schielzeth (2013). A general and simple method for obtaining R2 from generalized linear mixed-effects models. Methods in Ecology and Evolution 4(2), 133–142.

Nandram, B., J. Sedransk, and L. Pickle (1999). Bayesian analysis of mortality rates for US health service areas. Sankhyā: The Indian Journal of Statistics, Series B 4(2), 145–165.

Newey, W. K. and J. L. Powell (1987). Asymmetric least squares estimation and testing. Econometrica: Journal of the Econometric Society 55(4), 819–847.

Office for National Statistics UK (2017). Methodology for measuring uncertainty in ons local authority mid-year population estimates: 2012 to 2016. shorturl.at/psHV4. Accessed: 2022-03-23.

Okungu, V., J. Chuma, and D. McIntyre (2017). The cost of free health care for all Kenyans: Assessing the financial sustainability of contributory and non-contributory financing mechanisms. International Journal for Equity in Health 16(1), 1–13.

Otieno, P. O., E. O. A. Wambiya, S. F. Mohamed, H. P. P. Donfouet, and M. K. Mutua (2019). Prevalence and factors associated with health insurance coverage in resource-poor urban settings in Nairobi, Kenya: a cross-sectional study. BMJ Open 9(12), e031543.

Patterson, H. D. and R. Thompson (1971). Recovery of inter-block information when block sizes are unequal. Biometrika 58(3), 545–554.

Pawloski, L. R., K. M. Curtin, C. Gewa, and D. Attaway (2012). Maternal–child overweight/obesity and undernutrition in Kenya: A geographic analysis. Public Health Nutrition 15(11), 2140–2147.

Pfeffermann, D. (2002). Small area estimation-new developments and directions. International Statistical Review 70(1), 125–143.

Pfeffermann, D. (2013). New important developments in small area estimation. Statistical Science 28(1), 40–68.

Pfeffermann, D. and M. Sverchkov (1999). Parametric and semi-parametric estimation of regression models fitted to survey data. Sankhyā: The Indian Journal of Statistics, Series B 61(1), 166–186.

Prasad, N. N. and J. N. Rao (1990). The estimation of the mean squared error of small-area estimators. Journal of the American Statistical Association 85(409), 163–171.

Pratesi, M. (2016). Area-level spatio-temporal small area estimation models. In M. Pratesi (Ed.), Analysis of poverty data by small area estimation, Chapter 11. John Wiley & Sons.

Rao, J. N. K. and I. Molina (2015). Small area estimation. Wiley Online Library.

Riley, L., R. Guthold, M. Cowan, S. Savin, L. Bhatti, T. Armstrong, and R. Bonita (2016). The world health organization stepwise approach to noncommunicable disease risk-factor surveillance: Methods, challenges, and opportunities. American Journal of Public Health 106(1), 74–78.

Rojas-Perilla, N., S. Pannier, T. Schmid, and N. Tzavidis (2020). Data-driven transformations in small area estimation. Journal of the Royal Statistical Society: Series A (Statistics in Society) 183(1), 121–148.

Salvucci, V., G. Betti, F. Gagliardi, et al. (2012). Multidimensional and fuzzy measures of poverty and inequality at national and regional level in Mozambique. Technical report, Department of Economics, University of Siena.

Sambo, L. G., J. M. Kirigia, and J. N. Orem (2013). Health financing in the African region: 2000–2009 data analysis. International Archives of Medicine 6(1), 1–17.

Särndal, C.-E., B. Swensson, and J. Wretman (2003). Model assisted survey sampling. Springer Science & Business Media.

Schaible, W. L. (2013). Indirect estimators in US federal programs, Volume 108. Springer Science & Business Media.

Schilling, J., F. E. Opiyo, and J. Scheffran (2012). Raiding pastoral livelihoods: Motives and effects of violent conflict in north-western Kenya. Pastoralism: Research, Policy and Practice 2(1), 1–16.

Schmid, T., F. Bruckschen, N. Salvati, and T. Zbiranski (2017). Constructing sociodemographic indicators for national statistical institutes by using mobile phone data: Estimating literacy rates in Senegal. Journal of the Royal Statistical Society: Series A (Statistics in Society) 180(4), 1163–1190.

Sebire, N. J., M. Jolly, J. Harris, J. Wadsworth, M. Joffe, R. Beard, L. Regan, and S. Robinson (2001). Maternal obesity and pregnancy outcome: A study of 287, 213 pregnancies in London. International Journal of Obesity 25(8), 1175–1182.

Sen, A. (1976). Poverty: An ordinal approach to measurement. Econometrica: Journal of the Econometric Society 44(2), 219–231.

Steyn, N. P. and Z. J. Mchiza (2014). Obesity and the nutrition transition in Sub-Saharan Africa. Annals of the New York Academy of Sciences 1311(1), 88–101.

Transparency International Kenya (2016). The Kenya county gorvenance status report 2016. Technical report, Transparency International Kenya.

Tzavidis, N., S. Marchetti, and R. Chambers (2010). Robust estimation of small-area means and quantiles. Australian & New Zealand Journal of Statistics 52(2), 167–186.

Tzavidis, N., L.-C. Zhang, A. Luna, T. Schmid, and N. Rojas-Perilla (2018). From start to finish: A framework for the production of small area official statistics. Journal of the Royal Statistical Society: Series A (Statistics in Society) 181(4), 927–979.

Wang, H., N. Otoo, and L. Dsane-Selby (2017). Ghana national health insurance scheme. The World Bank.

World Bank (2018). Policy options to advance the big 4: Unleashing Kenya's private sector to drive inclusive growth and accelerate poverty reduction. Technical report, The World Bank.

World Bank (2020). Poverty and shared prosperity 2020: Reversals of fortune. The World Bank.

World Bank (2021). Poverty and equity brief, Africa eastern and southern.

World Health Organization (2005a). Sustainable health financing, universal coverage and social health insurance. Technical report, World Health Organization.

World Health Organization (2005b). Who steps surveillance manual: The who stepwise approach to chronic disease risk factor surveillance. Technical report, World Health Organization.

World Health Organization (2013). Research for universal health coverage: World health report 2013. Technical report, World Health Organization.

World Health Organization (2021). World health organization obesity and overweight fact sheet. `https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight`. Accessed: 25.05.2020.

World Health Organization et al. (2010). The Abuja declaration: Ten years on. Technical report, World Health Organization.

Zhongming, Z., L. Linong, Z. Wangqiang, L. Wei, et al. (2021). Practical guidebook on data disaggregation for the sustainable development goals. Asian Development Bank.

Summaries

Abstracts in English

Abstract: Estimation of disaggregated poverty and inequality indicators with application to the Kenya Integrated Household Budget Survey

Achieving the Sustainable Development Goals (SDG) of no poverty and reduced inequality are key priorities for the Kenyan government. Information about poverty and inequality is important for decision making and policy analysis. Providing targeted interventions on poverty and inequality requires reliable information at disaggregated levels. Although sample surveys are sufficient sources of data for large sample sizes, they are not sufficient for small samples. Therefore, in this paper two model-based small area estimation methods are used to estimate poverty and inequality indicators in Kenya by linking data from the Kenya Integrated Household Budget Survey 2015 and the Kenya Population and Housing Census 2009. Particularly, the Empirical Best Predictor (EBP) and a M-quantile model are explored. For practical reasons, four indicators are estimated i.e. the Mean, Head Count Ratio, Poverty Gap and the Gini coefficient. Since the EBP assumes normality assumption, three transformations are explored: the logarithmic, Log-shift and the Box-Cox. To assess the uncertainty in estimation a parametric bootstrap is used to estimate the Mean Squared Error (MSE). is used as another way of dealing with deviations from normality of model errors terms through robust regression techniques. The MSE is estimated employing a non-parametric MSE estimator. The two models give similar results, especially for the Mean, Head Count Ratio and Poverty Gap. For the Gini coefficient though both model-based approaches do not agree with each other and the Direct estimates are more reasonable. Overall this paper shows that county-level estimates obtained through model-based small area estimation methods are more precise than Direct estimates based solely on the survey data.

Keywords:Box-Cox transformation, Bootstrap, Empirical Best Predictor, M-quantile, Poverty mapping, Small Area Estimation

Abstract: Small area estimation of health insurance coverage for Kenyan counties

Health insurance is important in disease management, access to quality health care and attaining Universal Health Care. National and regional data on health insurance coverage needed for policy-making is mostly obtained from household surveys; however, estimates at lower administrative units like at the county level in Kenya are highly variable due to small sample sizes. Small area estimation combines survey and census data using a model to increases the effective sample size and therefore provides more precise estimates. In this study we estimate the health insurance coverage for Kenyan counties using a binary M-quantile small area model for women $(n = 14,730)$ and men $(n = 12,007)$ aged 15 to 49 years old. This has the advantage that we avoid specifying the distribution of the random effects and distributional robustness is automatically achieved. The response variable is derived from the Kenya Demographic and Health Survey 2014 and auxiliary data from the Kenya Population and Housing Census 2009. We estimate the mean squared error using an analytical approach based on Taylor series linearisation. The national direct health insurance coverage estimates are 18% and 21% for women and men respectively. With the current health insurance schemes, coverage remains low across the 47 counties. These county-level estimates are helpful in formulating decentralized policies and funding models.

Keywords: Binary M-quantile, Direct estimation, Health insurance coverage, Universal Health Care, Taylor series linearisation.

Abstract: Estimating county-level overweight prevalence in Kenya using small area methodology

Public health surveillance of overweight prevalence is essential to assess the extent of the problem, identify regions and groups most affected and inform policymaking. However, the needed reliable data at disaggregated levels is lacking in Kenya. The Kenya STEPwise Survey for Non-communicable Diseases and Risk Factors (KSSNDRF) was nationally representative. It was used to obtain various indicators of non-communicable diseases and risk factors including overweight. However, due to small sample sizes at lower levels like at the county, overweight prevalence estimates are statistically imprecise (i.e., high variance). Therefore, to increase the effective sample size we combine data from the KSSNDRF and the Kenya Population and Housing Census by model-based small area methods. In particular, we fit an arcsine square-root transformed Fay-Herriot model. To transform back to the original scale, we use a bias-corrected back transformation. For this model, we smooth the design variance using Generalized Variance Functions. We compute the mean squared error estimates using a bootstrap procedure. We found that counties within urban areas — including the major towns like Nairobi, Nakuru, Nyeri and Mombasa — have a higher prevalence of overweight compared to rural counties. Although the paper focuses on overweight prevalence in Kenya, the presented method can also be applied to other indicators in developing countries with similar data sources.

Keywords: Direct estimation, Fay-Herriot model, Prevalence mapping, Sample surveys, Transformations.

Kurzzusammenfassungen auf Deutsch

Zusammenfassung: Schätzung disaggregierter Armuts- und Ungleichheitsindikatoren mit Anwendung auf die Kenya Integrated Household Budget Survey

Das Erreichen der Ziele für nachhaltige Entwicklung (Sustainable Development Goals, SDG) sind zentrale Prioritäten der Kenianischen Regierung. Informationen über Armut und Ungleichheit sind wichtig für die Entscheidungsfindung und Politikanalyse. Die Bereitstellung gezielter Interventionen zu Armut und Ungleichheit erfordert zuverlässige Informationen auf disaggregierter Ebene. Obwohl Stichprobenerhebungen für große Stichprobenumfänge ausreichende Datenquellen sind, reichen sie für kleine Stichproben nicht aus. Daher werden in diesem Beitrag zwei modellbasierte Schätzmethoden für kleine Gebiete verwendet, um Armuts- und Ungleichheitsindikatoren in Kenia zu schätzen, indem Daten aus der Kenya Integrated Household Budget Survey 2015 und dem Kenya Population and Housing Census 2009 verknüpft werden. Insbesondere werden der Empirical Best Predictor (EBP) und das M-Quantil-Modell untersucht. Aus praktischen Gründen werden vier Indikatoren geschätzt, der Mittelwert, die Armutsrate, die Armutslücke und der Gini-Koeffizient. Da das statistische Modell des EBPs von einer Normalitätsannahme, werden drei Transformationen untersucht: die logarithmische Transformation, die Log-Verschiebung und die Box-Cox-Transformation. Zur Abschätzung der Unsicherheit in Form des mittleren quadratischen Fehlers (MSE) wird ein parametrischer Bootstrap verwendet. Das M-Quantil-Modell wird als eine weitere Möglichkeit verwendet, um mit Abweichungen von der Normalität der Modellfehlerterme durch robuste Regressionstechniken zu umgehen. Der MSE wird unter Verwendung eines nichtparametrischen MSE-Schätzers geschätzt. Die beiden Modelle liefern ähnliche Ergebnisse, insbesondere für den Mittelwert, die Armutsrate und die Armutslücke. Für den -Koeffizienten stimmen die beiden modellbasierten Ansätze jedoch nicht überein und die direkten Schätzungen sind sinnvoller. Insgesamt zeigt dieser Beitrag dass, Schätzungen auf Kreisebene, die durch modellbasierte Schätzmethoden für kleine Gebiete erzielt werden, genauer als direkte Schätzungen sind, die ausschließlich auf den Erhebungsdaten basieren.

Schlüsselwörter: Box-Cox-Transformation, Bootstrap, empirischer bester Prädiktor, M-Quantil, Armutskartierung, Small-Area-Methoden.

Zusammenfassung: Schätzung des Krankenversicherungsschutzes für kenianische Bezirke

Krankenversicherungen sind wichtig für das Gesundheitsmanagement, den Zugang zu qualitativ hochwertiger Gesundheitsversorgung und das Erreichen einer universellen Gesundheitsversorgung. Nationale und regionale Daten zum Krankenversicherungsschutz, die für die Politikgestaltung benötigt werden, werden hauptsächlich aus Haushaltsbefragungen gewonnen. Schätzungen für niedrigere Verwaltungseinheiten wie auf Bezirksebene sind in Kenia aufgrund kleiner Stichprobengrößen sehr unterschiedlich und nicht verlässlich. Small Area Estimation (SAE) kombiniert Erhebungs- und Volkszählungsdaten mithilfe eines statistischen Modells, um die effektive Stichprobengröße zu erhöhen, und liefert daher genauere Schätzungen. In dieser Studie schätzen wir den Krankenversicherungsschutz für kenianische Bezirke unter Verwendung eines binären M-Quantil-SAE-Modells für Frauen $(n = 14.730)$ und Männer $(n = 12.007)$ im Alter von 15 bis 49 Jahren. Dies hat den Vorteil, dass wir vermeiden, die Verteilung der Zufallseffekte zu spezifizieren und die Verteilungsrobustheit automatisch erreicht wird. Die Antwortvariable wird aus der Kenya Demographic and Health Survey 2014 und Hilfsdaten aus dem Kenya Population and Housing Census 2009 abgeleitet. Wir schätzen den mittleren quadratischen Fehler unter Verwendung eines analytischen Ansatzes, der auf der Linearisierung von Taylorreihen basiert. Die nationale Krankenversicherung beträgt 18% und 21% für Frauen bzw. Männer. Mit den derzeitigen Krankenversicherungssystemen bleibt die Abdeckung in den 47 Bezirken gering. Diese Schätzungen auf Bezirksebene sind hilfreich bei der Formulierung dezentraler Richtlinien und Finanzierungsmodelle.

Schlüsselwörter: Binäres M-Quantil, direkte Schätzung, Krankenversicherungsschutz, Universellen Gesundheitsversorgung, Linearisierung der Taylor-Reihe.

Zusammenfassung: Schätzung der Prävalenz von Übergewicht auf Bezirksebene in Kenia unter Verwendung einer regionalen Schätzmethode.

Die regionale Bestimmung von Übergewicht durch öffentliche Behörden ist essentiell, um das Ausmaß des Problems zu beurteilen, die am stärksten betroffenen Regionen und Gruppen zu identifizieren und die Politik zu informieren. In Kenia fehlen jedoch die erforderlichen zuverlässigen Daten auf disaggregierter Ebene. Die Kenya STEPwise Survey for Non-communicable Diseases and Risk Factors (KSSNDRF) ist landesweit repräsentativ. Sie wird verwendet, um verschiedene Indikatoren für nicht übertragbare Krankheiten und Risikofaktoren, einschließlich Übergewicht, zu erhalten. Aufgrund kleiner Stichprobenumfänge auf niedrigeren Ebenen wie etwa Landkreise sind die Prävalenzschätzungen für Übergewicht jedoch statistisch ungenau (d. h. eine hohe Varianz). Um die effektive Stichprobengröße zu erhöhen, kombinieren wir daher Daten aus dem KSSNDRF und der Kenya Population and Housing Census 2009 mit modellbasierten Methoden für kleine Gebiete. Insbesondere passen wir ein arcsin-transformiertes Fay-Herriot-Modell an. Um zurück in die ursprüngliche Skala zu transformieren, verwenden wir eine Bias-korrigierte Rücktransformation. Für dieses Modell glätten wir die Varianz des Stichprobenfehlers unter Verwendung von verallgemeinerten Varianzfunktionen. Die Schätzung des mittleren quadratischen Fehlers erfolgt unter Verwendung eines Bootstrap-Verfahrens. Wir fanden heraus, dass Bezirke innerhalb städtischer Gebiete — einschließlich der großen Städte wie Nairobi, Nakuru, Nyeri und Mombasa — im Vergleich zu ländlichen Bezirken eine höhere Prävalenz von Übergewicht aufweisen. Obwohl sich dieser Beitrag auf die Prävalenz von Übergewicht in Kenia konzentriert, kann die vorgestellte Methode auch auf andere Indikatoren in Entwicklungsländern mit ähnlichen Datenquellen angewendet werden.

Schlüsselwörter: Direkte Schätzung, Fay-Herriot-Modell, Prävalenzkartierung, Stichprobenerhebung, Transformationen.

Declaration of Authorship

I certify that this work contains no material which has been accepted for the award of any other degree or diploma in my name, in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text.

Berlin, March 20, 2023

Noah Cheruiyot Mutai

March 20, 2023