

The copyright of this thesis vests in the author. No quotation from it or information derived from it is to be published without full acknowledgement of the source. The thesis is to be used for private study or non-commercial research purposes only.

Published by the University of Cape Town (UCT) in terms of the non-exclusive license granted to UCT by the author.

POVERTY MAPPING: WHAT CAN WE
REALLY LEARN?

JESSE NAIDOO

Supervisor:

Associate Professor Martin Wittenberg (Economics, University of Cape Town)

Submitted in fulfillment of the requirements
of the degree of Master of Commerce (Economics)
at the University of Cape Town

August 2010

Abstract

I examine the problem of poverty mapping. I explain the difficulties of small-area estimation arising from data sparseness and argue that any attempt to overcome this fundamental problem has to take the form of a conditional homogeneity assumption. I estimate poverty maps of South Africa's provinces using the technique developed by Elbers, Lanjouw and Lanjouw Elbers et al. [2003a] on combined data from 1995 (the October Household Survey and Income and Expenditure Survey) and 1996 (the Population Census). I demonstrate that the results are highly sensitive to specification, severely biased in finite samples, and almost certain to fail to estimate the poverty headcount consistently. To deal with these problems, I develop a numerically simple nonparametric small-area estimator which is consistent in all areas under weaker conditions than the Elbers, Lanjouw and Lanjouw estimator. I derive a formula for the (conditional) variance of the nonparametric estimator and I provide evidence (from a Monte Carlo simulation) that its estimated standard errors are small and negligibly biased.

To Stella

University of Cape Town

Acknowledgements

When first I met Professor Martin Wittenberg, I realised instantly that I could learn a lot from him. I did not, however, realise exactly how much I would. I am in his debt for all the times he has corrected me or pointed me in a different (and invariably more fruitful) direction, and thus for all the mistakes he has prevented me from making. I also owe him a deep personal debt for his advice, his mentorship, and the example of honesty, clarity, and rigor of thought that I have seen him set every time we talk. I would not be pursuing a career in economics were it not for Professor Wittenberg's influence.

I want to thank my family - Kevin, Eamonn, Claire and Joshua - for being a consistent source of joy in my life. Of course, our Wednesday night suppers, being a consistent source of food, helped too.

Lastly, I want to thank my dear friends David Watson and Ian Grose, and of course my incomparably wonderful girlfriend Claudia Gastrow for their support, their cajoling, and their tolerance of my distraction and irritation.

Contents

1	Introduction to Poverty Mapping	1
1.1	What Are Poverty Maps?	1
1.2	Who Uses Poverty Maps, and Why?	3
1.3	Why is Small-Area Estimation Necessary?	4
1.3.1	A Prelude on Notation	4
1.3.2	The Problem: Sparse Data	5
1.3.3	A Potential Solution: Combining Census and Survey Data	5
1.4	Conceptual Framework	7
1.4.1	Welfare Measures	7
1.4.2	“Idiosyncratic” vs Sampling Error	8
1.5	The Geometry of Small-Area Estimation	9
1.5.1	The Reaggregation Constraint	9
1.5.2	The Fundamental Problem	11
2	The Data	13
2.1	Context: South Africa’s Changing Administrative Geography	13
2.2	1995 October Household Survey/Income and Expenditure Sur- vey	14
2.2.1	Sample Design	14
2.2.2	Province-Level Expenditure Distributions	15
2.3	1996 Population Census	16
2.3.1	Sample Design	16
2.4	Data Construction and Cleaning	17
2.4.1	Household-level Covariates	18
2.4.2	District-level Covariates	18
2.5	Is Small-Area Estimation Suitable?	18
2.5.1	The Survey Data Is Indeed Sparse, and the Census Data Is Indeed Plentiful	20

2.5.2	The Marginal Distribution of the Covariates Is Indeed Heterogenous	22
3	The ELL Technique	25
3.1	The Basic Idea	25
3.2	Estimators, Vagueness and Methodological Choice	26
3.3	Computational Details	27
3.3.1	Assumptions	28
3.3.2	Recreate the Conditional Distribution	29
3.3.3	Integrate With Respect to the Conditional Distribution	31
3.3.4	Estimate The Standard Errors	32
3.4	What Can Properly Be Considered An “ELL” Estimate?	33
3.4.1	Implementing ELL Requires Arbitrary Choices	38
3.4.2	Implementation	39
4	Weaknesses of ELL	41
4.1	Sensitivity to Specification	42
4.1.1	Point Estimates: Small Areas	42
4.1.2	Intra-Regional Rankings	43
4.1.3	Point Estimates: (Reaggregated) Regional	44
4.2	Finite-Sample Bias	47
4.2.1	Existence	47
4.2.2	Magnitude	52
4.3	Consistency	55
4.3.1	A Simple Illustration of the Inadequacy of the R^2 Criterion	55
4.3.2	Direct Evidence: Specification Choice Shifts The Conditional Mean	56
4.3.3	Specification Error and Sampling Error: A Decomposition	59
5	A Nonparametric Estimation Technique	61
5.1	Introduction and Motivation	61
5.1.1	The Setup	61
5.1.2	Assumptions	62
5.1.3	Deriving the Estimator from the Analogy Principle	63
5.2	Properties of the Nonparametric Estimator	65
5.2.1	Consistency	65
5.2.2	Reaggregation	67
5.2.3	Finite-Sample Distribution	69

5.3	Comparison with ELL: A Monte Carlo Experiment	72
5.3.1	Consistency	73
5.3.2	Bias and Variance	77
5.4	Empirical Application: Poverty in the Free State	80
6	Conclusions	84
A	Further Results	86
A.1	Summary Statistics: First-Stage R^2 Values	86
A.2	Sensitivity: Area Headcount	89
A.3	Sensitivity: Rankings	99
B	Descriptive Statistics	110
B.1	Consumption Data	110
B.2	Household-Level Covariates	112
B.3	Area-Level Controls (Census Covariates)	125

University of Cape Town

List of Tables

2.1	Direct Estimates (from IES Data) of Headcount, by Province	15
2.2	Variable Definitions - Household-Level Covariates	19
2.3	Variable Definitions - District-Level Covariates	20
2.4	Sample Sizes - Northern Cape	21
2.5	A Simple "Significance" Check of the Covariates	23
2.6	Census Sample Means, Northern Cape	24
3.1	Selected Poverty Mapping Papers and their (Authors') Methodological Choices	38
4.1	Estimates Over 50 Random Specifications, North West	43
4.2	Within-Province Rankings Over 50 Random Specifications, North West	44
4.3	Implied Regional Headcount Over 50 Random Specifications, by Province	46
4.4	Percentage of Areas Rejecting $d_s = 0$, by Province and Significance Level	53
4.5	Summary Statistics - Lower Bounds for Bias, By Province	54
5.1	Means of Sampling Distributions, ELL and Nonparametric Estimates	77
5.2	Summary Statistics - Estimated Standard Errors, ELL Estimator	78
5.3	Summary Statistics - Estimated (Conditional) Standard Errors, $\sqrt{\widehat{V}[\widehat{\mu}_a N_a, s_x]}$	79
5.4	Estimated Headcounts Based on Alternative Predictors	82
A.1	Summary Statistics: R^2 Values Over Alternative Specifications	86
A.2	R^2 Over 200 Bootstrap Repetitions, by Province	88
A.3	Estimates Over 50 Random Specifications, W Cape	90

A.4	Estimates Over 50 Random Specifications, E Cape	92
A.5	Estimates Over 50 Random Specifications, N Cape	93
A.6	Estimates Over 50 Random Specifications, Free State	94
A.7	Estimates Over 50 Random Specifications, KwaZulu-Natal	96
A.8	Estimates Over 50 Random Specifications, Gauteng	97
A.9	Estimates Over 50 Random Specifications, Mpumalanga	98
A.10	Estimates Over 50 Random Specifications, Limpopo	99
A.11	Within-Province Rankings Over 50 Random Specifications, W Cape	100
A.12	Within-Province Rankings Over 50 Random Specifications, E Cape	102
A.13	Within-Province Rankings Over 50 Random Specifications, N Cape	103
A.14	Within-Province Rankings Over 50 Random Specifications, Free State	105
A.15	Within-Province Rankings Over 50 Random Specifications, KwaZulu-Natal	106
A.16	Within-Province Rankings Over 50 Random Specifications, Gauteng	107
A.17	Within-Province Rankings Over 50 Random Specifications, Mpumalanga	108
A.18	Within-Province Rankings Over 50 Random Specifications, Limpopo	109
B.1	Summary Statistics - Log Monthly Total Expenditure, by Province (IES Data)	111
B.2	Summary Statistics - HH Controls, W Cape (Census Data)	112
B.3	Summary Statistics - HH Controls, W Cape (IES Data)	113
B.4	Summary Statistics: HH Controls, E Cape (Census Data)	113
B.5	Summary Statistics: HH Controls, E Cape (IES Data)	115
B.6	Summary Statistics - HH Controls, N Cape 1	115
B.7	Summary Statistics - HH Controls, N Cape (IES Data)	116
B.8	Summary Statistics - HH Controls, Free State (Census Data)	116
B.9	Summary Statistics - HH Controls, Free State (IES Data)	117
B.10	Summary Statistics - HH Controls, KZN (Census Data)	118
B.11	Summary Statistics - HH Controls, KZN (IES Data)	118
B.12	Summary Statistics - HH Controls, North West (Census Data)	119
B.13	Summary Statistics - HH Controls, North West (IES Data)	120
B.14	Summary Statistics - HH Controls, Gauteng (Census)	120
B.15	Summary Statistics - HH Controls, Gauteng (IES Data)	121

B.16 Summary Statistics - HH Controls, Mpumalanga (Census Data)	122
B.17 Summary Statistics - HH Controls, Mpumalanga (IES Data)	122
B.18 Summary Statistics - HH Controls, Limpopo (Census Data)	123
B.19 Summary Statistics - HH Controls, Limpopo (IES Data)	124
B.20 Summary Statistics - Area Controls, W Cape (Census Data)	125
B.21 Summary Statistics - Area Controls, W Cape (IES Data)	125
B.22 Summary Statistics - Area Controls, E Cape (Census Data)	126
B.23 Summary Statistics - Area Controls, E Cape (IES Data)	127
B.24 Summary Statistics - Area Controls, N Cape (Census Data)	128
B.25 Summary Statistics - Area Controls, N Cape (IES Data)	128
B.26 Summary Statistics - Area Controls, Free State (Census Data)	129
B.27 Summary Statistics - Area Controls, Free State (IES Data)	130
B.28 Summary Statistics: Area Controls, KZN (Census Data)	130
B.29 Summary Statistics - Area Controls, KZN (IES Data)	131
B.30 Summary Statistics - Area Controls, North West (Census Data)	132
B.31 Summary Statistics - Area Controls, North West (IES Data)	132
B.32 Summary Statistics - Area Controls, Gauteng (Census Data)	133
B.33 Summary Statistics - Area Controls, Gauteng (IES Data)	134
B.34 Summary Statistics - Area Controls, Mpumalanga (Census Data)	135
B.35 Summary Statistics - Area Controls, Mpumalanga (IES Data)	135
B.36 Summary Statistics - Area Controls, Limpopo (Census Data)	136
B.37 Summary Statistics - Area Controls, Limpopo (IES Data)	137

List of Figures

1.1	Poverty Map of the Free State in 1996. Source: Alderman et al. [2002]. Permanent URL: http://go.worldbank.org/PYZKB2ADO0	2
1.2	The Reaggregation Constraint: $\omega_1\mu_1 + \omega_2\mu_2 = \mu_R$. Only pairs (μ_1, μ_2) on the illustrated line are consistent with the region-level estimates.	12
2.1	Kernel Density Plots - KwaZulu-Natal; Western, Eastern, and Northern Cape.	16
2.2	Kernel Density Plots - Free State, North West, Gauteng, Mpumalanga, and Limpopo.	17
4.1	Distribution of Estimates for Kudumane (North West) Over 50 Random Specifications	45
4.2	Density of Implied Headcount Over 50 Random Specifications, Eastern Cape	48
4.3	Joint Distribution of Bootstrapped Estimates for Komga (Eastern Cape)	50
4.4	Marginal Bootstrap Densities for Komga (Eastern Cape)	50
4.5	Marginal Bootstrap Density for Difference in Estimates, Komga (Eastern Cape)	51
4.6	Marginal Bootstrap Density for Difference in Estimates, Reaggregated Headcount (KwaZulu-Natal)	52
4.7	Estimated Conditional Mean Log Expenditure Under Different Specifications, Bizana (Eastern Cape)	57
4.8	Marginal Bootstrap Densities for Bizana (Eastern Cape)	58
5.1	Sampling Distributions - Nonparametric Estimator, Area 3	74
5.2	Sampling Distributions - ELL Estimator, Area 3	75
5.3	Sampling Distributions - Reaggregated Nonparametric Estimates	76

5.4 (log) Consumption, by Gender of Head 83

University of Cape Town

Chapter 1

Introduction to Poverty Mapping

1.1 What Are Poverty Maps?

Poverty maps are exactly what they sound like: representations of the extent of poverty throughout a given geographical region. In fact, poverty maps are slightly more general than that: there are “inequality maps,” which chart the geographic variations in a given measure of inequality over a certain region, too.

In this thesis, I will use the term “welfare” to refer both to inequality measures like the Gini coefficient and to mean-sensitive functionals of the income distribution like the poverty gap; the distinction is not important for my purposes. Nonetheless, my concern in this work is primarily with the poverty headcount (a welfare measure which is introduced in section 1.4.1 below), so I will continue to use the term “poverty map” even though the econometric arguments I present here generalise easily to other welfare measures.

Since the turn of the millennium, an empirical literature on poverty maps has emerged. This literature primarily consists of the estimation of poverty maps for various developing countries, although there are some papers that introduce and examine the methodology used in the production of such maps. I review this literature in section 1.2. However, since my focus in this thesis is methodological - on the reliability of poverty maps rather than any country-specific issue - I do not recount the results of the papers which only present estimated maps.

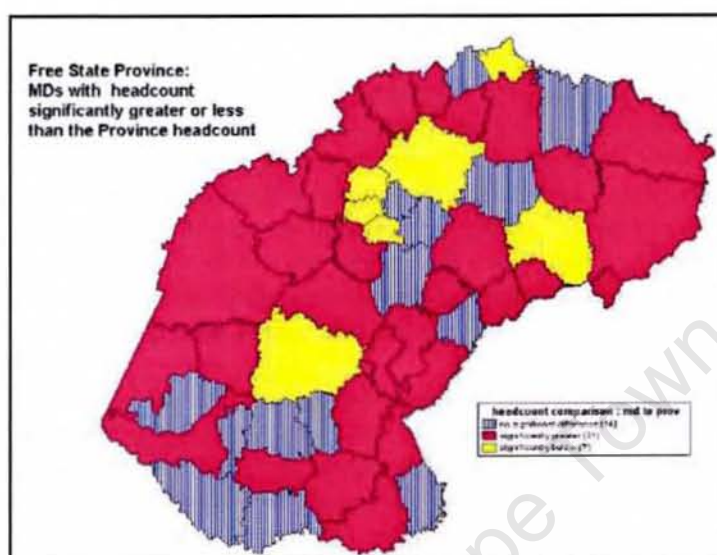


Figure 1.1: Poverty Map of the Free State in 1996. Source: Alderman et al. [2002]. Permanent URL: <http://go.worldbank.org/PYZKB2ADO0>

The thesis is organised as follows: in this introductory chapter I explain some background concepts concerning small-area estimation and welfare measures; I show how the methodological challenges of poverty mapping arise from the demands of policymakers and of the academic literature, and I suggest a way of viewing the econometric problem of poverty mapping in geometric terms. In chapter 2, I introduce the datasets that I use in the empirical work and I discuss the region-level poverty estimates for South Africa in 1996. In chapter 3, I review the existing methodology and highlight its vulnerability to, roughly speaking, specification error. Motivated by the discussion of these vulnerabilities, I explore the sensitivity to misspecification, the finite-sample bias, and the likely inconsistency of the existing poverty-mapping methodology in chapter 4. Finally, in chapter 5, I propose a nonparametric small-area estimator for welfare measures that is consistent under strictly weaker assumptions than the existing technique.

1.2 Who Uses Poverty Maps, and Why?

Policymakers all over the world, but especially in developing countries, want to target the poor geographically. In South Africa, a clause in the Constitution¹ requires that nationally raised revenue be divided “equitably” between national, provincial and local governments. The Constitution further explicitly requires Parliament to interpret “equitable” in terms of “the fiscal capacity and efficiency of the provinces and municipalities, [the] developmental and other needs of provinces, local government and municipalities, [and] economic disparities within and among the provinces.”

The South African government has implemented this clause by creating the “equitable share” grant, of which R25.6 billion - roughly \$3 billion - went to local governments (see National Treasury, Republic of South Africa [2009a], Parliament of the Republic of South Africa [2009] for further details) in 2009. The equitable share grant has several components, but the majority of equitable share funds - about 70%, according to Department of Local Government, Western Cape Provincial Administration [2002], Loots [2004] - are allocated to municipalities in proportion to their levels of poverty.² The National Treasury estimates that the equitable share grant accounts for 17.5% of municipal operating revenue across the nation, though this hides significant inter-regional variation - specifically, large urban municipalities are able to raise funds through property taxes and utility provision; rural municipalities, which have poorer populations and far less commercial activity, depend much more heavily on the equitable share grant. In this context, reliable estimates of poverty at a fine level of disaggregation are very important from a political and - assuming at least some of the grant money filters through to the population at large - a welfare perspective.

South Africa is not alone in its attempts: Henninger and Snel [2002] and World Bank [2007] outline the antipoverty programs in - among others - Guatemala, Nicaragua, Vietnam, Brazil, Albania, Morocco, and Indonesia which have used the (mostly World Bank - generated) poverty maps to target the poor geographically. Some World Bank researchers - e.g. Hentschel and Lanjouw [1998] - have advocated for the use of poverty maps to plan infrastructural investments.

¹Parliament of the Republic of South Africa [1996]: chapter 13, section 214.

²A full explanation of the equitable share formula can be found in Annex W1 to the 2009 National Budget Review (National Treasury, Republic of South Africa [2009b]).

Not only policymakers want small-area estimates of welfare. For instance, the growth literature has increasingly recognised the salience of welfare distribution: Banerjee and Duflo [2003] is just one prominent example. Secondly, reliable estimates of inter-regional welfare distributions are clearly a prerequisite for many lines of inquiry in political economy, public economics and economic geography. And finally, there are reasons to suspect that welfare distribution, broadly conceived, affects other social and economic phenomena, like crime, investment and migration. In fact, at least one study - Demombynes and Özler [2005] - has already used small-area estimates of welfare calculated in exactly the manner described below to examine the spatial distribution of crime in South Africa.

Thus reliable, disaggregated, estimates of welfare measures would be useful as inputs into other areas of research, as well as to policymakers, who might find such estimates helpful not only for antipoverty initiatives, but also to inform other types of policy debates.

1.3 Why is Small-Area Estimation Necessary?

1.3.1 A Prelude on Notation

I partition the population of interest into several “regions”. Monetary variables - either (log) income or expenditure - are denoted y . I define a “region” to be the lowest administrative level for which we have representative data on the distribution of expenditure, y . In the present application (on South African data), this means that I choose regions to be provinces. A given region consists of a number of small areas, indexed by the subscript a ($1 \leq a \leq A$); I define “area” to mean “magisterial district,” since that is the finest administrative level by which the data I have can be grouped.

I deliberately confuse households³ and individuals. Since welfare measures are almost always defined over individuals, yet survey data is almost universally collected at the household level, the caveat that the data need to be weighted by household size is ever-present here. That said, I index households with a subscript i , whence $1 \leq i \leq N_a$, where N_a is the number of households in area a . Household-level covariates of consumption (or income) that appear in both the census and survey data are represented by \mathbf{x}_i .

³Statistics South Africa defines two persons to be members of the same household if they sleep in a common dwelling for four or more days per week and regularly shares meals with the other members.

1.3.2 The Problem: Sparse Data

Fundamentally, the problem is that the data we have are too sparse at the level (area) we want to examine to deliver estimates that have reasonable precision.

The problem of data sparseness arises because of the following two facts:

1. Reliable information on individual welfare (i.e. expenditure or income) is expensive to collect, particularly in developing countries where much economic activity is informal.
2. There are detailed (household-level) surveys that do measure this information accurately (such as South Africa's Income and Expenditure Surveys), but because of the costs of data collection, they are designed to be representative of the population only at fairly high levels of aggregation, like the province or nation.

Unfortunately, this means that for most of the small areas (magisterial districts) in a given region (province), we will have no observations at all; for many others, there may only be a handful. Direct estimates of welfare measures at the small-area level are therefore likely to have standard errors so large as to render them practically useless. Hence, welfare measures can only be reliably *directly* estimated at these high levels of aggregation (the "region" level).

However, census data often contains many of the same variables describing a household's demographic composition, physical living conditions (e.g. the availability of running water), as well as its economic characteristics (e.g. number of members employed, the levels of education attained by household members). It is the existence of these covariates \mathbf{x} that suggests linking small, high-quality datasets with large, more crude ones.

1.3.3 A Potential Solution: Combining Census and Survey Data

Any candidate small-area estimation method - not just the current ELL technique, which is described further in chapter 3 - must impose *some* sort of homogeneity assumption: we know something about the distribution of

y in the region, but not much about it in the constituent areas. Only if the region is homogenous (at least, conditional on some observable covariates \mathbf{x}), then the distribution of y can be reconstructed from the distribution of \mathbf{x} in each area a .

By imposing a conditional homogeneity assumption, it may be possible to reconstruct the distribution of y in each small area; for if we know the conditional distribution $y|\mathbf{x}$, and we have the marginal distribution of \mathbf{x} in each small area (as we will with a census), then we can obtain the marginal distribution of y by multiplying and integrating. Below I explain why this reconstruction is possible given the structure of the available data.

Writing $f_a(y, \mathbf{x})$ for the joint distribution of (y, \mathbf{x}) in area a , $f_R(y, \mathbf{x})$ for the region-level joint distribution, and letting $K = \dim(\mathbf{x})$, the conditional homogeneity assumption states that

$$\forall a, a' \in R: f_a(y|\mathbf{x}) = f_{a'}(y|\mathbf{x}) \quad (1.1)$$

Note that the conditional homogeneity assumption (1.1) implies that the area-level conditional density $f_a(y|\mathbf{x})$ coincides with the region-level density, since

$$\begin{aligned} f_R(y|\mathbf{x}) &= \sum_{a=1}^A \omega_a f_a(y|\mathbf{x}) \\ &= f_a(y|\mathbf{x}) \end{aligned} \quad (1.2)$$

where ω_a is the population share of area a , so that $\sum_{a=1}^A \omega_a = 1$.

$$\begin{aligned} f_a(y) &= \int_{\mathbb{R}^K} f_a(y, \mathbf{x}) d\mathbf{x} \\ &= \int_{\mathbb{R}^K} f_a(y|\mathbf{x}) f_a(\mathbf{x}) d\mathbf{x} \\ &= \int_{\mathbb{R}^K} f_R(y|\mathbf{x}) f_a(\mathbf{x}) d\mathbf{x} \end{aligned} \quad (1.3)$$

where the final equality follows from (1.2). Now $f_R(y|\mathbf{x})$ could potentially be estimated from the survey data, and $f_a(\mathbf{x})$ is readily available from the census data, so that the unobserved area-specific expenditure distribution

$f_a(y)$, which delivers the welfare measure, can be recovered from the combined survey and census data.

This is the essential insight that motivates the existing technique, which I will refer to as “ELL,” after the World Bank researchers who first explained it in Elbers et al. [2002, 2003a], though a less general version of the technique appeared earlier in Hentschel et al. [2000].

In section 2.5, I confirm that the above conditions - sparse data on y at the area level, plentiful data on \mathbf{x} at the area level, and intra-regional variation in the marginal distribution of \mathbf{x} - hold, establishing at least a minimal justification for the use of small-area estimation.

1.4 Conceptual Framework

1.4.1 Welfare Measures

Welfare measures, abstractly conceived, are functionals

$$W : \mathbb{M} \rightarrow \mathbb{R}$$

where \mathbb{M} is a set of probability measures. W may not be defined for *all* probability measures; for example, mean income obviously cannot be computed for those distributions for which the expectation does not exist. Elements of \mathbb{M} are (superpopulation) distributions of income or expenditure, y , where $y \in [0, \infty)$ or \mathbb{R} (if one wishes to work on a logarithmic scale).

At best, though, we could calculate the value of a function

$$W(\cdot) : \mathbb{R}^{N_a} \rightarrow \mathbb{R} \quad (1.4)$$

For example, the Foster-Greer-Thorbecke (FGT) class of welfare measures are given by

$$W_\alpha(y_1, y_2 \dots y_{N_a}) = \frac{1}{N_a} \sum_{i=1}^{N_a} \left[\left(1 - \frac{y_i}{z^*}\right)^\alpha \times \mathbf{1}_{\{y_i < z^*\}} \right] \quad (1.5)$$

where $z^* > 0$ is a “poverty line”. When $\alpha = 0$, we call W the *poverty headcount*; when $\alpha = 1$, we call the resulting measure the *poverty gap*.

Another example is the class of Atkinson welfare measures (for $\alpha \geq 0$):

$$W_\alpha(y_1, y_2 \dots y_{N_a}) = \frac{1}{N_a} \sum_{i=1}^{N_a} \frac{y_i^{1-\alpha}}{1-\alpha} \quad (1.6)$$

for $\alpha \neq 1$, and

$$\log(W_\alpha(y_1, y_2 \dots y_{N_a})) = \frac{1}{N_a} \sum_{i=1}^{N_a} \log(y_i) \quad (1.7)$$

for $\alpha = 1$.⁴ Deaton [1997] provides a detailed discussion of the implications of the choice of measure for the analysis of social welfare.

In this context, it makes sense to distinguish between the realised value of a measure and its expected value: even if we had a complete enumeration of every individual's expenditure - a vector \mathbf{y} - the world might have been otherwise, so, for example, there is a difference between the empirical mean \bar{y} and the population mean $\mathbb{E}[y]$, *even if the sample is exhaustive*.

1.4.2 "Idiosyncratic" vs Sampling Error

Suppose that the conditional distribution of expenditure and the household-level covariates (at the area or region level; given the conditional homogeneity assumption, the distinction is irrelevant) is fully characterised by some parameter vector β_0 . If we denote the conditional density of $\mathbf{y}|\mathbf{X}$ by $f(y_1, \dots, y_{N_a} | \mathbf{x}_1 \dots \mathbf{x}_{N_a})$, or, using vector notation for brevity, as $f(\mathbf{y}|\mathbf{X}_a, \beta)$, we have

$$\begin{aligned} \mathbb{E}[W(\mathbf{y})|\mathbf{X}_a] &= \int_{\mathbb{R}^{N_a}} W(\mathbf{y}_a) f(\mathbf{y}_a|\mathbf{X}_a, \beta) d\mathbf{y}_a \\ &= g(\beta, \mathbf{X}_a) \end{aligned} \quad (1.9)$$

where we have evaluated g at \mathbf{X}_a , the $(N_a \times \dim(\mathbf{x}))$ matrix of values of \mathbf{x}_i that is actually realised in area a . I will write μ_a for the true value of $\mathbb{E}[W(\mathbf{y})|\mathbf{X}_a]$, i.e. define

$$\mu_a = g(\beta_0, \mathbf{X}_a) \quad (1.10)$$

⁴Equivalently, of course, W_0 is the geometric mean:

$$W_0(y_1, y_2 \dots y_{N_a}) = \sqrt[N_a]{y_1 y_2 \dots y_{N_a}} \quad (1.8)$$

Imposing a parametric structure on the joint distribution $f(y, \mathbf{x})$ allows us to replace the estimation of a joint distribution - which is impossible to do nonparametrically given the sparseness of data on y - with the estimation of a parameter that supposedly describes $f(y|\mathbf{x})$. Then, given an estimate $\hat{\beta}$ of β , estimating μ_a reduces to the problem of evaluating $g(\cdot, \cdot)$ at $(\hat{\beta}, \mathbf{X}_a)$.

Notice that

$$W(\mathbf{y}_a) - \hat{\mu}_a = [W(\mathbf{y}_a) - \mu_a] + [\mu_a - \hat{\mu}_a] \quad (1.11)$$

The first term on the right, the difference between the value of W realised in the world and the conditional expectation of W is what has come to be called “idiosyncratic error”. The magnitude of this component of the error $W(\mathbf{y}_a) - \mu_a$ shrinks as N_a , the size of the small-area population, rises. This is because disturbances from distinct households tend to cancel one another out.

The second term above, the difference $\mu_a - g(\hat{\beta}, \mathbf{X}_a)$, arises because of sampling error in the estimation of $\hat{\beta}$, is called *model error*. The magnitude of the model error is determined by the sampling distribution of $\hat{\beta}$, which in turn is governed by s , the size of the sample over which the estimation is performed.

1.5 The Geometry of Small-Area Estimation

1.5.1 The Reaggregation Constraint

As I discussed in section 1.3, reasonably precise estimates of welfare measures are obtainable at the region level, since the available high-quality household datasets are often designed to be representative of the population at the region level.

For the following section, suppose that the welfare measure is additively separable over households. That is,

$$W_\alpha(y_1, y_2 \dots y_M) = \sum_{i=1}^M w(y_i) \quad (1.12)$$

for some function $w(\cdot)$ with scalar domain; the FGT, Atkinson and Generalised Entropy classes of welfare measures all satisfy the requirement (1.12).

This restriction has two convenient consequences. One is that the (conditional) expectation of W depends only on the marginal distribution of y_i and not on the covariance structure of the y 's. This is particularly computationally convenient since then the computation of an N_a -dimensional integral collapses into the computation of N_a one-dimensional integrals.

Secondly, additive separability provides a baseline check of any small-area estimates. Since expectations are linear, the regional welfare measure is a convex combination of the small-area estimates. Let $N_R = N_1 + \dots + N_A$ be the population of the region and, as above, let $\omega_a = N_a/N_R$ be the population share of area a . Since

$$W(y_1, \dots, y_{N_R}) = \sum_{a=1}^A \left[\sum_{i \in \text{area } a} w(y_i) \right] \quad (1.13)$$

we have, taking expectations, a linear constraint on the values of $\mu = (\mu_1 \dots \mu_A)$ that are consistent with the region-level survey data.

For example, the poverty headcount takes

$$w(y_i) = \frac{1}{N_R} \mathbf{1}_{\{y_i < z^*\}} \quad (1.14)$$

so that, letting \mathbf{y}_R denote the N_R -vector of incomes in the region and \mathbf{y}_a denote the N_a -vector of incomes in area a , we have

$$\begin{aligned} W(\mathbf{y}_R) &= \frac{1}{N_R} \sum_{i=1}^{N_R} \mathbf{1}_{\{y_i < z^*\}} \\ &= \sum_{a=1}^A \frac{N_a}{N_R} \left[\frac{1}{N_a} \sum_{i \in \text{area } a} \mathbf{1}_{\{y_i < z^*\}} \right] \\ &= \sum_{a=1}^A \omega_a W(\mathbf{y}_a) \end{aligned} \quad (1.15)$$

and hence (taking expectations conditional on some covariates \mathbf{x}):

$$\begin{aligned} \mu_R &= \mathbb{E}[\mathbf{1}_{\{y < z^*\}} | \mathbf{X}_R] \\ &= \sum_{a=1}^A \omega_a \mu_a \end{aligned} \quad (1.16)$$

where μ_R is the region-level headcount.

The above argument requires something perfectly natural of any candidate vector of small-area estimates: that separable measures, when estimated at the area level, should reaggregate up to the same value as the direct (region-level) estimate. Equation (1.16) constrains feasible estimates to lie in an $A - 1$ -dimensional (affine) subspace of \mathbb{R}^A .

If there are also range restrictions on the area-level estimates, we have even more information about true vector of small-area measures (μ_1, \dots, μ_A) . In the case of the FGT class of measures, each μ_a necessarily lies in the (closed) unit interval $[0, 1]$, so we must have (denoting by P the plane defined by equation 1.16)

$$(\mu_1, \dots, \mu_A) \in [0, 1]^A \cap P \quad (1.17)$$

We know that the set $[0, 1]^A \cap P \subset \mathbb{R}^A$ is nonempty, because $\mu_a = \mu_R$ for all a is certainly feasible: it satisfies the reaggregation constraint (1.16), since $\sum_{a=1}^A \omega_a = 1$, and $\mu_R \in [0, 1]$ so that $(\mu_R, \mu_R, \dots, \mu_R) \in [0, 1]^A$.

1.5.2 The Fundamental Problem

Ultimately the project of poverty mapping is to estimate the vector (μ_1, \dots, μ_A) . I have argued above that there are certain restrictions - both *a priori* and empirical - that we can impose on the possible estimates. Still, these restrictions allow for a significant variety of estimates. An example of this situation is depicted in Figure 1.2 below for $A = 2$, with $\omega_1 = 0.4$, $\omega_2 = 0.6$, and $\mu_R = 0.3$. The straight line in Figure 1.2 is the set of small-area estimates that are consistent with the reaggregation constraint.

Poverty maps need to identify a “sensible” choice of $(\hat{\mu}_1, \dots, \hat{\mu}_A)$. The problem, of course, is in interpreting “sensible”. To economists, a natural step would be to pose the choice of $(\hat{\mu}_1, \dots, \hat{\mu}_A)$ as an optimisation problem, much in the same spirit as maximum likelihood estimation. However, absent a probability model of the data-generating process $f(\mathbf{y}_a)$ for each area a , there is no way of choosing which vectors in the feasible set are more or less “likely”.

In chapter 3 I review the existing methodology by which almost all of the poverty maps in the literature have been estimated. This technique, as will

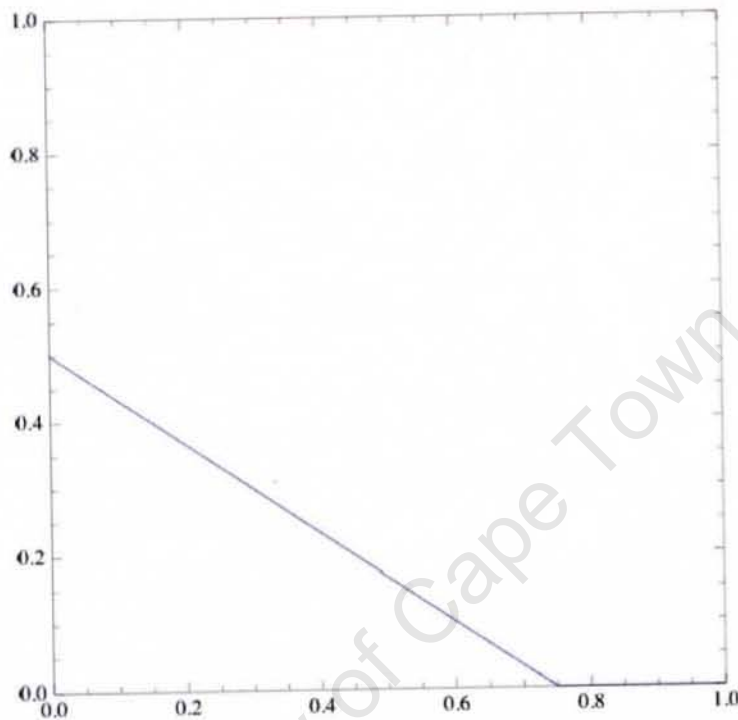


Figure 1.2: The Reaggregation Constraint: $\omega_1\mu_1 + \omega_2\mu_2 = \mu_R$. Only pairs (μ_1, μ_2) on the illustrated line are consistent with the region-level estimates.

become clear later, imposes a parametric model on the joint distribution (y, \mathbf{x}) , but it is unclear to me whether the estimates that result from this technique can be posited as the solutions to an optimisation problem with an objective function that is maximised at the true values (μ_1, \dots, μ_A) .

Chapter 2

The Data

2.1 Context: South Africa's Changing Administrative Geography

Apartheid, the legal structure of racial discrimination and segregation that was enforced in South Africa from 1948 until 1991, produced a dysfunctional system of overlapping administrative hierarchies. These parallel bureaucracies were created as a political conceit, to give substance to the white government's official claim that the different races should "develop separately".¹

When the three datasets (introduced below) were collected, South Africa was partitioned into 354 *magisterial districts*, as defined by the judiciary. Magisterial districts were nested in nine provinces. However, there were several parallel systems of local administration, a legacy of the previous racially segregated government system. For details, see Dube [2005]. In 1997, the democratically elected government consolidated these systems into a single sub-national administrative hierarchy, consisting of nine provinces, 47 *district councils* (most, but not all, of which are contained in a single province), and 283 *local municipalities*. Local municipalities, luckily, *are* nested in district councils.

The 10% census sample and the 1995 October Household Survey (described below) do have information on magisterial district, which allows me to attach (magisterial) district-level means to observations in the survey data, as

¹The National Party government of the time even went so far as to declare certain areas of the land - the so-called *bantustans* - "independent foreign countries," which were supposedly the "homelands" of the various ethnic groups of the black African population.

encouraged by the poverty-mapping literature.

2.2 1995 October Household Survey/Income and Expenditure Survey

Originally intended to provide a basis for inflation data, the Income and Expenditure Surveys are a series of household-level surveys, covering patterns of consumption and the composition of income. An IES has been collected by Statistics South Africa every five years since 1995. The 1995 IES was collected as the second phase of the OHS of the same year, by surveying the same households as had been selected for the OHS. Thus the IES/OHS is a merged dataset.

2.2.1 Sample Design

The OHSs were a series of household-level surveys - covering the labour market experiences of the population, migration, household welfare (access to amenities and goods ownership, for example), and other demographic information - that were collected annually from 1993 to 1999. Specifically, the population (as recorded in the 1991 census) was stratified by race, urban/rural category and province. Then, 3000 enumerator areas were sampled, and ten households were randomly chosen within each of the selected enumerator areas, making for a total sample of 30 000 households. Non-response was very low, with only 405 households refusing to cooperate. The final OHS sample thus contained 29 595 households, representing a total of 130 787 persons.

Because the poverty line in this paper depends only on total household welfare and not on per-capita equivalents, I use the logarithm of total monthly household consumption, as measured in the IES, as the dependent variable. The household covariates (education of members, demographic structure etc.) come from the OHS. Because the IES was conducted after the OHS (in December 1995), there was some attrition. Also, some households do not match between the two surveys; I therefore lose some observations in merging the IES and the OHS. Furthermore, I decided to drop the households with missing values for any of the variables in the subsequent analysis. In the end, I was left with a sample of 27 830 households, representing 122 607 individuals.

2.2.2 Province-Level Expenditure Distributions

Since the IES/OHS is designed to be representative at the province level, I estimate the poverty headcount for each province directly as the mean of a dummy variable (which is 1 for households with expenditure less than R800/month, and 0 otherwise), weighted by household size and the OHS sampling weights. The results are documented in Table 2.1.

Province	Headcount (Std. Error)
W Cape	0.11 (0.0097)
E Cape	0.452 (0.0125)
N Cape	0.352 (0.0287)
Free State	0.476 (0.0182)
KwaZulu-Natal	0.194 (0.0107)
North West	0.386 (0.0268)
Gauteng	0.066 (0.0083)
Mpumalanga	0.246 (0.0205)
Limpopo	0.353 (0.0208)

Table 2.1: Direct Estimates (from IES Data) of Headcount, by Province

To get a rough sense of the distributions at the region level, I plotted kernel density estimates of log consumption expenditure.² The resulting figures are below, in Figures 2.1 and 2.2.

As expected, the urban provinces (Gauteng and the Western Cape) have expenditure distributions that lie substantially to the right of the rest of those for South Africa's more rural provinces like the Eastern Cape and Limpopo. Naturally, this fact is reflected in the lower estimated headcounts for those provinces in Table 2.1.

²I used the default settings in STATA 9: an Epanechnikov kernel with bandwidth

$$h = 0.9 \min\{s(y), \text{IQR}/1.34\} \times n^{-1/5}$$

where $s(y)$ is the sample standard deviation, IQR is the sample interquartile range, and n is the sample size.

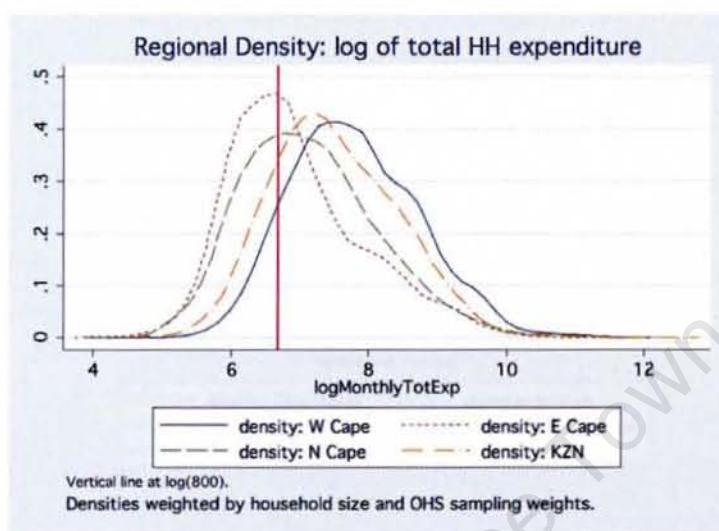


Figure 2.1: Kernel Density Plots - KwaZulu-Natal; Western, Eastern, and Northern Cape.

2.3 1996 Population Census

2.3.1 Sample Design

In 1998, Statistics South Africa released the 10% sample of unit records, which was a systematic sample of the full census data, after stratification on province, district council and local authority.³ This data was collected in October 1996, and was intended to be an exhaustive sample of all persons inside the borders of the Republic on Census night (October 9th – 10th).

The census collected information on households' demographic structure; on variables describing employment and labour market outcomes; and on their living conditions and other economic variables.

The public release of the census data includes the institutional population (persons in hospitals, prisons, boarding schools, workers' hostels, military

³Unfortunately, the geography information in the public release of the 1996 census does not conform to the new administrative divisions, even though the sampling process involves stratification on district council.

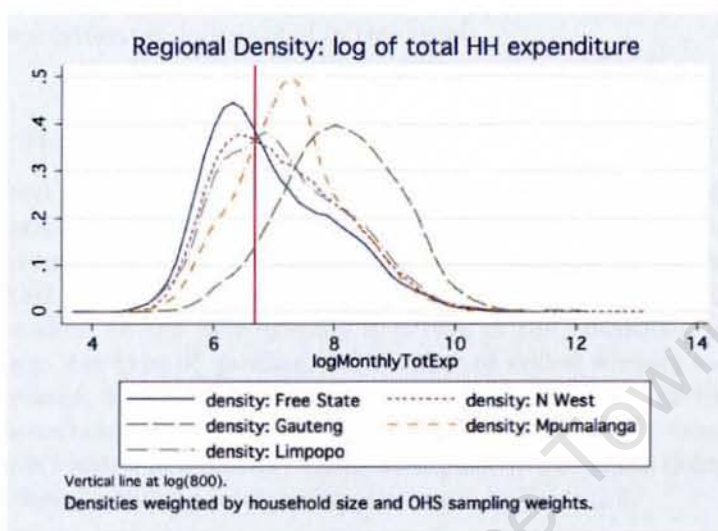


Figure 2.2: Kernel Density Plots - Free State, North West, Gauteng, Mpumalanga, and Limpopo.

barracks, etc.) only in the person-level data file. Such observations, lacking a clear analogue of the census' definition of "household," obviously failed to merge with the household-level data correctly. Hence, I have omitted them from my analysis. There were 12 995 such persons in the person-level dataset, out of a total 3 481 931 individuals in the 10% sample.

2.4 Data Construction and Cleaning

Since there is existing work on the small-area estimation of poverty maps in South Africa, and my aim in this thesis is to examine the methodology of poverty mapping, I attempt to recreate the data used in Alderman et al. [2002, 2000] so that my work at least engages the existing literature. In this section, I document the data preparation process. Of course, I may not have succeeded entirely in recreating the datasets they used. However, my aim in chapter 4 is methodological: I examine the sensitivity of the ELL technique to alternative specifications, rather than trying to replicate existing results. There may be differences in data cleaning and construction between this thesis and Alderman et al. [2002, 2000], but those differences cannot be responsible for the divergence *across specifications, on the same dataset* in

small-area estimates documented in this thesis.

2.4.1 Household-level Covariates

I examined the census and Income and Expenditure Survey (IES) metadata and identified all the variables that contained comparable information. In fact, this turns out to be the same as the set of variables used in Alderman et al. [2002, 2000]. There are 16 such household-level variables, comprised of information on the demographic structure of the household, economic status (e.g. the type of dwelling, the number of skilled workers resident in the household, whether the household owns a telephone), and on the nature of the household's neighborhood (e.g. urban/rural dummies, whether the dwelling is located in a former "tribal homeland"). The exact definitions of each of these household-level variables appears in Table 2.2.

2.4.2 District-level Covariates

I then computed the mean value of the above household-level variables, as well as of other indicators available in the census but not in the October Household Survey (OHS) - such as whether the household owns its dwelling - over each magisterial district in the *census* data. Since I have geographical information in both datasets, I was able to attach the area means to the IES observations. The construction of those census-only district means I computed are listed in Table 2.3.

I provide descriptive statistics on the covariates, as well as on the dependent variable, at the region level for both the IES and Census data in appendix B.

2.5 Is Small-Area Estimation Suitable?

Recall from section 1.3.2 that the core idea of small-area estimation is to deduce the joint distribution of (y, \mathbf{x}) in each area from the marginal distribution $f_a(\mathbf{x})$ by imposing a conditional homogeneity assumption, which I showed implies that $f_a(y|\mathbf{x}) = f_R(y|\mathbf{x})$ for all areas a in the region R .

As I argued in section 1.3.2, this approach is necessary because enough information on the expenditure distribution (the y 's) is available at the region

Variable	Definition or Comments
logHHsize	log(number of members of household).
aHH	Dummy: all household members African
wHH	Dummy: all household members white
fDw	Dummy: dwelling is house, apartment, retirement village; includes rooms in shared property (e.g. hostels)
rpP	total number of rooms/household size
sFac	Dummy: flush or chemical toilet, or pit latrine <i>with</i> ventilation (excludes non-ventilated pit latrines) on the same site as dwelling
elecL	Dummy: dwelling has electric lighting
rCol	Dummy: local authority removes refuse
hTel	Dummy: dwelling has fixed-line telephone in working order
nPrEd	Number of household members with complete primary education
nProf	Number of household members employed as professionals (ISCO 1-digit codes 2-3)
nSk	Number of household members employed as skilled workers (ISCO 1-digit codes 6-8)
fhHH	Dummy: household head is female
farm	Dummy: enumeration area is classified as "farm"
urban	Dummy: enumeration area is classified as "urban"
tribal	Dummy: enumeration area is classified as "tribal" (indicates former tribal authority)

Table 2.2: Variable Definitions - Household-Level Covariates

level, but not at the area level, while the census data will contain many observations on the covariates \mathbf{x} for all areas. For a nontrivial disaggregation of the expenditure distribution, the marginal distribution of the covariates has to differ between areas in a given region - we must have $f_a(\mathbf{x}) \neq f_{a'}(\mathbf{x})$ for at least one pair $a \neq a'$ - for if not, the areas are nothing but the region

Variable	Definition or Comments
waterServices	Dummy: household has piped water inside dwelling or on site
propertyOwnedByHH	Dummy: household owns dwelling

Table 2.3: Variable Definitions - District-Level Covariates

writ small.

In this section I confirm that these conditions - sparseness in y at the area level, large sample size in \mathbf{x} at the area level, and inter-area variations in the distribution of \mathbf{x} - do in fact hold for the three datasets that I use. Since - as I argue below - these conditions hold, a small-area estimation technique like ELL is appropriate, at least at a first glance. (Of course there are significant problems in the implementation of ELL, as I chronicle in Chapter 4, but those criticisms concern the exogeneity assumptions of ELL, and have little to do with the data itself.)

2.5.1 The Survey Data Is Indeed Sparse, and the Census Data Is Indeed Plentiful

Consider Table 2.4 below, which displays the number of observations at both the household and individual level in the two datasets for the Northern Cape province. (The number of individuals is actually irrelevant given that my choice of dependent variable is *total* household expenditure, but I include it for completeness nonetheless.)

The Northern Cape is not the worst province in terms of sparse IES/OHS data; the Eastern Cape, for example, has at least one magisterial district with no survey observations at all. Even so, there are several areas with fewer than 40 observations. A direct estimate of the headcount would be \hat{p}_a , the sample proportion of poor households. \hat{p}_a , as the sample mean (over s_a observations) of the Bernoulli(p_a) random variable $\mathbf{1}_{\{y_i < z^*\}}$, has variance

$$\text{Var}[\hat{p}_a] = \frac{p_a(1-p_a)}{s_a} \quad (2.1)$$

and hence the direct estimate has standard error

$$\text{se}[\hat{p}_a] = \frac{1}{\sqrt{s_a}} \sqrt{p_a(1-p_a)} \quad (2.2)$$

Magisterial District	IES		Census	
	Households	Persons	Households	Persons
Namakwaland	80	257	1486	5823
Calvinia	39	144	458	1520
Sutherland	29	94	94	278
Williston	30	83	108	348
Carnarvon	40	122	211	782
Prieska	50	249	420	1621
Britstown	40	140	124	507
Colesberg	60	232	332	1436
De Aar	50	167	573	2375
Hanover	30	136	94	392
Hopetown	30	104	255	1063
Noupoort	30	107	154	707
Philipstown	40	133	211	830
Richmond	39	154	136	533
Gordonia	138	603	2601	11337
Kenhardt	40	126	315	1175
Barkley-West	80	309	729	3214
Hartswater	50	155	841	2988
Herbert	70	364	484	2159
Warrenton	50	199	524	2111
Kimberley	149	555	4151	17504
Kuruman	57	170	541	1948
Postmasburg	78	321	1275	4891
Hay	40	154	246	1008
Fraserburg	30	71	95	351
Victoria-West	50	188	220	933
Total	1419	5337	16678	67834

Table 2.4: Sample Sizes - Northern Cape

Now, above we have $s_a = 29$ for Sutherland. For the purposes of illustration, suppose that the true headcount, p_a , is 30%. Then, a direct estimate based on the IES sample would have a standard error of

$$\frac{1}{\sqrt{29}} \times \sqrt{0.3 \times 0.7} \approx 0.085 \quad (2.3)$$

or 8.5%. Remember what the standard error represents: it is not an upper bound to the sampling error, but rather one-half of the range into which

about 68% of the estimates would fall (if \hat{p}_a were normally distributed). Obviously, the use of a central limit theorem as justification for a normal approximation is particularly implausible in this situation, but as a heuristic, it implies that nearly a third of the directly estimated headcounts will fall *outside* of the already very wide range from 21.5% to 38.5%. I consider this to be an unacceptably large margin for error, though of course others may differ.

By contrast, the number of census observations in even the smallest districts is substantially larger. Sutherland, again, has the fewest observations in the province, but the sample size in the census is 94, which (if we had expenditure data in the census) would deliver an estimate with standard error

$$\frac{1}{\sqrt{95}} \times \sqrt{0.3 \times 0.7} \approx 0.047 \quad (2.4)$$

or 4.7%, implying a far tighter band of estimates.

2.5.2 The Marginal Distribution of the Covariates Is Indeed Heterogenous

The premise of small-area estimation is that the inter-area heterogeneity in the distribution of y is attributable entirely to inter-area heterogeneity in the distribution of \mathbf{x} - that is the content of the conditional homogeneity assumption. Since the true data-generating process is unknown, we have no hope of verifying that the y -heterogeneity is *wholly* due to the heterogeneity in \mathbf{x} . However, we can at least verify that the distribution of the covariates \mathbf{x} is different across distinct areas in the region - that \mathbf{x} -heterogeneity at least *exists* - so that it is at least possible that y -heterogeneity is produced by \mathbf{x} -heterogeneity.

Below I will illustrate this heterogeneity using two particular correlates of household expenditure: a dummy for white households (defined as households with all members classified white), and (the natural logarithm of) household size.

First, I confirm that these two variables are indeed empirically significant (in the statistical sense) correlates of household expenditure. To do this I performed an OLS regression of log household expenditure on those two variables. The results are in Table 2.5 below, and indeed these two variables seem like good candidates for “determinants” of log expenditure - for

example, these two variables alone “explain” 46.6% of the variation in log household expenditure, and the estimated coefficient is highly statistically significant for both variables.

VARIABLES	logMonthlyTotExp
logHHsize	0.515*** (0.0332)
wHH	1.726*** (0.0499)
Constant	6.168*** (0.0476)
Observations	1419
R^2	0.466

Standard errors in parentheses
 *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Table 2.5: A Simple “Significance” Check of the Covariates

Next, I check that these two variables have different marginal distributions across different areas in the Northern Cape. Table 2.6 contains the (un-weighted) means of household size and the race dummy over each magisterial district in the province. Clearly, there are substantial differences between areas: the average household in Carnarvon, for instance, is nearly a full person smaller than the average household in Noupoort (the means are 3.7 and 4.6, respectively). The story is similar for the race dummy: the proportion of white households ranges from 31.2% in Kuruman to 10.5% in Herbert.

Given the strong statistical association between household expenditure and these covariates, as documented in Table 2.5, it seems at least plausible that a small-area estimation technique might be successful on this data.

Magisterial District	Household Size	Dummy: White Household
Namakwaland	3.919	0.153
Calvinia	3.319	0.207
Sutherland	2.957	0.277
Williston	3.222	0.241
Carnarvon	3.706	0.152
Prieska	3.860	0.148
Britstown	4.089	0.185
Colesberg	4.325	0.105
De Aar	4.145	0.227
Hanover	4.170	0.138
Hopetown	4.169	0.204
Noupoort	4.591	0.117
Philipstown	3.934	0.166
Richmond	3.919	0.169
Gordonia	4.359	0.186
Kenhardt	3.730	0.171
Barkley-West	4.409	0.096
Hartswater	3.553	0.222
Herbert	4.461	0.105
Warrenton	4.029	0.143
Kimberley	4.217	0.182
Kuruman	3.601	0.312
Postmasburg	3.836	0.237
Hay	4.098	0.138
Fraserburg	3.695	0.305
Victoria-West	4.241	0.209
Total	4.067	0.183

Table 2.6: Census Sample Means, Northern Cape

Chapter 3

The ELL Technique

3.1 The Basic Idea

There are two basic steps to the ELL technique, first introduced in Elbers et al. [2002, 2003a]. In the first stage, a model of the conditional distribution $y|\mathbf{x}$ must be estimated. Typically this is done by generalised least squares, although some papers use ordinary least squares. Indexing households by h and survey clusters by c , the feasible GLS estimation is performed by first estimating

$$y_{ch} = \mathbf{x}_{ch}\beta_0 + u_{ch} \quad (3.1)$$

over the *survey* observations at the *region* level. The residuals u_{ch} are typically presumed to obey a random-effects structure:

$$u_{ch} = \eta_c + \varepsilon_{ch} \quad (3.2)$$

with η_c independent of ε_{ch} . If this true, then the OLS residuals should be de-meaned over the survey clusters to form estimates of the cluster effect, η , and the household-specific disturbance, ε . To “allow” for heteroskedasticity in ε , a model of the squared residuals is then fitted, which yields an estimate of the household-specific variance for each census household and leads to “normalised” first-stage residuals $\hat{\varepsilon}_{ch}^*$. Typically the model is of logistic form with an upper bound set equal to the (arbitrary) level $1.05 \times \max_{c,h} \{\hat{\varepsilon}_{ch}^2\}$.

In the second stage, a simulated error term \tilde{u}_{ch} is drawn from the assumed distribution for each *census* household, which yields a complete census of imputed log expenditures as

$$\tilde{y}_{ch} = \mathbf{x}_{ch}\hat{\beta} + \tilde{u}_{ch} \quad (3.3)$$

for the r^{th} simulation draw.¹ The value of the welfare measure in each area a is then computed directly from the simulated values as $W(\tilde{\mathbf{y}}_a^r)$. The simulation step is repeated R times. The mean of $W(\tilde{\mathbf{y}}_a^r)$ over these simulations is $\hat{\mu}_a$, the ELL estimate of the (conditional expectation of) W in area a . The standard deviation of $W(\tilde{\mathbf{y}}_a^r)$ over the simulations is the ELL estimate of the standard error of $\hat{\mu}_a$.

Because $W(\cdot)$ is frequently nonlinear in \mathbf{y} , Elbers et al. [2003a] suggests integrating the estimated $\hat{\mu}_a$ over the sampling distribution of $\tilde{\beta}$. Since this is unknown, researchers hoping to use the technique must simulate draws $\tilde{\beta}$ from the asymptotic distribution of the first-stage estimators.

Effectively, ELL estimates are Monte Carlo integrals:

$$\begin{aligned}\hat{\mu}_a &= \frac{1}{R} \sum_{r=1}^R W(\tilde{\mathbf{y}}^r) \\ &\approx \int_{\mathbb{R}^K} \left[\int_{\mathbb{R}^{N_a}} W(\mathbf{y}) \hat{f}(\mathbf{y}|\mathbf{X}_a, \tilde{\beta}) d\mathbf{y} \right] \hat{f}^a(\tilde{\beta}|\mathbf{X}_R, \hat{\beta}) d\tilde{\beta}\end{aligned}\quad (3.4)$$

where $K = \dim(\beta_0)$, $\hat{f}(\mathbf{y}|\mathbf{X}_a, \tilde{\beta})$ is the estimated conditional density of log expenditure based on the parameter estimate $\tilde{\beta}$, and $\hat{f}^a(\tilde{\beta}|\mathbf{X}_R, \hat{\beta})$ is the (estimated) asymptotic sampling density of $\tilde{\beta}$. Both of these densities are obviously determined by the first-stage specification.

3.2 Estimators, Vagueness and Methodological Choice

I emphasize that at no point does the procedure outlined in the original papers Elbers et al. [2002, 2003a], which have become the methodological basis for this literature, insist on the use of any particular assumptions on functional form, error structure, estimation technique, or simulation procedure (i.e. whether to simulate distinct cluster effects for each census cluster, or for some higher level of aggregation). Consequently those papers do not describe an *estimator* in the technical sense (i.c. a measurable function of

¹A distribution for the residuals has to be chosen - more or less arbitrarily - by the researcher. Several papers use the empirical distribution of the first-stage residuals, but some authors use parametric distributions - typically the normal or t distributions (scaled to have the same variance as the first-stage residuals). In addition, the researcher must choose whether to simulate the "cluster effect" $\tilde{\eta}$ at the census cluster level or at a higher level.

the observed data).

Therefore, my results are vulnerable to the criticism that I have not calculated my estimates according to the *true* poverty-mapping methodology, but according to an apparently similar, though distinct, technique. In section 3.3, I describe the diversity of methodological choices consistent with the ELL technique and the associated computations in more detail, and I show that my choices in implementing the ELL technique are consistent with the most popular practices in the poverty-mapping literature.

Of course, the reader may judge for herself if the results are driven mostly by arbitrary methodological choices; but this is *exactly the point I attempt to make in chapter 4*.

3.3 Computational Details

Recall that there are two basic steps to the ELL technique: the estimation of a model of $y|\mathbf{x}$, and the numerical integration of $W(\cdot)$ with respect to the implied estimate of the conditional density. Below I describe in more detail how to perform these calculations, as well as briefly explaining (in section 3.3.4) the way that the standard errors of ELL estimates are calculated.

First, though, I introduce some notation. Echoing section 1.4.2, let $K = \dim(\mathbf{x})$, and define $g : \mathbb{R}^K \rightarrow \mathbb{R}$ by

$$\begin{aligned} g(\beta|\mathbf{X}_a) &= \mathbb{E}[W(\mathbf{y})|\mathbf{X}_a] \\ &= \int_{\mathbb{R}^{N_a}} W(\mathbf{y}_a) f(\mathbf{y}_a|\mathbf{X}_a, \beta) d\mathbf{y}_a \end{aligned} \quad (3.5)$$

where $f(\mathbf{y}|\mathbf{X}_a, \beta)$ is the true conditional density of the expenditure vector for area a .

If we simulate draws of $\hat{\beta}$ from its *true* sampling distribution $f(\cdot|\mathbf{X}_R, \beta_0)$, and the *true* conditional density of $\mathbf{y}_a|\mathbf{X}_a$ is actually in the parametric family described by β , we would get the estimate

$$\begin{aligned}
\hat{\mu}_a &= \frac{1}{R} \sum_{r=1}^R W(\tilde{\mathbf{y}}^r) \\
&\approx \int_{\mathbb{R}^K} g(\tilde{\beta}) f(\tilde{\beta} | \mathbf{X}_R, \beta_0) d\tilde{\beta} \\
&= \mathbb{E} \left[\mathbb{E} \left[W(\mathbf{y}) | \mathbf{X}, \hat{\beta} \right] | \mathbf{X} \right] \\
&= h(\beta_0 | \mathbf{X}_a, \mathbf{X}_R)
\end{aligned} \tag{3.6}$$

where I have stressed that the (density of) the sampling distribution of $\hat{\beta}$ depends on \mathbf{X}_R , the matrix of observed covariates from the *survey* data at the *region* level.

3.3.1 Assumptions

The ELL technique (as in Elbers et al. [2002, 2003a]) proceeds from the following assumptions:

ASSUMPTION 3.1 (Exogeneity). *The expenditure-generating process (for a given region) has a conditional mean linear in the covariates, \mathbf{x} :*

$$\begin{aligned}
y_i &= \mathbb{E}[y_i | \mathbf{x}_i] + u_i \\
&= \mathbf{x}_i \beta_0 + u_i
\end{aligned} \tag{3.7}$$

i.e. $\mathbb{E}[u_i | \mathbf{x}_i] = 0$ holds over i .

While of course the set of possible regressors is limited to those variables which are present in both the survey and the census data, exactly which variables should be included in \mathbf{x} is almost never discussed, as I document in section 3.4 below. To my knowledge, no poverty mapping paper even discusses the problem of consistently estimating β_0 (and some explicitly spurn identification).

ASSUMPTION 3.2 (Random Effects). *The error term (for a given region) is the sum of two independent components: a “location effect”, η_c , and a household-specific error (“idiosyncratic effect”), ε_{ch} :*

$$u_i = \eta_c + \varepsilon_{ch} \tag{3.8}$$

(where household i is the h^{th} one in cluster c).

ASSUMPTION 3.3 (Logistic-form Heteroskedasticity of ε).

$$\text{Var}[\varepsilon_i | \mathbf{x}_i] = A \left(\frac{\exp[\mathbf{z}_i \alpha]}{1 + \exp[\mathbf{z}_i \alpha]} \right) + B \quad (3.9)$$

for some $A > 0$, $B \geq 0$, $\alpha \in \mathbb{R}^p$ and a p -dimensional vector \mathbf{z} , which is a measurable function of \mathbf{x} .

Notice how the data-generating process described by (3.7) and (3.2) entails a homogeneity assumption: the differences in the distribution of y between areas is attributable entirely to the differences in the distribution of \mathbf{x} . While this is probably untrue, as Tarozzi [2008] argues, it is only important for my purposes insofar as it causes \mathbf{x} to be endogenous. In fact, if we think of area heterogeneity as arising from omitted area dummies and their interactions with the household covariates, then the first stage-estimation of ELL (in step 1 below) is analogous to an inconsistent random-effects estimation when a fixed-effects model is appropriate.

3.3.2 Recreate the Conditional Distribution

Given the above assumptions on the data-generating process, Elbers et al. [2003a] suggests estimating μ_a by the following steps:

1. Estimate β - by OLS or GLS - in the model

$$y_i = \mathbf{x}_i \beta + u_i \quad (3.10)$$

over the survey observations at the region level.

2. Use the covariates \mathbf{x} to get fitted values $\hat{y} = \mathbf{x} \hat{\beta}$ for all the *census* observations at the *area* level.
3. Add simulated error terms $\tilde{\mathbf{u}}^r$. Elbers et al. [2003a] suggests several options for the choice of distribution from which to draw $\tilde{\mathbf{u}}^r$. Under the assumption that $\eta \equiv 0$, one could use the empirical distribution of the OLS residuals.

If GLS estimation is used in step 1, one must estimate the cluster effects η and the household-specific disturbances ε by demeaning the OLS residuals over the survey clusters:

$$\hat{\eta}_c = \frac{1}{n_c} \sum_{h=1}^{n_c} \hat{u}_{ch} \quad (3.11)$$

$$\hat{\varepsilon}_{ch} = \hat{u}_{ch} - \hat{\eta}_c \quad (3.12)$$

Given the estimated household-specific errors, $\hat{\varepsilon}_i$, Elbers et al. [2003a] propose estimating the model of the heteroskedasticity (3.9) by imposing

$$A = 1.05 \times \max_{c,h} \{\hat{\varepsilon}_{ch}^2 : 1 \leq c \leq C, 1 \leq h \leq n_c\}$$

$$B = 0$$

which implies (writing σ_i^2 for $\text{Var}[\varepsilon_i | \mathbf{x}_i]$),

$$\log \left(\frac{\sigma_i^2}{A - \sigma_i^2} \right) = \mathbf{z}_i \alpha \quad (3.13)$$

Adding an exogenous error term ν to the right-hand side of (3.13) implies α can be estimated from the regression

$$\log \left(\frac{\hat{\varepsilon}_i^2}{A - \hat{\varepsilon}_i^2} \right) = \mathbf{z}_i \alpha + \nu_i \quad (3.14)$$

Rearranging (3.13) and neglecting terms which are $O(h^3)$ in a Taylor approximation of

$$\sigma_i = \sqrt{\frac{A \exp[\mathbf{z}_i \alpha] \exp[\nu_i]}{1 + \exp[\mathbf{z}_i \alpha] \exp[\nu_i]}} \quad (3.15)$$

about $\nu = 0$ yields the estimated household-specific standard deviation

$$\mathbb{E}[\sigma_i | \mathbf{z}_i] \approx \sqrt{\frac{A\lambda}{1+\lambda}} + \frac{1}{2!} \text{Var}[\nu] \cdot \left[\frac{1}{2(1+\lambda)^2} \sqrt{\frac{A\lambda}{1+\lambda}} \right] \cdot \left[\frac{1-\lambda}{1+\lambda} \sqrt{\frac{A\lambda}{1+\lambda}} - \frac{1}{2} \right]$$

$$= \hat{\sigma}_i \quad (3.16)$$

where $\lambda = \exp[\mathbf{z}_i \hat{\alpha}]$.²

²Some authors expand σ^2 about $\nu = 0$, from which follows the formula:

$$\hat{\sigma}_i^2 \approx \frac{A\lambda}{1+\lambda} + \frac{1}{2!} \text{Var}[\nu] \left(\frac{A\lambda(1-\lambda)}{(1+\lambda)^3} \right)$$

for the estimated idiosyncratic variance - for example, Lanjouw et al. [2007], but this is incorrect since the nonlinearity of $\sigma = \sqrt{\sigma^2}$ means that $\sqrt{\mathbb{E}[\sigma^2]} \neq \mathbb{E}[\sigma]$, even though it is the latter that we need to use in standardizing the residuals.

Next, one normalizes the estimated idiosyncratic errors by dividing by the respective $\hat{\sigma}_i$ and demeaning them over the s survey observations, i.e.

$$\hat{\varepsilon}_i^* = \frac{1}{\hat{\sigma}_i} \hat{\varepsilon}_i - \frac{1}{s} \sum_{i=1}^s \frac{\hat{\varepsilon}_i}{\hat{\sigma}_i} \quad (3.17)$$

Within the GLS framework, one can employ the presumed random-effects structure of the data by simulating η and ε^* separately (since they are assumed independent of one another), either from the empirical distributions or from a parametric distribution scaled to have the same variance as the empirical distributions. Elbers et al. [2003a] suggests using “standardized normal, t , or other distributions” for this purpose.

Regardless of the distribution from which the disturbances $\tilde{\eta}_c$ and $\tilde{\varepsilon}_{ch}^*$ are drawn, one has to choose the level of aggregation at which to simulate η ; choosing the cluster level assigns a randomly drawn $\tilde{\eta}_c^r$ to each census household n in cluster c . On the other hand, if one believes that there are “location effects” that apply at a higher level than the cluster, one could choose to assign the same $\tilde{\eta}^r$ to each household in a larger group, such as the small area level or at some intermediate level of aggregation, depending on the geographical information available in the census.

3.3.3 Integrate With Respect to the Conditional Distribution

- Repeat step 3 R times, obtaining R complete censuses of log expenditure,

$$\tilde{\mathbf{y}}^r = \mathbf{x}\hat{\beta} + \tilde{\mathbf{u}}^r$$

- Perform Monte Carlo integration on $W(\mathbf{y}_a)$, thus calculating

$$\begin{aligned} \tilde{\mu}_a &= \frac{1}{R} \sum_{r=1}^R W(\tilde{\mathbf{y}}_a^r) \\ &\approx \int_{\mathbb{R}^{N_a}} W(\mathbf{y}) \hat{f}(\mathbf{y}|\mathbf{X}_a, \hat{\beta}) d\mathbf{y} \\ &= \hat{g}(\hat{\beta}|\mathbf{X}_a) \end{aligned} \quad (3.18)$$

which is our estimate of μ_a at the “area” level. Here, $\widehat{f}(\mathbf{y}|\mathbf{X}_a, \widehat{\beta})$ is the density from which the simulated values $\widetilde{\mathbf{y}}^r$ have been drawn.³

6. Because g is, in general, a nonlinear function of β , there would be some bias associated with the evaluation of $g(\widehat{\beta})$, *even if we knew the true conditional density* $f(\mathbf{y}|\mathbf{X})$. Elbers et al. [2003a] suggests that “... using simulation to integrate over the model parameter estimates $[\widehat{\beta}]$... yields an unbiased estimator.” That means that we should calculate the r^{th} imputed value of (log) expenditure for household i as:

$$\widetilde{y}_i^r = \mathbf{x}_i \widetilde{\beta}^r + (\widetilde{\eta}_i^r + \widehat{\sigma}_i \varepsilon_i^{*,r}) \quad (3.20)$$

Of course, in practice we have to use the asymptotic sampling distribution

$$\widetilde{\beta} \sim N(\widehat{\beta}, \widehat{\text{aVar}}(\widehat{\beta})) \quad (3.21)$$

which we get from the first-stage regression (3.1). Denote by $\widehat{f}^a(\cdot|\mathbf{X}_R)$ the density of the presumed asymptotic sampling distribution of $\widetilde{\beta}$; then what can actually be calculated is not $h(\widehat{\beta}|\mathbf{X}_a, \mathbf{X}_R)$ as in (3.6), but

$$\begin{aligned} \widehat{\mu}_a &= \frac{1}{R} \sum_{r=1}^R W(\widetilde{\mathbf{y}}^r) \\ &\approx \int_{\mathbb{R}^k} \widehat{g}(\widetilde{\beta}) \widehat{f}^a(\widetilde{\beta}|\mathbf{X}_R, \widehat{\beta}) d\widetilde{\beta} \\ &= \widehat{h}(\widehat{\beta}|\mathbf{X}_a, \mathbf{X}_R) \end{aligned} \quad (3.22)$$

3.3.4 Estimate The Standard Errors

7. The standard error of $\widehat{\mu}$, Elbers et al. [2003a] suggests, should be estimated by the standard deviation of the simulated $W(\widetilde{\mathbf{y}}_a^r)$ over the R simulations:

³Conditional on \mathbf{X}_a , there is a one-to-one correspondence between densities for \mathbf{u} and densities for \mathbf{y} :

$$f_{\mathbf{y}}(\mathbf{y}|\beta, \mathbf{X}_a) = f_{\mathbf{u}}(\mathbf{y} - \mathbf{X}_a\beta) \quad (3.19)$$

which we easily obtain from a change of variables $\mathbf{u} = \mathbf{y} - \mathbf{X}_a\beta$.

$$\widehat{\text{se}}(\widehat{\mu}_a) = \sqrt{\frac{1}{R} \sum_{r=1}^R (W(\tilde{y}_a^r) - \widehat{\mu}_a)^2} \quad (3.23)$$

If $\widehat{\mu}_a$ is consistent for μ_a - remember, in this context, this means that $\widehat{\mu}_a \xrightarrow{p} \mu_a$ as the population size, N_a , and as the survey sample size, s , grow without bound - then, according to Elbers et al. [2003a], (3.23) is a consistent estimate of

$$\sqrt{\text{Var}[W(\mathbf{y}_a)|\mathbf{X}_a, \beta_0] + \text{Var}[g(\widehat{\beta})|\mathbf{X}_a]} \quad (3.24)$$

where we have neglected the “computational error” associated with numerical integration, since this error can be made arbitrarily small by choosing R as large as necessary. The variances in (3.24) are with respect to the joint variability in the superpopulation (over alternate realisations of the population) and in the survey sample (for a given population).

3.4 What Can Properly Be Considered An “ELL” Estimate?

Since my aim in this paper is to examine the sensitivity and consistency of the estimates produced by the ELL technique, I should ensure that my calculations are actually “ELL” estimates. Given the diversity of possible implementations allowed by the original paper Elbers et al. [2003a] I find it impossible to say definitively whether I have actually implemented the technique that has come to be called “ELL” or merely a similar, but distinct, technique. Instead, I have tried to ensure that my calculations conform to the standards of the existing literature.

I reviewed some of the papers in the poverty mapping literature and tabulated their authors’ choices with respect to the choice of first-stage estimation method, the distribution from which to draw \tilde{u} , the level at which to apply the “location effect”, and the criteria used in the specification of the first-stage model. The results are below, in Table 3.1.

In my reading of this literature, the primary requirement of the first-stage model appears to be that it should have “predictive power,” which has been interpreted by the authors of the method themselves and the World Bank

poverty Mapping Team as “high first-stage R^2 ”.

For example, we read in Lanjouw et al. [2007]:

OLS Regression results from the first-stage models are given in Appendix 2, Tables A1-A10. Across the ten pseudo-surveys used here, the R^2 ranges from 0.415-0.53 (see Table 1). The explanatory power of the models in this analysis is in the general range of models from past applications. The R^2 for models for particular strata ranged from 0.45 to 0.77 in Ecuador . . . The explanatory power achieved with the PROGRESA models is rather good given that the households in the PROGRESA communities are more homogenous than those within a stratum in a typical application.

In fact, some authors go so far as to explicitly dismiss concerns about the identification of β_0 , as in Alderman et al. [2002]:

The explanatory power of the nine regressions ranged from an adjusted R^2 of 0.47 (Eastern Cape) to 0.72 (Free State), with the median adjusted R^2 equal to 0.64.

. . . Finally, note that from a methodological standpoint it does not matter whether these variables are exogenous.

or in Minot and Baulch [2005]

Because our main interest is predicting the value of $\ln(y)$ rather than assessing the impact of each explanatory variable, we are not concerned about the possible endogeneity of some of the explanatory variables.

The implicit argument here seems to be that if the first-stage model “predicts” y well in-sample, as measured by the R^2 , then

- (a) it will predict y well out-of-sample too, and so
- (b) the simulated \tilde{y} will be about right, at least over many simulations, and thus
- (c) $\hat{\mu}_a = \hat{h}(\hat{\beta}|\mathbf{X}_a, \mathbf{X}_R)$ will be close to $\mu_a = g(\beta_0, \mathbf{X}_a)$.

I do not think that this argument really stands up to scrutiny. For one thing, area heterogeneity (i.e. intra-regional differences in β_0) will tend to undermine the step from (a) to (b), since the first-stage model may perform poorly in some areas but well on aggregate. Secondly, even supposing that (conditional) area homogeneity holds, if \mathbf{x} is endogenous, the first-stage model will not yield consistent estimates of β_0 , and then there is no guarantee that $\hat{\mu}_a$ is consistent for μ_a in any area a .

Most importantly, though, a high first-stage R^2 is no guarantee of the consistent estimation of β_0 . Unfortunately, as I argue in section 4.3, it is consistent estimation of β_0 , not the quality of the in-sample prediction, that matters for the accuracy of the second-stage estimates.

Although some papers assert that $\mathbb{E}[u_i|\mathbf{x}_i] = 0$ holds over i - such as Elbers et al. [2003a, 2008] - most do not. Yet, there is little or no attention devoted to building a case for the suitability of the first stage model of consumption (or income) in any of the papers in this literature, as can be seen in Table 3.1.

Paper	Country	First Stage	Errors	Level of $\tilde{\eta}$	Num. Reps	Specification Criteria
Elbers et al. [2003a, 2002]	Ecuador	GLS	normal, $t(5)$, empirical	not stated	not stated	not stated; high R^2 lauded
Hentschel et al. [2000]	Ecuador	OLS	normal ⁴	-	∞	not stated; high R^2 lauded
Alderman et al. [2002, 2000]	South Africa	GLS	normal	cluster	100	not stated; high R^2 lauded
Lanjouw et al. [2007]	Mexico	GLS	t , empirical	cluster	not stated	not stated; high R^2 lauded
Minot and Baulch [2005]	Vietnam	OLS	normal	-	∞	not stated; high R^2 lauded
Gibson et al. [2005]	Papua New Guinea	GLS	unclear ⁵	unclear	100	unclear; some model selection ⁶

Continued on next page...

⁴This paper effectively sets $R = \infty$, since for homoskedastic normal errors, there is an exact formula for (the conditional expectation of) the headcount.

⁵"Estimates of the distributions ... are obtained from the residuals ... and from an auxiliary equation that explains the heteroscedasticity in the household-specific part of the residual." - Gibson et al. [2005], p. 7

⁶To quote the paper, "The particular specification of the model resulted from a detailed model discovery process, with many sensitivity checks ... Briefly, the modelling started just with household characteristics, restricting it to those for which there were also variables available in the Census. After removing irrelevant variables the model was augmented ..." I can only assume the authors mean "statistically insignificant at a conventional significance level" by "irrelevant," which means that some variant of backward selection was used to arrive at a final specification.

Table 3.1 (continued from previous page)

Paper	Country	First Stage	Errors	Level of $\tilde{\eta}$	Num. Reps	Specification Criteria
Baschieri et al. [2005]	Azerbaijan	GLS	unclear ⁷	unstated	100	informal ⁸
Elbers et al. [2008]	Brazil	GLS	unclear	cluster, area	not stated	not stated; high R^2 lauded
Baschieri and Falkingham [2005]	Tajikistan	GLS	unclear	unstated	100	informal ⁹
Simler and Nhate [2002]	Mozambique	GLS	t	cluster	100	stepwise regression
Carletton et al. [2007]	Albania	GLS	not stated	not stated	100	not stated
Fujii [2003]	Cambodia	GLS	empirical	cluster	100	informal ("reasonable fit")
Healy et al. [2003], Jitsuchon and Richter [2007]	Thailand	GLS	normal, t , empirical	cluster	not stated	stepwise regression
Lanjouw [2004], Litvack [2007]	Morocco	GLS	not stated	not stated	not stated	not stated
Vishwanath and Yoshida [2007]	Sri Lanka	not stated	not stated	not stated	not stated	not stated
Ndeng'e et al. [2003]	Kenya	GLS	not stated	not stated	not stated	stepwise regression

Continued on next page...

⁷ "For each household we draw simulated disturbance terms . . . from their corresponding distribution[s]."

⁸ "In some strata, where the selected variables on the strict test of comparability did not yield a reasonable high R -square, the criteria for selection of the regression variables were relaxed. . . The final specification included only those variables that were significant at least at 90 per cent level and the quarterly dummy variables."

⁹ "In some strata, where the selected variables did not yield a reasonable R -square, the criteria for selection of the regression variables were relaxed . . . To improve the explanatory power of the consumption model [it] was decided to include both census mean variables and some selected environmental variables."

Table 3.1 (continued from previous page)

Paper	Country	First Stage	Errors	Level of $\tilde{\eta}$	Num. Reps	Specification Criteria
Elbers et al. [2007]	Ecuador; Madagascar; Cambodia	GLS	normal, t	cluster	not stated	not stated

Table 3.1: Selected Poverty Mapping Papers and their (Authors') Methodological Choices

3.4.1 Implementing ELL Requires Arbitrary Choices

There are several points at which the method outlined in Elbers et al. [2003a] allows for the individual researcher's discretion. Specifically, anyone hoping to construct a poverty map by this method must choose

- (a) a first-stage estimation technique;
- (b) a distribution from which to draw the residuals $\tilde{\mathbf{u}}^r$;
- (c) as part of the decision in (b), the level (cluster, area, or some intermediate level of aggregation) at which to apply the simulated "cluster effect" $\tilde{\eta}_c$;
- (d) if GLS is chosen in the first-stage, and if the empirical distribution of the residuals is chosen in (b), whether to draw the standardized household residuals $\tilde{\varepsilon}^*$ from the clusters corresponding to the simulated cluster effects $\tilde{\eta}$ or from the full distribution of the (cluster-demeaned) residuals;
- (e) the number of simulations, R ;
- (f) exactly which covariates \mathbf{x} to use in the first-stage "prediction model".

Given the breadth of discretion one must exercise before calculating $\hat{\mu}_a$, it is easily possible for two different researchers to obtain different estimates, even if they use the same data, the same random-number generator with the same seed value, and the same number of repetitions R .

3.4.2 Implementation

Apart from the specification of the first-stage model, I kept the following choices constant across all estimations that follow, in Chapter 4:

- (a) I used GLS estimation, first obtaining an estimate of $\hat{\beta}$ by OLS and then estimating the cluster effects $\hat{\eta}$ and $\hat{\varepsilon}$ from the resulting residuals.
- (b) I drew both the cluster effects and the standardized household errors from their respective empirical distributions. For the heteroskedasticity model, I chose throughout to use all the household-level variables in \mathbf{x} .
- (c) I simulated the cluster effect $\tilde{\eta}_c$ at the area level (magisterial district). According to Lanjouw and Ravallion [2006], doing so “allows” for high-level spatial correlation.
- (d) I chose to draw the household idiosyncratic error for census households from the set of normalised first-stage residuals that correspond to the survey cluster from which their simulated cluster effect, $\tilde{\eta}_c$, was drawn. According to Elbers et al. [2003a], this approach “allows for nonlinear relationships between location and household unobservables.”
- (e) I used 100 simulations to perform the Monte Carlo integration.

The specification \mathbf{x} is, in my reading of the poverty-mapping literature, never motivated. However, the implicit criterion used in almost all of the papers in this literature is that the first-stage regression model (or the OLS stage of feasible GLS) should have a “high” R^2 statistic, though cutoff values are never explicitly stated. Nonetheless, I respect this convention: as I document in the appendix, section A.1, no estimate in this paper is derived from a first-stage model with an R^2 lower than 0.442. Even that value is unusually low: the vast majority of the first-stage models I estimated returned R^2 -statistics higher than 0.5.

There is one other caveat about the models used to “predict” consumption in the poverty mapping literature: that district-level means (which can be obtained from the census data) be included as regressors. A spate of papers (see Elbers et al. [2003b, 2002], Lanjouw et al. [2007], Lanjouw and Ravallion [2006]) by the creators of the ELL technique emphasize that area means should be included in the first-stage regression to “capture” cluster level effects. I follow their instructions: every estimate in this paper is based on a

first-stage specification that includes at least 10 area-level means. I therefore consider the estimates that follow in this thesis to be admissible in terms of the implicit criteria of the poverty-mapping literature.

University of Cape Town

Chapter 4

Weaknesses of ELL

I calculate the poverty headcount using the 1995 IES as the survey dataset and the 1996 South African population census by the ELL technique, using the specific methodological choices described in section 3.4.2. For comparability with the results of Alderman et al. [2002, 2000], I use the following poverty line:

A person is poor if they live in a household with total expenditure less than R800/month.

Say there are K_H household-level covariates and K_D area-level means, and we pick k_H and k_D of each. Then there are

$$\binom{K_H}{k_h} \times \binom{K_D}{k_d}$$

possible choices of first-stage specification.¹

I sample 50 such specifications at random and compute $\hat{\mu}_a$ (for each a) given each specification. I generate random first-stage specifications by drawing a random subset of 75% of the possible household covariates and 75% of the district-level mean variables. For comparison, I also calculated estimates of the headcount using every variable in the dataset that was not dropped due to collinearity. I call this last specification the “maximal model”.

¹In practice, this varies between regions because some variables end up getting dropped in some provinces but not in others. For example, in the Western Cape $K_D = 18$, $K_H = 15$, so with $k_H = \text{round}(0.75 \times K_H) = 11$ and similarly for $k_d = 14$, we have $1365 \times 3060 = 4\,176\,900$ possible specifications.

Space constraints prevent me from displaying all of the results in this thesis. I document those results not displayed below in appendix A.

4.1 Sensitivity to Specification

4.1.1 Point Estimates: Small Areas

In Table 4.1, I present summary statistics for the magisterial districts in the North West province over the 50 randomly generated specifications; I ranked the areas in descending order of the estimated headcount (under the maximal model). I want to highlight two features of the distributions of estimates over the different specifications.

Firstly, the range of estimates that can be obtained is very large. For Kudumane, for example, one specification leads to a low (by South African standards) headcount of 27.8%, while another specification leads to the spectacularly high headcount of 75.9%. By a judicious choice of specification, a researcher could throw over 48% of the residents of this area into (or out of) poverty. Although Kudumane is the worst example of this sensitivity in the North West province, it is not without peer. Even the Brits district, which has the narrowest range of estimates in the province (17%), the interquartile range is a substantial 4.9%.

Secondly, the instability is not merely an artifact of a few poorly chosen models. For most of the districts in Table 4.1, the interquartile range of the headcount estimates is high too, generally on the order of 8%, but for several areas it is above 10%. To see this, look at Figure 4.1, where I plot kernel density estimates of the distribution of headcount estimates for selected areas over the 50 random specifications. A casual glance at Figure 4.1 indicates that it is easily possible to obtain very different estimates of the headcount just by picking different specifications.

Magisterial District	HC (maximal)	Mean	Min	Max	Range	IQR
Phokwani	0.735	0.540	0.264	0.738	0.474	0.156
Kudumane	0.706	0.565	0.278	0.759	0.481	0.126
Wolmaransstad	0.609	0.607	0.480	0.727	0.248	0.082
Huhudi	0.590	0.577	0.328	0.744	0.416	0.087
Schweizer-Reneke	0.588	0.668	0.554	0.761	0.207	0.069

Continued on next page...

Table 4.1 (continued from previous page)

Magisterial District	HC (maximal)	Mean	Min	Max	Range	IQR
Mmabatho	0.569	0.488	0.224	0.659	0.435	0.130
Vryburg	0.522	0.377	0.219	0.515	0.297	0.083
Lichtenburg	0.481	0.445	0.309	0.586	0.278	0.096
Ventersdorp	0.429	0.390	0.284	0.570	0.286	0.063
Mankwe	0.371	0.420	0.220	0.582	0.362	0.111
Potchefstroom	0.364	0.315	0.208	0.399	0.191	0.076
Madikwe	0.349	0.386	0.214	0.589	0.375	0.121
Christiana	0.328	0.345	0.226	0.454	0.229	0.073
Brits	0.291	0.272	0.205	0.378	0.173	0.049
Delareyville	0.288	0.410	0.276	0.587	0.311	0.067
Ga-Rankuwa	0.241	0.279	0.208	0.383	0.174	0.052
Temba	0.231	0.279	0.172	0.556	0.384	0.074
Klerksdorp	0.221	0.229	0.120	0.352	0.232	0.053
Rustenburg	0.193	0.306	0.156	0.413	0.257	0.056

Table 4.1: Estimates Over 50 Random Specifications, North West

4.1.2 Intra-Regional Rankings

The instability of the headcount estimates cannot be blamed on a rank-preserving region-wide shift in the estimated headcount. To see this, I compared the rankings of magisterial districts within the province across specifications. The ranges of rankings obtained from the various specifications are as dramatic as those for the point estimates. The ranges observed in Table 4.2, for many of the districts, imply that mere specification choice can not merely shift, but practically reverse the relative rankings of the areas.

Magisterial District	Rank (maximal)	Mean	Min	Max	Range	IQR
Phokwani	1	4.9	1	12	11	3
Kudumane	2	4.0	1	11	10	3
Wolmaransstad	3	3.3	1	8	7	2
Huhudi	4	3.5	1	9	8	2
Schweizer-Reneke	5	1.5	1	5	4	1
Mmabatho	6	6.7	2	17	15	3
Vryburg	7	11.2	4	18	14	4

Continued on next page...

Table 4.2 (continued from previous page)

Magisterial District	HC (maximal)	Mean	Min	Max	Range	IQR
Lichtenburg	8	7.8	2	15	13	4
Ventersdorp	9	10.1	5	15	10	3
Mankwe	10	9.3	5	16	11	2
Potchefstroom	11	13.9	8	19	11	4
Madikwe	12	11.1	5	19	14	4
Christiana	13	12.7	6	19	13	3
Brits	14	16.1	8	19	11	3
Delareyville	15	9.4	4	16	12	3
Ga-Rankuwa	16	16.0	9	19	10	2
Temba	17	15.5	3	19	16	5
Klerksdorp	18	18.0	14	19	5	2
Rustenburg	19	14.8	11	19	8	3

Table 4.2: Within-Province Rankings Over 50 Random Specifications, North West

Consider the Temba district, for example. In a province of only 19 magisterial districts, a careful choice of specification could make Temba appear either relatively very well-off (the least poor area in the province), or bitterly poor (the third poorest). Again, Temba is not atypical; for 14 out of the 19 areas in Table 4.2, the range of the rankings over the 50 random specifications is greater than 10. This means that for such an area, there is a pair of specifications j, j' such that the ELL estimate under j puts the area in the poorest half of the province; under the specification j' , the area would be considered in the richest half of the intra-provincial ranking.

4.1.3 Point Estimates: (Reaggregated) Regional

Since the headcount is additively separable, I reaggregated the estimated headcounts in each area, weighting by the population size of each, to obtain the implied regional headcount for each specification. This provides a direct check of the reliability of the ELL estimates, since the IES data is representative at the provincial level.² Summary statistics on the distribution of the

²As mentioned in Chapter 2, the sampling frame of the 1995 IES/OHS was based on the 1991 population census, which is another possible source of difference between the estimates, since the inter-area distribution of the covariates as measured in the 1996 census may have changed significantly in the intervening five years. This is a possibility

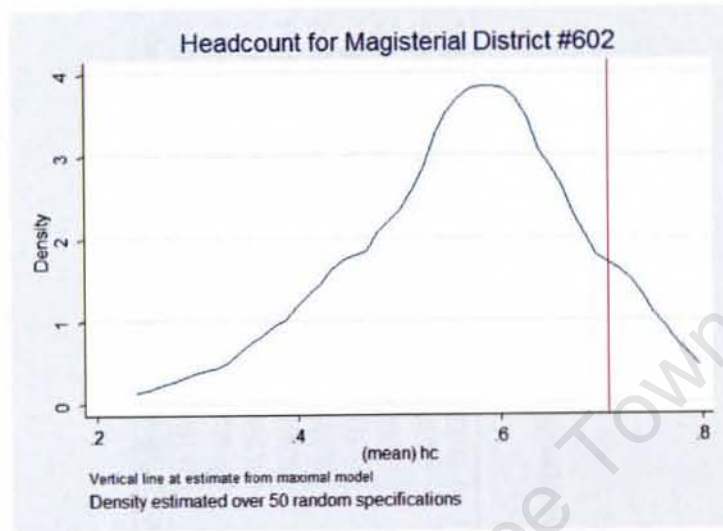


Figure 4.1: Distribution of Estimates for Kudumane (North West) Over 50 Random Specifications

implied provincial headcount for each of the nine provinces appear in Table 4.3; for comparison, I include the headcount estimates from the IES data, adjusting the standard errors for the clustered sample design.

that I ignore.

Province	HC (maximal) ^a	Direct Estimate (std. error)	Mean	Std. Dev.	Min	Max	Range	IQR
W Cape	0.0982	0.11 (0.0097)	0.1042	0.0188	0.077	0.1503	0.0733	0.0224
E Cape	0.446	0.452 (0.0125)	0.4536	0.0252	0.4246	0.5113	0.0867	0.0445
N Cape	0.2737	0.352 (0.0287)	0.2889	0.0241	0.2545	0.3697	0.1152	0.0277
Free State	0.4056	0.476 (0.0182)	0.4092	0.0208	0.3751	0.4684	0.0933	0.0273
KwaZulu-Natal	0.2105	0.194 (0.0107)	0.2074	0.0142	0.1839	0.2489	0.0649	0.0173
North West	0.3756	0.386 (0.0268)	0.3759	0.0384	0.2898	0.4612	0.1713	0.0522
Gauteng	0.0992	0.066 (0.0083)	0.0958	0.0102	0.0703	0.1162	0.0458	0.0123
Mpumalanga	0.2497	0.246 (0.0205)	0.2487	0.0171	0.2186	0.2895	0.0709	0.0257
Limpopo	0.332	0.353 (0.0208)	0.325	0.0199	0.2799	0.3757	0.0958	0.0244

Table 4.3: Implied Regional Headcount Over 50 Random Specifications, by Province

^aMaximal specification.

As Table 4.3 shows, the ELL estimates contradict the direct (IES) estimates, at least for some specifications. For example, the direct estimate of the headcount for the Eastern Cape is 45.2%. This is very close to the same as the mean of the implied ELL estimates over the random specifications, 45.4%. However, the ELL estimates range as high as 51.1%, more than six standard deviations above the IES estimate. For the Free State, the average ELL estimate is 40.9%, while the IES data suggests that the headcount is substantially higher - 47.6%. The worst performer, though, is Gauteng: the IES estimate, 6.6%, is entirely outside of the range of ELL estimates. The *lowest* ELL estimate of Gauteng's headcount, 7.0%, is about half a standard deviation higher than the IES estimate.

Though the reaggregated estimates have narrower ranges than the area-specific ones, there is still substantial variation across specifications. This is evident in Table 4.3. At the time of the 1996 Census, Gauteng had a population of approximately 6.9 million. A shift of 4.5% in Gauteng's headcount estimate would therefore imply the transition of about 310 500 persons in (or out) of poverty; and Gauteng is the *least* sensitive of the provinces.

In Figure 4.2, I display the density of the implied headcount for the Eastern Cape over the randomly generated specifications. For the Eastern Cape, specification choice is powerful enough to either throw the equivalent of a mid-sized city - over half a million people - into poverty, or to lift them out of it. (The Eastern Cape had a population of approximately 6.1 million at the time.)

I remind the reader that *all* of these estimates are based on first-stage models that have "high predictive power," and that the poverty mapping literature has almost universally adopted this informal criterion as its sole methodological principle. In section 4.3.1, I explain why this criterion is inadequate, and I show how high R^2 values can coexist with very poor models (in the sense of consistent parameter estimation).

4.2 Finite-Sample Bias

4.2.1 Existence

The differences documented above are so large - for some areas, the range of possible estimates is on the order of 0.5 or even larger - that it seems at least plausible that distinct specifications are not centred on the same values. If

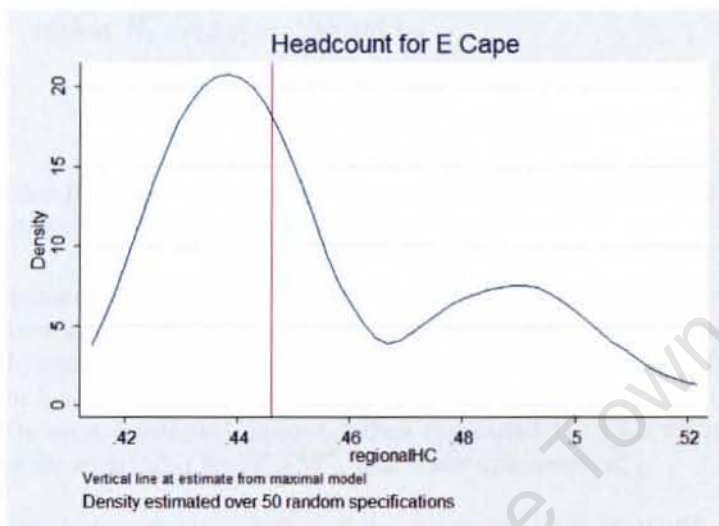


Figure 4.2: Density of Implied Headcount Over 50 Random Specifications, Eastern Cape

the latter holds, then at least some implementations of ELL yield biased estimates. As trivial as this point seems, it has been completely ignored by the literature.

Either all choices of specification lead to unbiased estimates, or at least some do not. The same comments hold with respect to the consistency of the estimates, and I will discuss the conditions under which ELL estimates will be consistent in section 4.3. Below, I use this logical truism to test for the presence of finite-sample bias indirectly.

Consider a pair of specifications for the first-stage model; call them \mathbf{X} and \mathbf{W} . If both estimates $\widehat{\mu}_a(\mathbf{X})$ and $\widehat{\mu}_a(\mathbf{W})$ are unbiased (for a given area) then the expectation of their difference must be zero. Define

$$\begin{aligned} m_x &= \mathbb{E}[\widehat{\mu}_a(\mathbf{X})] \\ m_w &= \mathbb{E}[\widehat{\mu}_a(\mathbf{W})] \end{aligned}$$

Say the sample size of the survey is s . A natural test statistic for $H_0 : m_x =$

$m_w = 0$ (against $H_1 : m_x \neq m_w$) would be

$$\begin{aligned}\hat{d}_s &= \widehat{\mu}_a(\mathbf{X}) - \widehat{\mu}_a(\mathbf{W}) \\ &= [\widehat{\mu}_a(\mathbf{X}) - \mu_a] + [\mu_a - \widehat{\mu}_a(\mathbf{W})]\end{aligned}\quad (4.1)$$

since under H_0 , $\mathbb{E}[\hat{d}_s] = d_s = 0$. If we reject H_0 , then we know that at least one of the two estimators is biased.

I approximate the joint sampling distribution of $(\widehat{\mu}_a(\mathbf{X}), \widehat{\mu}_a(\mathbf{W}))$ - and, by implication, the sampling distribution of \hat{d}_s - by bootstrapping the estimates. For each region I chose two of the 50 randomly generated specifications. Then, for $b = 1, \dots, B = 200$, I resampled the IES observations with replacement. On each resampled dataset I then computed the ELL estimates for each specification, $\widehat{\mu}_{s,b}^*(\mathbf{X}), \widehat{\mu}_{s,b}^*(\mathbf{W})$, and their difference, $\widehat{d}_{s,b}^*$.

The resulting first-stage models performed well in terms of the R^2 statistic. All provinces have mean R^2 -values over 0.5, and *no* first-stage model obtains an R^2 lower than 0.47. Thus, I also consider all of the bootstrapped estimates to have satisfied the literature's criteria.

Consider Figures 4.3 and 4.4, which show the joint distribution of the estimated headcount for Komga, in the Eastern Cape, over the 200 bootstrap samples. The ranges of the two estimates relative to one another is the most striking feature of Figure 4.3: the scatter does not even come close to the diagonal. If these estimators had the same means, we would expect to see much of the scatter concentrated about the line of equality, where $\widehat{\mu}_a(\mathbf{X}) = \widehat{\mu}_a(\mathbf{W})$. Instead, every single pair of estimates satisfies the same strict inequality $\widehat{\mu}_a(\mathbf{X}) > \widehat{\mu}_a(\mathbf{W})$ (where $\widehat{\mu}_a(\mathbf{W})$ is plotted on the vertical axis).

The marginal densities for Komga are depicted in Figure 4.4. Notice how the support of the densities are disjoint, which implies that 0 will be outside of the support of the bootstrap density of \hat{d}_s . In fact we can see this directly in Figure 4.5; not *one* of the bootstrapped pairs of headcounts enjoys a discrepancy of less than 10%.

As with the sensitivity of the point estimates, the bias result holds at the region level too. I calculate the implied regional headcount under both specifications for each bootstrap sample to obtain an approximation to the sampling distribution of \widehat{d}_b^* at the region level. The resulting density for

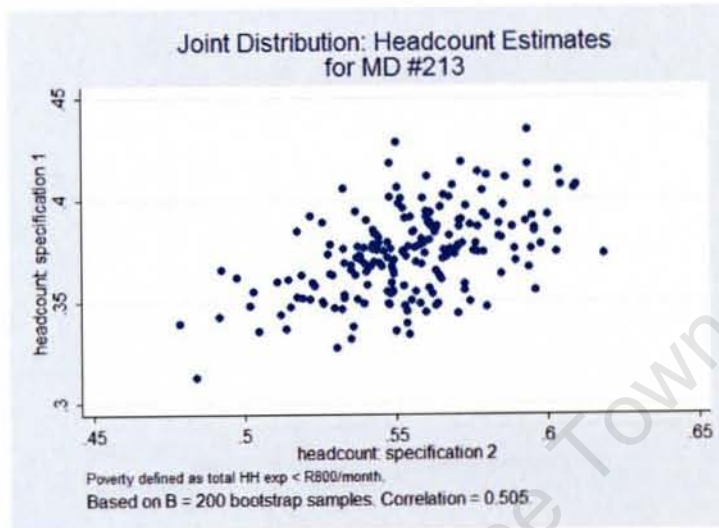


Figure 4.3: Joint Distribution of Bootstrapped Estimates for Komga (Eastern Cape)

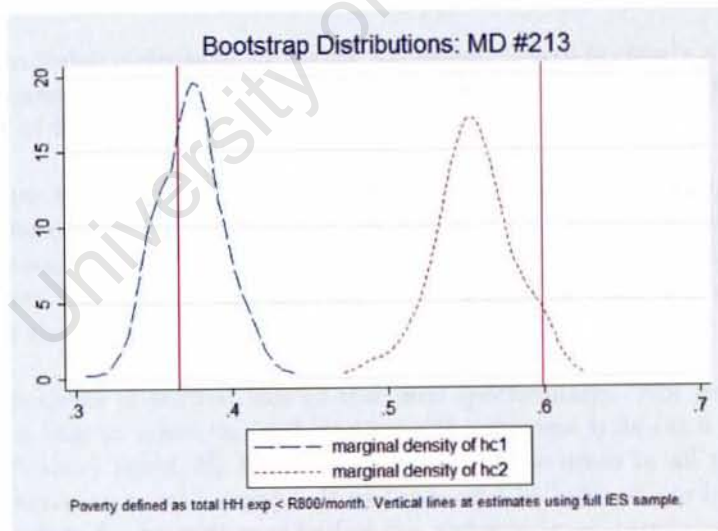


Figure 4.4: Marginal Bootstrap Densities for Komga (Eastern Cape)

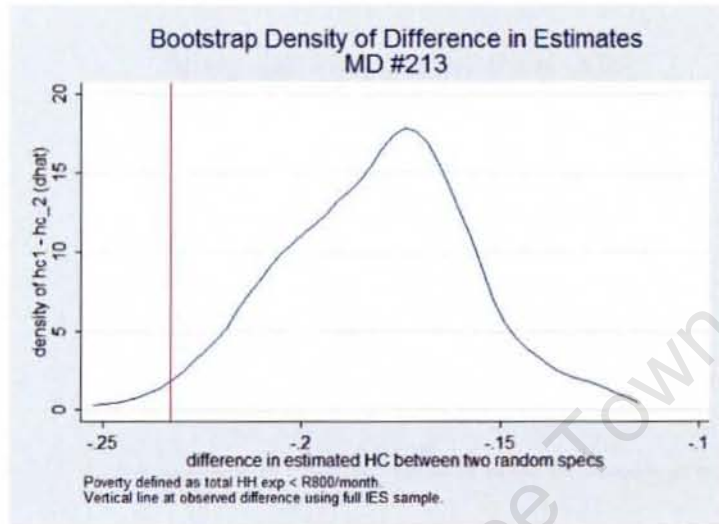


Figure 4.5: Marginal Bootstrap Density for Difference in Estimates, Komga (Eastern Cape)

KwaZulu-Natal is depicted in Figure 4.6 below. There is clearly a systematic difference between the two estimates; and again, 0 is outside of the support of the (approximate) sampling density of d_s .

I compute 80%, 90%, and 98% confidence intervals for the difference between the estimates by calculating the $100 \times (\alpha/2, 1 - \alpha/2)$ percentiles of the bootstrap distribution of the d_b^* for $\alpha = 0.1, 0.05$ and 0.01 . This allows me to test $H_0 : d_s = 0$ for each magisterial district and for each region. I tabulate the results of these tests in Table 4.4.

The hypothesis of mutual lack of bias fails spectacularly. Not *one* of the provinces fails to reject the null, and even the strictest tests (at a 2% level of significance) reject H_0 for more than half of the areas in all provinces save Mpumalanga and Limpopo. If we trade off a little bit of size for power, we can reject H_0 for well over half of the districts in *all* provinces at 10% significance, and for the Western Cape, we can reject $d = 0$ for *every single area*. And the most powerful test - at 20% significance - rejects H_0 in more than two-thirds of the areas in *every* province, with some provinces (the Free State, the Western Cape, and KwaZulu-Natal) confirming the presence

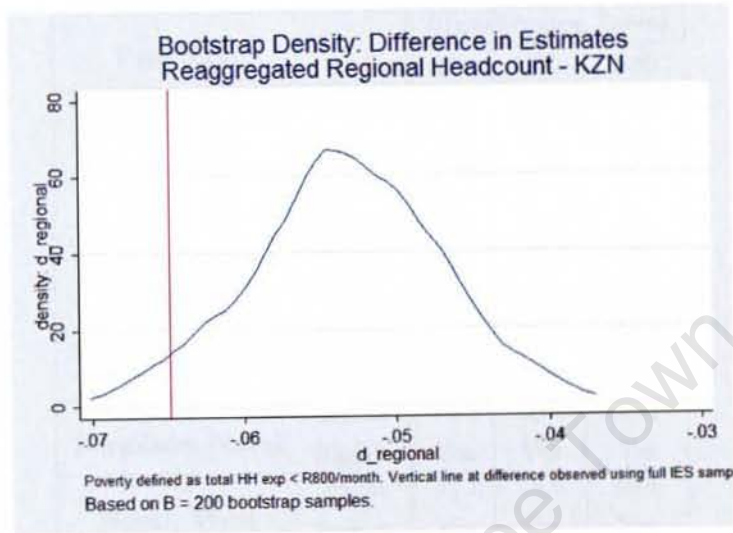


Figure 4.6: Marginal Bootstrap Density for Difference in Estimates, Reaggregated Headcount (KwaZulu-Natal)

of bias for over 90% of their magisterial districts.

4.2.2 Magnitude

Having confirmed the presence of finite-sample bias, I estimate a lower bound for its magnitude with half the absolute value of the mean of $\hat{d}_{s,b}^*$. This is truly a lower bound, because

$$\begin{aligned} \mathbb{E}[\hat{d}_s] &= \mathbb{E}[\hat{\mu}_a(\mathbf{X}) - \mu_a] - \mathbb{E}[\hat{\mu}_a(\mathbf{W}) - \mu_a] \\ &= \text{bias}(\hat{\mu}_a(\mathbf{X})) - \text{bias}(\hat{\mu}_a(\mathbf{W})) \end{aligned} \quad (4.2)$$

implying that

$$\begin{aligned} \frac{1}{2}|\mathbb{E}[\hat{d}_s]| &= \frac{1}{2}|\text{bias}(\hat{\mu}_a(\mathbf{X})) - \text{bias}(\hat{\mu}_a(\mathbf{W}))| \\ &\leq \frac{1}{2}(|\text{bias}(\hat{\mu}_a(\mathbf{X}))| + |\text{bias}(\hat{\mu}_a(\mathbf{W}))|) \\ &\leq \max\{|\text{bias}(\hat{\mu}_a(\mathbf{X}))|, |\text{bias}(\hat{\mu}_a(\mathbf{W}))|\} \end{aligned} \quad (4.3)$$

Province		Significance Level		
		2%	10%	20%
W Cape	Areas	97.6	100.0	100.0
	Region	Yes	Yes	Yes
E Cape	Areas	73.1	79.5	84.6
	Region	Yes	Yes	Yes
N Cape	Areas	50.0	69.2	80.8
	Region	Yes	Yes	Yes
Free State	Areas	84.6	90.4	94.2
	Region	Yes	Yes	Yes
KwaZulu-Natal	Areas	74.5	88.2	90.2
	Region	Yes	Yes	Yes
North West	Areas	73.7	78.9	89.5
	Region	Yes	Yes	Yes
Gauteng	Areas	54.2	62.5	75.0
	Region	Yes	Yes	Yes
Mpumalanga	Areas	45.2	58.1	67.7
	Region	Yes	Yes	Yes
Limpopo	Areas	41.9	61.3	67.7
	Region	Yes	Yes	Yes

Table 4.4: Percentage of Areas Rejecting $d_s = 0$, by Province and Significance Level

Thus, for each area (or region) the statistic

$$\hat{l}_s = \left| \frac{1}{2B} \sum_{b=1}^B \hat{d}_{s,b}^* \right| \quad (4.4)$$

is an approximate lower bound for the size of the bias of one of the estimators. I calculate this lower bound and I tabulate the summary statistics in Table 4.5.

The estimates in Table 4.5 tell a story that has now become familiar: the estimates for the North West province are particularly badly behaved, with half its areas having finite-sample biases of, at *best*, 7%, though the other provinces do not fare much better. For example, under one of the chosen specifications, the estimated headcount for every single area in Limpopo is

Province	Median	Min.	Max.	Regional Lower Bound
Western Cape	0.040	0.012	0.075	0.034
Eastern Cape	0.035	0.001	0.094	0.032
Northern Cape	0.039	0.001	0.098	0.037
Free State	0.048	0.000	0.098	0.047
KwaZulu-Natal	0.035	0.001	0.096	0.026
North West	0.070	0.004	0.196	0.051
Gauteng	0.017	0.002	0.055	0.016
Mpumalanga	0.022	0.001	0.102	0.022
Limpopo	0.025	0.002	0.093	0.017

Table 4.5: Summary Statistics - Lower Bounds for Bias, By Province

biased by at least 2%, and for some areas in that province, by at least 9%. And the situation is no better for the region-level estimates: some admissible specifications can yield implied headcounts that are biased by at least 4.7%, using the example of the Free State.

4.3 Consistency

Because the set of available regressors is constrained by those variables which are present and measured comparably in both the survey and the census data, the choice of specification is perforce atheoretic. Unfortunately the type of covariates which are likely to be included are very likely to be endogenous. For example, in the South African data I used in this paper, the possible regressors include variables on the household's amenities: whether it has a telephone, electric lighting, formal sanitation facilities, etc. Regardless of whether the dependent variable is expenditure or income (here, I have used expenditure) the possibility of simultaneity bias cannot be easily dismissed. Similarly, the demographic variables present in census data (like household size or the gender of the household head) are almost certainly correlated with the regression error in any model of consumption.

4.3.1 A Simple Illustration of the Inadequacy of the R^2 Criterion

First, I want to dismiss any persistent concerns that the high first-stage R^2 statistics indicate that the second-stage imputations will be close to the true values. Consider the following data-generating process:

$$y_i = x_i\beta_0 + \varepsilon_i \quad (4.5)$$

$$\text{Cov}(x, \varepsilon) = \alpha \quad (4.6)$$

with $\mathbb{E}[x] = 0 = \mathbb{E}[\varepsilon]$, $\text{Var}[x] = V_x$, and $\text{Var}[\varepsilon] = \sigma^2$.

Say we have a simple random sample of size s from this process and we compute $\hat{\beta}$ by OLS. Then the R^2 measure is

$$\begin{aligned} R_s^2 &= \frac{\frac{1}{s} \sum_{i=1}^s (x_i \hat{\beta}_s - \bar{y}_s)^2}{\frac{1}{s} \sum_{i=1}^s (y_i - \bar{y}_s)^2} \\ &= \frac{(\hat{\beta}_s)^2 \left(\frac{1}{s} \sum_{i=1}^s x_i^2 \right) - 2(\hat{\beta}_s)(\bar{y}_s)(\bar{x}_s) + \bar{y}_s^2}{\frac{1}{s} \sum_{i=1}^s (y_i - \bar{y}_s)^2} \end{aligned} \quad (4.7)$$

Define

$$\begin{aligned} \theta &= \text{plim}_{s \rightarrow \infty} \hat{\beta}_s \\ &= \beta_0 + \frac{\alpha}{V_x} \end{aligned} \quad (4.8)$$

The denominator of (4.7) is consistent for $\text{Var}[y] = \beta_0^2 V_x + 2\beta_0\alpha + \sigma^2$. Using Slutsky's Theorem and the fact that $(x_i^2)_{i=1}^\infty$ is an i.i.d. process when $(x_i)_{i=1}^\infty$ is, we see that the numerator, the "explained" sum of squares, has probability limit

$$\begin{aligned} \text{plim}_{s \rightarrow \infty} \frac{1}{s} \sum_{i=1}^s (\hat{y}_i - \bar{y}_s)^2 &= \theta^2 V_x - 2\theta \mathbb{E}[y] \mathbb{E}[x] + (\mathbb{E}[y])^2 \\ &= \left(\beta_0 + \frac{\alpha}{V_x} \right)^2 V_x \\ &= \beta_0^2 V_x + 2\beta_0\alpha + \alpha^2/V_x \end{aligned}$$

so that

$$\begin{aligned} R_\infty^2(\alpha) &= \text{plim}_{s \rightarrow \infty} R_s^2 \\ &= \frac{\beta_0^2 V_x + 2\beta_0\alpha + \alpha^2/V_x}{\beta_0^2 V_x + 2\beta_0\alpha + \sigma^2} \end{aligned} \quad (4.9)$$

Notice that $R_\infty^2 \rightarrow 1$ as $\alpha \rightarrow \pm\sigma\sqrt{V_x}$. This is perfectly intuitive: if x and ε are highly correlated, then x should "explain" much of the variation in y - the x part and most of the ε part, too! So a high R^2 might just indicate severe endogeneity, which also means that $|\theta - \beta_0| \gg 0$.

4.3.2 Direct Evidence: Specification Choice Shifts The Conditional Mean

Now, to see why first-stage consistency (and hence the choice of specification) has such a large impact on the estimates, despite the uniformly high R^2 values, consider the formula for the estimator $\hat{\mu}_a(\mathbf{X}_a)$ (neglecting the error due to numerical integration):

$$\hat{\mu}_a = \int_{\mathbb{R}^K} \left[\int_{\mathbb{R}^{N_a}} W(\mathbf{y}) \hat{f}(\mathbf{y}|\mathbf{X}_a, \tilde{\beta}) d\mathbf{y} \right] \hat{f}^\alpha(\tilde{\beta}|\mathbf{X}_R, \hat{\beta}) d\tilde{\beta}$$

This suggests two ways in which the ELL technique can fail.

Firstly, $\hat{f}^\alpha(\tilde{\beta}|\mathbf{X}_R)$, the estimated asymptotic distribution of $\tilde{\beta}$ - which we get from the first-stage regression - could be a poor approximation. Estimating $\tilde{\beta}$ inconsistently is a good way to ensure this. In particular, if

$\text{plim}_{s \rightarrow \infty} \hat{\beta} = \theta \neq \beta_0$, the distribution with respect to which we integrate the inner integral (which is a function of $\tilde{\beta}$) will concentrate probability mass on ever-smaller neighbourhoods of θ as the IES sample size, s , increases. This only matters if getting β right matters. It turns out that it does, which I will show in section 4.3.3 below.

Secondly, $\hat{f}(y_a | \mathbf{X}_a, \hat{\beta})$, the implied conditional density of log expenditure, could be a poor approximation to the true conditional density. Intuitively, if $\hat{\beta}$ is not consistent for β_0 but instead for $\theta \neq \beta_0$, the hyperplane $\mathbf{x}_i \hat{\beta}$ about which each household i 's simulated (log) expenditure varies will differ systematically from the true conditional mean $\mathbf{x}_i \beta_0$.

In fact, we can see this happening directly from the estimation results. In Figure 4.7 I exhibit the kernel density estimates of the marginal density of $\mathbf{x} \hat{\beta}^j$ for alternative specifications j for Bizana, in the Eastern Cape.

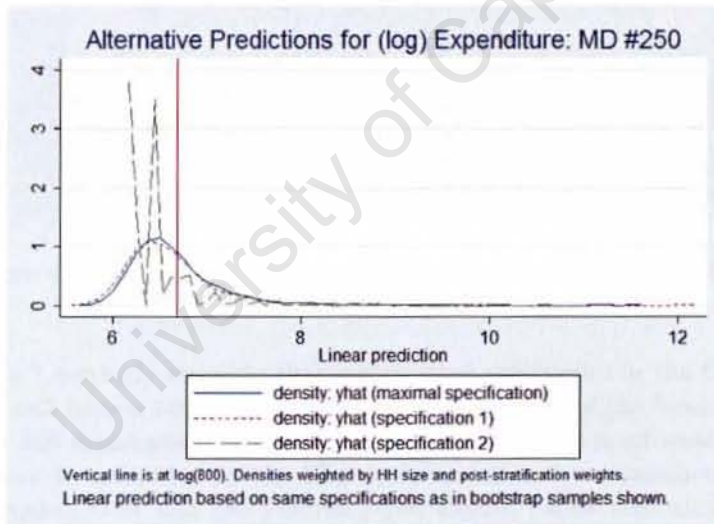


Figure 4.7: Estimated Conditional Mean Log Expenditure Under Different Specifications, Bizana (Eastern Cape)

The difference in the estimated conditional mean between the two specifications is visually obvious. For Bizana, the estimated density of the conditional mean under specification 2 is strongly concentrated below the poverty

line, while specification 1 appears to predict that a substantial minority of individuals will obtain incomes above the poverty line. And, indeed, we see this in Figure 4.8, which displays the marginal densities of the bootstrap distribution of the two headcount estimates for Bizana: the density for specification 2 puts most of its probability mass to the right of the density for specification 1.

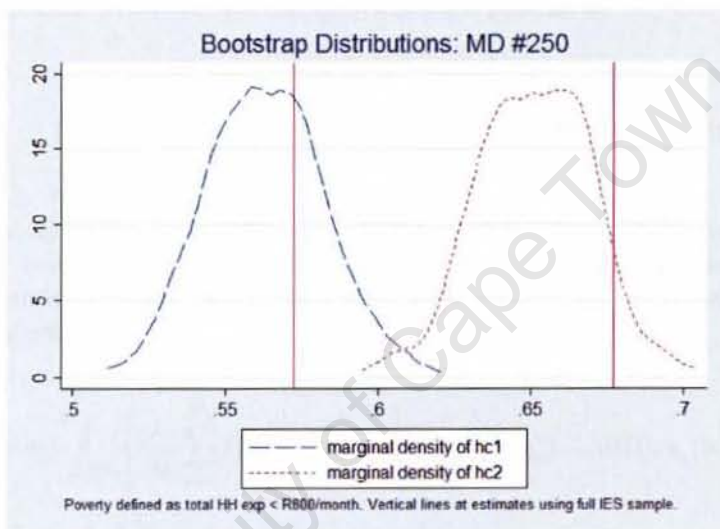


Figure 4.8: Marginal Bootstrap Densities for Bizana (Eastern Cape)

Figure 4.7 certainly suggests that inconsistent estimation in the first stage of ELL will have a large impact on the final estimates of the headcount. In fact we can decompose the difference between the true headcount and the headcount as estimated by the ELL technique into two components that I call *sampling error* and *specification error*. Below, I show that although the sampling error becomes negligible with large survey samples, the specification error is likely to persist asymptotically, *unless* the first-stage estimation is consistent.

4.3.3 Specification Error and Sampling Error: A Decomposition

Say $W(\cdot)$ is the headcount measure and the data generating process is given by

$$y_i = m(\mathbf{x}_i) + u_i \quad (4.10)$$

with the observations (y_i, \mathbf{x}_i) independent over i , where $m(\cdot)$ is the population regression function. We can also write the process as

$$y_i = \mathbf{x}_i\beta_0 + \varepsilon_i \quad (4.11)$$

where $\mathbf{x} \mapsto \mathbf{x}\beta_0$ is the best (in the mean-squared error sense) linear predictor of y .

Let $H(u, \mathbf{x})$ be the joint cumulative distribution function for the covariates and the true population residual u_i . If \mathbf{x} and u are not independent then the dependence of $H(\cdot, \mathbf{x})$ on \mathbf{x} is nontrivial. Then the *true* (conditional) expected value of W is:

$$\begin{aligned} \mu_a &= \int_{\mathbb{R}^{N_a}} \left[\frac{1}{N_a} \sum_{i=1}^{N_a} \mathbf{1}_{\{m(\mathbf{x}_i) + u_i < z^*\}} \right] dH(u_1, \mathbf{x}_{N_a}) \dots dH(u_{N_a}, \mathbf{x}_{N_a}) \\ &= \frac{1}{N_a} \sum_{i=1}^{N_a} \mathbb{P}(u_i < y_i - m(\mathbf{x}_i)) \\ &= \frac{1}{N_a} \sum_{i=1}^{N_a} H(z^* - m(\mathbf{x}_i), \mathbf{x}_i) \end{aligned} \quad (4.12)$$

What we are able to calculate with ELL, however, is (ignoring the numerical integration over the presumed sampling distribution of $\hat{\beta}$):

$$\hat{\mu}_a = \frac{1}{N_a} \sum_{i=1}^{N_a} \hat{F}^s(z^* - \mathbf{x}_i\hat{\beta}) \quad (4.13)$$

where $\hat{F}^s(\cdot)$ is the marginal cumulative distribution function chosen by the researcher, and the superscript s emphasizes that the function itself depends on the survey sample, either through rescaling (if a parametric distribution is imposed on u) or directly (if, say, the empirical distribution of the first-stage residuals is used).

Each term $H(z^* - m(\mathbf{x}_i), \mathbf{x}_i) - \widehat{F}^s(z^* - \mathbf{x}_i \widehat{\beta})$ in the difference (4.12) - (4.13) is identically equal to

$$\begin{aligned} H(z^* - m(\mathbf{x}_i), \mathbf{x}_i) - \widehat{F}^s(z^* - \mathbf{x}_i \widehat{\beta}) &= H(z^* - m(\mathbf{x}_i), \mathbf{x}_i) - H(z^* - \mathbf{x}_i \beta_0, \mathbf{x}_i) \\ &+ H(z^* - \mathbf{x}_i \beta_0, \mathbf{x}_i) - H(z^* - \mathbf{x}_i \widehat{\beta}, \mathbf{x}_i) \\ &+ H(z^* - \mathbf{x}_i \widehat{\beta}, \mathbf{x}_i) - \widehat{F}^s(z^* - \mathbf{x}_i \widehat{\beta}) \end{aligned} \quad (4.14)$$

The first term in (4.14) above might be called “specification error”: it is due to the discrepancy over the domain of \mathbf{x} between the true conditional mean function $m(\mathbf{x}_i)$ and the linear predictor $\mathbf{x} \mapsto \mathbf{x}\beta_0$. The second term could be called “sampling error,” and if H is continuous this term will vanish asymptotically, since $\widehat{\beta}$ is consistent for β_0 . I call the third term, $H(z^* - \mathbf{x}_i \widehat{\beta}, \mathbf{x}_i) - \widehat{F}^s(z^* - \mathbf{x}_i \widehat{\beta})$, “approximation error” since it arises due to the difference - at each point \mathbf{x}_i - between the chosen approximating distribution \widehat{F}^s and the joint CDF of the true errors, H .

It is not necessarily true that the approximation error will vanish asymptotically, even if the empirical distribution of the first-stage residuals is used.³

Moreover if H is differentiable we have

$$H(z^* - m(\mathbf{x}_i), \mathbf{x}_i) - H(z^* - \mathbf{x}_i \beta_0, \mathbf{x}_i) \approx \frac{\partial H}{\partial u}(c_i^*, \mathbf{x}_i) \times (\mathbf{x}_i \beta_0 - m(\mathbf{x}_i)) \quad (4.15)$$

by the mean value theorem, where we interpret \approx to mean that the difference converges to zero in probability and $c_i^* \in [z^* - m(\mathbf{x}_i), z^* - \mathbf{x}_i \theta]$.

Summing over the observations i shows that the magnitude of this “asymptotic bias” depends on several factors: the unknown joint CDF H ; the marginal distribution of \mathbf{x} in the area a ; and the difference between the conditional mean function $m(\mathbf{x}_i)$ and the linear predictor $\mathbf{x} \mapsto \mathbf{x}\beta_0$.

This dependence is probably the reason that ELL estimates are sensitive to specification: including different regressors in \mathbf{x} alters the magnitude of $m(\mathbf{x}_i) - \mathbf{x}_i \beta_0$ at each point \mathbf{x}_i in the domain of the covariates. This is why a “good” model, in the sense that it has a high first-stage R^2 , does not necessarily produce consistent estimates of the integrals of functions weighted by its estimated conditional density.

³I thank an anonymous referee for correcting an earlier version of my argument.

Chapter 5

A Nonparametric Estimation Technique

5.1 Introduction and Motivation

One of the central problems with the ELL technique, as I have argued in chapter 4, is that it relies too heavily on a finely specified parametric structure for the data-generating process. Departures from the presumed structure - specifically, the likely failure of exogeneity for the covariates of log household expenditure - are likely to yield small-area estimates that are inconsistent and severely biased in finite samples.

In this chapter I propose a new, nonparametric small-area estimator that is consistent in all areas under weaker conditions than ELL. In addition I derive a formula for the (conditional) variance of the estimator and I provide evidence (from a Monte Carlo simulation) that its estimated standard errors are small and negligibly biased. Finally I compute small-area estimates of the headcount for the Free State in 1995 and compare the results to the ELL estimates.

5.1.1 The Setup

As before, we have a joint data-generating process on log expenditure and a vector of covariates \mathbf{x} ; both the survey and the census data, I will assume, are drawn as simple random samples from the same superpopulation. Furthermore, I will assume that the census and survey samples are inde-

pendent; while this literally cannot be true (since censuses are exhaustive), the intersection of the census and survey samples is likely to be negligible. For example, the 1995 October Household Survey covered approximately 130 000 individuals; the 1996 population census (before adjusting for undercount) covered 36.2 million individuals, implying an overlap of at most 0.3%.

Given the sparse nature of data on expenditure and the “curse of dimensionality” inherent in nonparametric estimation, I will limit my discussion here to the scalar case, but clearly all of the results that follow can be extended to the case where $\dim(\mathbf{x}) > 1$.

I will assume that the scalar covariate x is discrete, taking on k distinct values; without loss of generality, suppose these are the first k integers. x need not be *categorical* - household size, for example, has a natural order - but since my focus is on estimating functionals of the distribution of y , the particular values of x are irrelevant here.

I denote the joint density of (y, x) (with respect to the product of Lebesgue and counting measure) $f_a(y, x)$ for area a , and $f_R(y, x)$ at the region level. Writing ω_a for the population share of area a , we have

$$f_R(y, x) \equiv \sum_{a=1}^A \omega_a f_a(y, x) \quad (5.1)$$

on $\mathbb{R} \times \{1, \dots, k\}$; I refer to this requirement as the “reaggregation constraint”.

My goal is to estimate the expected value of a welfare measure $W(\mathbf{y})$ for area a , i.e.

$$\begin{aligned} \mu_a &= \mathbb{E}[W(\mathbf{y})|a] \\ &= \int_{\mathbb{R}^{N_a}} W(\mathbf{y}) f_a(\mathbf{y}) d\mathbf{y} \end{aligned} \quad (5.2)$$

5.1.2 Assumptions

I make two primitive assumptions on the data-generating process:

ASSUMPTION 5.1 (Global Support). $f_R(x) > 0$ on $\{1, 2, \dots, k\}$.

ASSUMPTION 5.2 (Conditional Homogeneity). *For any pair of areas a, a' in the region, we have*

$$f_a(y|\mathbf{x}) \equiv f_{a'}(y|\mathbf{x}) \quad (5.3)$$

on $\mathbb{R} \times \{1, \dots, k\}$.

Note that when conditional homogeneity holds, the reaggregation constraint implies that $f_a(y|x) = f_R(y|x)$ too.

As with the ELL technique, the specification - exactly *which* X to choose out of the set of variables which are observed both in the census and survey data - is not restricted by the estimation technique I propose. Much like the identification assumptions in other empirical work, assumption 5.2 should be *motivated*, but it cannot be verified given the sparse nature of the survey data.

Notice that the assumptions I use here are strictly weaker than those needed for the consistency of ELL (see Chapter 4, especially section 4.3): write the true data-generating process as

$$y_i = m(x_i) + u_i \quad (5.4)$$

where $m(\cdot)$ is the conditional mean function, and

$$u_i \sim f_u(\cdot|x_i) \quad (5.5)$$

where $f_u(\cdot|x)$ is a family of (mean-zero) densities that potentially varies with the value of x , but *not* on the area a . (This is the allowance for heteroskedasticity in ELL.)

Under ELL, we further require $m(x_i) = x_i\beta_0$ for some fixed vector β_0 . Hence the assumptions which justify the use of ELL are a special, more restrictive case of the above conditional homogeneity assumption.

5.1.3 Deriving the Estimator from the Analogy Principle

If conditional homogeneity holds, then the joint distribution of (y, x) in each area a can easily be obtained by multiplying the common conditional density $f_R(y|x)$ by the area-specific marginal distribution $f_a(x)$, which can easily be

obtained from the abundant census data.

The law of total expectations gives us - for any random variable W - that

$$\begin{aligned}\mathbb{E}[W|a] &= \mathbb{E}[\mathbb{E}[W|X, a]|a] \\ &= \sum_{x=1}^k \mathbb{P}(X = x|a) \mathbb{E}[W|X = x, a] \\ &= \sum_{x=1}^k \mathbb{P}(X = x|a) \mathbb{E}[W|X = x]\end{aligned}\quad (5.6)$$

where the final equality follows from a conditional homogeneity assumption like (5.3): that given X , the area a is irrelevant (with respect to W). As I argue below, each term in (5.6) can be consistently estimated from the existing census and survey data.

To put this insight in more concrete terms, let

$$W(y_1, \dots, y_n) = \frac{1}{n} \sum_{i=1}^n w(y_i) \quad (5.7)$$

be an additively separable welfare measure. Since the linearity of the integral means that the expected value of W will depend only on the marginal densities for each household (or individual), so that I can ignore the covariance structure of the (unobserved) area-specific expenditure vector \mathbf{y}_a , unlike in the ELL technique, where arbitrary functional forms for heteroskedasticity have to be used.

Applying the analogy principle, $\mathbb{E}[W(y_1, \dots, y_{s_x})|X = x]$ should be estimated by the sample mean of $w(\cdot)$ over the survey observations j for which $X_j = x$ (of which there are s_x), since

$$\begin{aligned}\mathbb{E}[W(y_1, \dots, y_{s_x})|X = x] &= \frac{1}{s_x} \sum_{j=1}^{s_x} \mathbb{E}[w(y_j)|X = x] \\ &= \mathbb{E}[w(y_j)|X = x]\end{aligned}\quad (5.8)$$

on the assumption that the survey observations are all drawn from the same population (so that $\mathbb{E}[w(y_j)|X = x]$ is constant over j).

Then, in formula (5.6), $\mu_a = \mathbb{E}[W(\mathbf{y})|a]$ should be estimated by weighting $\mathbb{E}[w(\mathbf{y})|X = x]$ by the area-specific probabilities $p_a^x = \mathbb{P}(X = x|a)$, or, replacing p_a^x by the consistent estimate $\hat{p}_a^x = N_a^x/N_a$ we have:

$$\begin{aligned}\hat{\mu}_a &= \sum_{x=1}^k \frac{N_a^x}{N_a} \left[\frac{1}{s_x} \sum_{\{j: X=x\}} w(y_j) \right] \\ &= \sum_{x=1}^k \hat{p}_a^x \hat{w}_x\end{aligned}\quad (5.9)$$

Of course, the above formula should be modified in the case of weighted data or for person-level (rather than household-level) measures.

5.2 Properties of the Nonparametric Estimator

5.2.1 Consistency

THEOREM 5.1. *Let $s = \sum_{x=1}^k s_x$ be the size of the survey sample and let N_a be the size of the census population. If conditional homogeneity holds, then*

$$\text{plim}_{s \rightarrow \infty} \left(\text{plim}_{N_a \rightarrow \infty} \hat{\mu}_a \right) = \mu_a \quad (5.10)$$

Proof. Step 1. Assume, for now, that

$$\text{plim}_{s \rightarrow \infty} \hat{w}_x = \mathbb{E}[W(\mathbf{y})|X = x] \quad (5.11)$$

If this is so, then applying the weak law of large numbers as $N_a \rightarrow \infty$ we have that $\hat{p}_a^x \rightarrow p_a^x$ (in probability), i.e.

$$\begin{aligned}\text{plim}_{s \rightarrow \infty} \left(\text{plim}_{N_a \rightarrow \infty} \hat{\mu}_a \right) &= \sum_{x=1}^k p_a^x \mathbb{E}[W(\mathbf{y})|X = x, a] \\ &= \mathbb{E}[\mathbb{E}[W(\mathbf{y})|X = x, a]|a] \\ &= \mathbb{E}[W(\mathbf{y})|a] \\ &= \mu_a\end{aligned}\quad (5.12)$$

Step 2. I establish that

$$\begin{aligned}
 \text{plim}_{s \rightarrow \infty} \widehat{\mu}_a &= \sum_{x=1}^k \frac{N_a^x}{N_a} \text{plim}_{s \rightarrow \infty} \left[\sum_{\{j: X=x\}} \frac{1}{s_x} w(y_j) \right] \\
 &= \sum_{x=1}^k \widehat{p}_a^x \int_{-\infty}^{\infty} w(y) f_R(y|X=x) dy \\
 &= \sum_{x=1}^k \widehat{p}_a^x \int_{-\infty}^{\infty} w(y) f_a(y|X=x) dy \quad (5.13)
 \end{aligned}$$

where the final equality follows from the conditional homogeneity assumption.

This is not obvious because it is s , not s_x , which is increasing without bound. Write

$$\widehat{w}_x = \frac{1}{s_x} \sum_{\{j: X=x\}} w(y_j) \quad (5.14)$$

To show that (5.13) holds, it is enough to prove that

$$\widehat{w}_x \longrightarrow \int_{-\infty}^{\infty} w(y) f_a(y|X=x) dy \quad (5.15)$$

in mean square, i.e.

$$\text{Var} \left[\widehat{w}_x - \int_{-\infty}^{\infty} w(y) f_a(y|X=x) dy \right] \longrightarrow 0 \quad (5.16)$$

as $s \rightarrow \infty$.

Firstly, notice that by the law of total expectations,

$$\begin{aligned}
 \mathbb{E}[\widehat{w}_x] &= \mathbb{E}[\mathbb{E}[\widehat{w}_x | s_x]] \\
 &= \mathbb{E} \left[\int_{-\infty}^{\infty} w(y) f_a(y|X=x) dy \right] \\
 &= \int_{-\infty}^{\infty} w(y) f_a(y|X=x) dy \quad (5.17)
 \end{aligned}$$

Secondly, we have the variance decomposition

$$\begin{aligned}
 \text{Var}[\widehat{w}_x] &= \mathbb{E}[\text{Var}[\widehat{w}_x | s_x]] + \text{Var}[\mathbb{E}[\widehat{w}_x | s_x]] \\
 &= \mathbb{E}[\text{Var}[\widehat{w}_x | s_x]] + 0 \\
 &= \mathbb{E} \left[\frac{\text{Var}[w(y)]}{s_x} \right] \quad (5.18)
 \end{aligned}$$

Now

$$\text{Var}[\hat{w}_x | s_x] = \begin{cases} \text{Var}[w(y)]/s_x & \text{if } s_x > 0 \\ 0 & \text{if } s_x = 0 \end{cases} \quad (5.19)$$

Conditional on s , s_x is binomially distributed with parameter $\mathbb{P}(X = x) = p_x \in (0, 1)$. The variance of \hat{w}_x , therefore, is proportional to

$$\mathbb{E} \left[\frac{1}{s_x} | s_x > 0 \right] = \frac{1}{1 - (p_x)^s} \sum_{r=1}^s \binom{s}{r} (p_x)^r (1 - p_x)^{s-r} \frac{1}{r} \quad (5.20)$$

I need to show that the expression in (5.20) converges to 0 as $s \rightarrow \infty$. To see this, consider the following sequence of functions:

$$h_s(z) = -\frac{1}{s}z + \left(1 + \frac{1}{s}\right) \quad (5.21)$$

Notice that $1/z \leq h_s(z)$ on $[1, s]$. Also, by the assumption of global support for x , $\mathbb{P}(s_x = 0) < 1$. Therefore,

$$\begin{aligned} \mathbb{E} \left[\frac{1}{s_x} | s_x > 0 \right] &\leq \mathbb{E} [h_s(s_x) | s_x > 0] \\ &= -\frac{1}{s} \mathbb{E} [s_x | s_x > 0] + \left(1 + \frac{1}{s}\right) \\ &= -\frac{1}{s} \left[\frac{sp_x}{1 - (p_x)^s} \right] + 1 + \frac{1}{s} \\ &= -\frac{p_x}{1 - (p_x)^s} + 1 + \frac{1}{s} \\ &\rightarrow 1 - p_x \end{aligned} \quad (5.22)$$

Since p_x can be chosen arbitrarily close to 1, we have that

$$\lim_{s \rightarrow \infty} \text{Var} \left[\hat{w}_x - \int_{-\infty}^{\infty} w(y) f_a(y | X = x) dy \right] = 0 \quad (5.23)$$

□

5.2.2 Reaggregation

Since additively separable welfare measures can be estimated directly at the regional level as

$$\hat{\mu}_R = \frac{1}{s} \sum_j w(y_j) \quad (5.24)$$

we have a natural constraint on the small-area estimates: that the reaggregated small-area estimates should agree with the direct region-level estimate, i.e.

$$\mu_R = \sum_{a=1}^A \omega_a \mu_a \quad (5.25)$$

The nonparametric estimator satisfies this requirement asymptotically, in the following sense:

THEOREM 5.2. *Let $N_R = \sum_{a=1}^A N_a$ be the size of the census sample, and write $\hat{\omega}_a = N_a/N_R$ for the census estimate of ω_a . Then if assumptions (5.1) - (5.2) hold, we have that*

$$\text{plim}_{N_a \rightarrow \infty} \left(\text{plim}_{s \rightarrow \infty} \left| \hat{\mu}_R - \sum_{a=1}^A \hat{\omega}_a \hat{\mu}_a \right| \right) = 0 \quad (5.26)$$

Proof. Note that

$$\begin{aligned} \sum_{a=1}^A \hat{\omega}_a \hat{\mu}_a &= \sum_{a=1}^A \frac{N_a}{N_R} \sum_{x=1}^k \frac{N_a^x}{N_a} \hat{w}_x \\ &= \sum_{a=1}^A \sum_{x=1}^k \frac{N_a^x}{N_R} \hat{w}_x \\ &= \sum_{x=1}^k \hat{p}_R^x \hat{w}_x \end{aligned} \quad (5.27)$$

where $N_R^x = \sum_{a=1}^A N_a^x$, and $\hat{p}_R^x = N_R^x/N_R$ is the census estimate of $\mathbb{P}(X = x)$ at the region level.

Also, we have that

$$\begin{aligned} \hat{\mu}_R &= \frac{1}{s} \sum_{x=1}^k s_x \hat{w}_x \\ &= \sum_{x=1}^k \hat{q}_R^x \hat{w}_x \end{aligned} \quad (5.28)$$

where $\hat{q}_R^x = s_x/s$. Both \hat{p}_R^x and \hat{q}_R^x are weakly consistent for $\mathbb{P}(X = x)$ as $s, N_R \uparrow \infty$ (respectively), so that

$$\begin{aligned}
 \text{plim}_{N_a \rightarrow \infty} \text{plim}_{s \rightarrow \infty} \left(\left| \hat{\mu}_R - \sum_{a=1}^A \hat{\omega}_a \hat{\mu}_a \right| \right) &= \text{plim}_{N_a \rightarrow \infty} \text{plim}_{s \rightarrow \infty} \left(\left| \sum_{x=1}^k (\hat{p}_R^x - \tilde{p}_R^x) \hat{w}_x \right| \right) \\
 &\leq \text{plim}_{N_a \rightarrow \infty} \text{plim}_{s \rightarrow \infty} \left(\sum_{x=1}^k |\hat{p}_R^x - \tilde{p}_R^x| |\hat{w}_x| \right) \\
 &= 0 \tag{5.29}
 \end{aligned}$$

□

5.2.3 Finite-Sample Distribution

Lack of Bias

The estimator defined in (5.9) is unbiased under the assumption that the census and survey observations are independent, since

$$\begin{aligned}
 \mathbb{E}[\hat{\mu}_a] &= \sum_{x=1}^k \mathbb{E}[\hat{p}_a^x \hat{w}_x] \\
 &= \sum_{x=1}^k (\mathbb{E}[\hat{p}_a^x] \mathbb{E}[\hat{w}_x] + \text{cov}(\hat{p}_a^x, \hat{w}_x)) \\
 &= \mu_a + \sum_{x=1}^k \text{cov}(\hat{p}_a^x, \hat{w}_x) \tag{5.30}
 \end{aligned}$$

and independence implies that $\text{cov}(\hat{p}_a^x, \hat{w}_x) = 0$ for all x .

Variance

LEMMA 5.3. Let X_1, \dots, X_N be a collection of random variables with $\text{Var}[X_n] < \infty$ for each $1 \leq n \leq N$. Define

$$S_N = \sum_{n=1}^N X_n \tag{5.31}$$

Then, writing $V[\cdot]$ for the variance, we have

$$V[S_N] = \sum_{n=1}^N V[X_n] + 2 \sum_{n=2}^N \sum_{m=1}^{n-1} \text{cov}(X_n, X_m)$$

Proof. Since $S_N = X_N + S_{N-1}$, we have

$$\begin{aligned} V[S_N] &= V[X_N] + V[S_{N-1}] + 2\text{cov}(X_N, S_{N-1}) \\ &= V[X_N] + V[S_{N-1}] + 2 \sum_{n=1}^{N-1} \text{cov}(X_N, X_n) \end{aligned} \quad (5.32)$$

Write

$$\begin{aligned} \Delta_N &= V[S_N] - V[S_{N-1}] \\ &= V[X_N] + 2 \sum_{n=1}^{N-1} \text{cov}(X_N, X_n) \end{aligned} \quad (5.33)$$

Then

$$\begin{aligned} V[S_N] &= V[X_1] + \sum_{j=2}^N \Delta_j \\ &= V[X_1] + \sum_{j=2}^N \left(V[X_j] + 2 \sum_{n=1}^{j-1} \text{cov}(X_j, X_n) \right) \\ &= \sum_{n=1}^N V[X_n] + 2 \sum_{n=2}^N \sum_{m=1}^{n-1} \text{cov}(X_n, X_m) \end{aligned} \quad (5.34)$$

□

THEOREM 5.4. *Let*

$$w_x = \mathbb{E}[\hat{w}_x] = \int_{-\infty}^{\infty} w(y) f_a(y|X=x) dy \quad (5.35)$$

and

$$V_x = s_x \cdot V[\hat{w}_x] = V[w(y)|X=x] \quad (5.36)$$

Then the conditional variance of $\hat{\mu}_a$ is

$$\begin{aligned} V[\hat{\mu}_a | (N_a)_{a=1}^A, (s_x)_{x=1}^k] &= \sum_{x=1}^k \left[(w_x)^2 \frac{p_a^x(1-p_a^x)}{N_a} + \frac{V_x}{s_x} \left(\frac{p_a^x(1-p_a^x)}{N_a} + (p_a^x)^2 \right) \right] \\ &\quad - 2 \sum_{x=1}^k \sum_{m=1}^{x-1} \frac{p_a^x p_a^m}{N_a} w_x w_m \end{aligned} \quad (5.37)$$

Proof. By Lemma 5.3, we have

$$\begin{aligned} V[\hat{\mu}_a] &= V\left[\sum_{x=1}^k \hat{p}_a^x \hat{w}_x\right] \\ &= \sum_{x=1}^k V[\hat{p}_a^x \hat{w}_x] + 2 \sum_{x=1}^k \sum_{m=1}^{x-1} \text{cov}(\hat{p}_a^x \hat{w}_x, \hat{p}_a^m \hat{w}_m) \end{aligned} \quad (5.38)$$

By the assumed independence of the survey and the census, $h_1(\hat{w}_x)$ is independent of $h_2(\hat{p}_a^x)$ for any pair of measurable functions $h_1(\cdot), h_2(\cdot)$. Then, the terms in the first summation are

$$\begin{aligned} V[\hat{p}_a^x \hat{w}_x] &= \mathbb{E}[(\hat{p}_a^x \hat{w}_x)^2] - \mathbb{E}[\hat{p}_a^x \hat{w}_x]^2 \\ &= \mathbb{E}[(\hat{p}_a^x)^2] \mathbb{E}[(\hat{w}_x)^2] - \mathbb{E}[\hat{p}_a^x]^2 \mathbb{E}[\hat{w}_x]^2 \\ &= \left\{ \mathbb{E}[(\hat{p}_a^x)^2] \mathbb{E}[(\hat{w}_x)^2] - \mathbb{E}[\hat{p}_a^x]^2 \mathbb{E}[(\hat{w}_x)^2] \right\} \\ &\quad + \left\{ \mathbb{E}[\hat{p}_a^x]^2 \mathbb{E}[(\hat{w}_x)^2] - \mathbb{E}[\hat{p}_a^x]^2 \mathbb{E}[\hat{w}_x]^2 \right\} \\ &= \mathbb{E}[(\hat{w}_x)^2] V[\hat{p}_a^x] + \mathbb{E}[\hat{p}_a^x]^2 V[\hat{w}_x] \\ &= (V[\hat{w}_x] + (w_x)^2) V[\hat{p}_a^x] + \mathbb{E}[\hat{p}_a^x]^2 V[\hat{w}_x] \\ &= \left(\frac{V_x}{s_x} + (w_x)^2 \right) \frac{p_a^x(1-p_a^x)}{N_a} + (p_a^x)^2 \frac{V_x}{s_x} \\ &= (w_x)^2 \frac{p_a^x(1-p_a^x)}{N_a} + \frac{V_x}{s_x} \left(\frac{p_a^x(1-p_a^x)}{N_a} + (p_a^x)^2 \right) \end{aligned} \quad (5.39)$$

Next, the assumption of simple random sampling for the survey data gives us that $\text{cov}(\hat{w}_x, \hat{w}_m) = 0$. So the $(x, m)^{th}$ term in the second sum is

$$\begin{aligned} \mathbb{E}[\hat{p}_a^x \hat{w}_x \hat{p}_a^m \hat{w}_m] - \mathbb{E}[\hat{p}_a^x \hat{w}_x] \mathbb{E}[\hat{p}_a^m \hat{w}_m] &= \mathbb{E}[\hat{p}_a^x \hat{p}_a^m] \mathbb{E}[\hat{w}_x \hat{w}_m] - \mathbb{E}[\hat{p}_a^x] \mathbb{E}[\hat{w}_x] \mathbb{E}[\hat{p}_a^m] \mathbb{E}[\hat{w}_m] \\ &= \{ \mathbb{E}[\hat{p}_a^x \hat{p}_a^m] \mathbb{E}[\hat{w}_x \hat{w}_m] - \mathbb{E}[\hat{p}_a^x \hat{p}_a^m] \mathbb{E}[\hat{w}_x] \mathbb{E}[\hat{w}_m] \} \\ &\quad + \{ \mathbb{E}[\hat{p}_a^x \hat{p}_a^m] \mathbb{E}[\hat{w}_x] \mathbb{E}[\hat{w}_m] - \mathbb{E}[\hat{p}_a^x] \mathbb{E}[\hat{w}_x] \mathbb{E}[\hat{p}_a^m] \mathbb{E}[\hat{w}_m] \} \\ &= \mathbb{E}[\hat{p}_a^x \hat{p}_a^m] \text{cov}(\hat{w}_x, \hat{w}_m) + \mathbb{E}[\hat{w}_x] \mathbb{E}[\hat{w}_m] \text{cov}(\hat{p}_a^x, \hat{p}_a^m) \\ &= \mathbb{E}[\hat{w}_x] \mathbb{E}[\hat{w}_m] \text{cov}(\hat{p}_a^x, \hat{p}_a^m) \\ &= w_x w_m \cdot \left(\frac{-p_a^x p_a^m}{N_a} \right) \end{aligned} \quad (5.40)$$

where the final equality follows from the fact that the vector $(N_a^1, \dots, N_a^x, \dots, N_a^k)$ is multinomially distributed with N_a trials and probabilities $(p_a^1, \dots, p_a^x, \dots, p_a^k)$,

so that

$$\begin{aligned}
 \text{cov}(\widehat{p}_a^x, \widehat{p}_a^m) &= \text{cov}\left(\frac{N_a^x}{N_a}, \frac{N_a^m}{N_a}\right) \\
 &= \frac{1}{(N_a)^2} \text{cov}(N_a^x, N_a^m) \\
 &= \frac{1}{(N_a)^2} \cdot -N_a p_a^x p_a^m \\
 &= \frac{-p_a^x p_a^m}{N_a}
 \end{aligned} \tag{5.41}$$

□

Of course, since N_a and s_x are random variables, the above formula (5.37) is an exact finite-sample variance only over those samples with the same census area population, and the same marginal allocation of survey outcomes across the k categories of x , $(s_1, s_2 \dots s_x \dots s_k)$. Thus (5.37) can be an under- or an over-estimate of the true (unconditional) variance. Since $\widehat{\mu}_a$ is unbiased for all possibly configurations of $(N_a)_{a=1}^A$ and $(s_x)_{x=1}^k$, we have

$$\begin{aligned}
 V[\widehat{\mu}_a] &= \mathbb{E}[V[\widehat{\mu}_a | N_1, \dots, N_A, s_1 \dots s_k]] + V[\mathbb{E}[\widehat{\mu}_a | N_1, \dots, N_A, s_1 \dots s_k]] \\
 &= \mathbb{E}[V[\widehat{\mu}_a | N_1, \dots, N_A, s_1 \dots s_k]]
 \end{aligned} \tag{5.42}$$

5.3 Comparison with ELL: A Monte Carlo Experiment

I simulated datasets according to the following process:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \tag{5.43}$$

with the marginal distribution of the (binary) regressor x varying by area a :

$$x_i | a \sim \text{Bernoulli}(p_a^x) \tag{5.44}$$

and endogeneity of the following form:

$$\varepsilon | x \sim N(\lambda(x), \sigma^2) \tag{5.45}$$

I used the parameter values $\beta_0 = 5$, $\beta_1 = 1$, and $\sigma = 1$ to generate log

incomes for households in $A = 4$ areas. I set the population shares at

$$\begin{aligned}\omega_1 &= 0.1 \\ \omega_2 &= 0.3 \\ \omega_3 &= 0.4 \\ \omega_4 &= 0.2\end{aligned}$$

and the area-specific marginal probabilities for x at

$$\begin{aligned}p_1^x &= \mathbb{P}(X = 1|a = 1) = 0.9 \\ p_2^x &= \mathbb{P}(X = 1|a = 2) = 0.4 \\ p_3^x &= \mathbb{P}(X = 1|a = 3) = 0.5 \\ p_4^x &= \mathbb{P}(X = 1|a = 4) = 0.7\end{aligned}$$

implying that

$$p^x = \mathbb{P}(X = 1) = \sum_a \omega_a p_a^x = 0.55 \quad (5.46)$$

I set $\lambda(1) = 1$. The restriction that $\mathbb{E}[\varepsilon] = 0$ then requires that

$$(1 - p^x)\lambda(0) + p^x\lambda(1) = 0 \quad (5.47)$$

i.e. $\lambda(0) = -p^x\lambda(1)/(1 - p^x) = -1.22$. Also, notice that

$$\text{cov}(x, \varepsilon) = \mathbb{E}[x\varepsilon] = p^x\lambda(1) = 0.55 \quad (5.48)$$

so that x is indeed endogenous.

I simulated 100 artificial datasets consisting of independent, unweighted “census samples” with $N_R = 10000$ and 100000 and “survey samples” of size $s = 1000$ and 5000. I calculated the nonparametric and ELL estimates for each area on each simulated dataset to characterize (an approximation to) the sampling distribution of each estimator.

5.3.1 Consistency

As argued in section 5.2.1 above, the nonparametric estimator is consistent for the true headcount despite the endogeneity of the regressor x . We can see this in Figure 5.1 below with the sampling distributions becoming more concentrated about the population value.

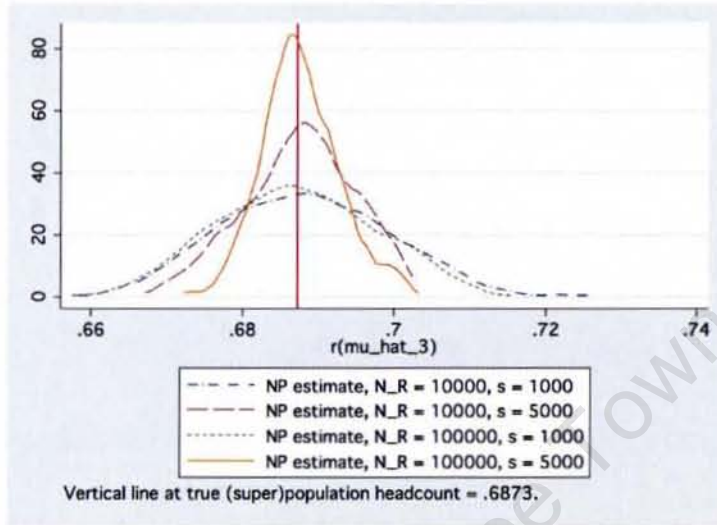


Figure 5.1: Sampling Distributions - Nonparametric Estimator, Area 3

It appears that the size of the survey matters far more for the dispersion of the estimates than does the size of the census sample. If we compare the sampling distributions for given s , increasing N_R tenfold (from 10 000 to 100 000) does not concentrate the distribution much. Increasing the survey size from 1000 to 5000, however, greatly reduces the sampling variation in the estimates. By contrast, the ELL estimator is not consistent in the presence of endogeneity. In Figure 5.2, it is clear that the sampling distributions are collapsing on a point that is very distant from the population value.

We also have consistency at the region level. That the reaggregated nonparametric estimates consistently estimate the region-level headcount should not be surprising: after all, the reaggregated estimate is simply a linear combination of consistent estimates, and continuous functions honour probability limits. In Figure 5.3 below, the collapse of the sampling densities on the true value is evident.

A clear result of these simulations is that the size of the survey sample (s) appears to be far more important than the size of the census (N_R) in determining the variability of the estimates. This parallels the finding of Elbers et al. [2003a], that the “idiosyncratic” error in the census accounts for very

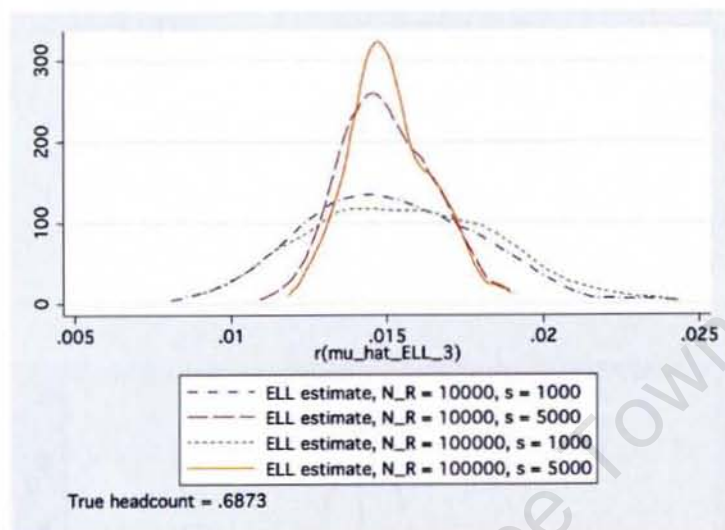


Figure 5.2: Sampling Distributions - ELL Estimator, Area 3

little (typically about 3%) of the total variation in the estimates; sampling error (or “model error” in the parlance of the original paper) accounts for the majority of the uncertainty in the estimates.

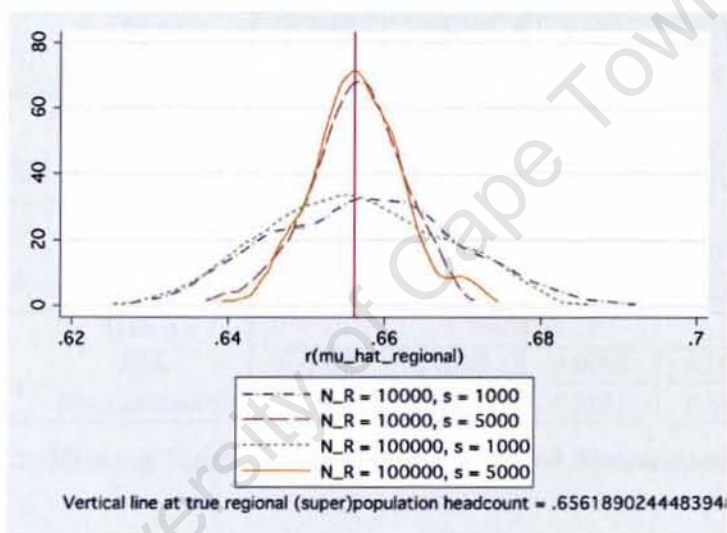


Figure 5.3: Sampling Distributions - Reaggregated Nonparametric Estimates

5.3.2 Bias and Variance

In Table 5.1 below, we see that while the nonparametric estimates are unbiased in each area (at least to two decimal places), the distribution of the ELL estimates are centred on points quite distant from the true area headcounts, bearing out the claim of section 5.2.3 above.

Area		$N_R = 10000$		$N_R = 100000$	
		$s = 1000$	$s = 5000$	$s = 1000$	$s = 5000$
$a = 1$	True μ_a	0.4385			
	ELL	0.003	0.003	0.003	0.003
	Nonparametric	0.4414	0.4381	0.4382	0.4393
$a = 2$	True μ_a	0.7495			
	ELL	0.0182	0.0180	0.0186	0.0181
	Nonparametric	0.751	0.7499	0.7493	0.7496
$a = 3$	True μ_a	0.6873			
	ELL	0.0151	0.015	0.0155	0.0151
	Nonparametric	0.6884	0.6875	0.6870	0.6877
$a = 4$	True μ_a	0.5629			
	ELL	0.0091	0.0089	0.0093	0.0091
	Nonparametric	0.5641	0.5617	0.5631	0.5636

Table 5.1: Means of Sampling Distributions, ELL and Nonparametric Estimates

By inspecting the average (across each of the 100 simulations) of the estimated standard errors for each area (and each sample size), it seems that the estimated standard errors for the ELL estimates are not always a reliable guide to the true variability of that estimator. The extent of this overstatement or understatement varies with the size of the census and survey sample and the characteristics of the area, though: for example, with $N_R = 10000$ and a survey size of $s = 5000$, the estimated standard error for the ELL estimator is 0.0017 in area one, while the true standard error (over superpopulation and sampling variation) is 0.0005, more than three times smaller. However, the estimated standard errors are quite close for area four, especially with larger sample sizes.

Table 5.2 contains summary statistics on the distribution of the estimated standard errors for the ELL estimators. I also report the true variability

of the estimates, $\sqrt{V[\hat{\mu}_a^{ELL}]}$ (as approximated by the standard deviation of the point estimates over the 100 simulations).

Area	$N_R = 10000$		$N_R = 100000$		
	$s = 1000$	$s = 5000$	$s = 1000$	$s = 5000$	
$a = 1$	True $\sqrt{V[\hat{\mu}_a^{ELL}]}$	0.0007	0.0005	0.0006	0.0002
	Mean	0.0017	0.0017	0.0007	0.0005
	Std. Dev.	0.0002	0.0002	0.0001	0.00005
$a = 2$	True $\sqrt{V[\hat{\mu}_a^{ELL}]}$	0.0034	0.0018	0.0035	0.0016
	Mean	0.0038	0.0027	0.0028	0.0014
	Std. Dev.	0.0006	0.0002	0.0007	0.0002
$a = 3$	True $\sqrt{V[\hat{\mu}_a^{ELL}]}$	0.0028	0.0015	0.0029	0.0013
	Mean	0.003	0.0021	0.0023	0.0012
	Std. Dev.	0.0006	0.0002	0.0006	0.0002
$a = 4$	True $\sqrt{V[\hat{\mu}_a^{ELL}]}$	0.0017	0.001	0.0017	0.0008
	Mean	0.0025	0.0021	0.0015	0.0009
	Std. Dev.	0.0004	0.0002	0.0003	0.00009

Table 5.2: Summary Statistics - Estimated Standard Errors, ELL Estimator

On the other hand, the estimated standard errors of the nonparametric estimator are, roughly speaking, accurate. We see this in Table 5.3, where for each area, the average of the estimated standard errors is accurate to four decimal places (where I have used the average true conditional variance as the unconditional variance $V[\hat{\mu}_a]$), with the exception of area one for the smallest sample sizes, where it is accurate to three decimal places.

Encouragingly, the standard deviation (across simulations) of the estimated conditional standard errors is on the order of 10^{-4} or smaller, while the coverage rate (using a confidence interval centred at the point estimate and having half-width of two standard errors) is 100% in all areas at all sample sizes. I conclude from this that the estimated standard errors as given in Theorem 5.4 are a useful and accurate measure of the variability of the nonparametric estimator.

Area		$N_R = 10000$		$N_R = 100000$	
		$s = 1000$	$s = 5000$	$s = 1000$	$s = 5000$
$a = 1$	True $\sqrt{V}[\hat{\mu}_a]$	0.0195	0.0102	0.0187	0.0085
	Mean	0.0194	0.0102	0.0187	0.0085
	Std. Dev.	0.0003	0.00009	0.0003	0.00007
$a = 2$	True $\sqrt{V}[\hat{\mu}_a]$	0.01	0.0066	0.0085	0.0041
	Mean	0.01	0.0066	0.0085	0.0041
	Std. Dev.	0.0002	0.00009	0.0002	0.00003
$a = 3$	True $\sqrt{V}[\hat{\mu}_a]$	0.0115	0.0068	0.0105	0.0049
	Mean	0.0115	0.0068	0.0105	0.0049
	Std. Dev.	0.0002	0.00007	0.0001	0.00004
$a = 4$	True $\sqrt{V}[\hat{\mu}_a]$	0.0158	0.0091	0.0146	0.0068
	Mean	0.0158	0.0091	0.0146	0.0068
	Std. Dev.	0.0003	0.00006	0.0002	0.00004

Table 5.3: Summary Statistics - Estimated (Conditional) Standard Errors, $\sqrt{\hat{V}[\hat{\mu}_a|N_a, s_x]}$

5.4 Empirical Application: Poverty in the Free State

As with any estimation technique, the results are only as plausible as the identifying assumptions. In the case of the nonparametric estimator I have developed in this chapter, the essential assumption is conditional homogeneity - that given the value of the binary regressor x , the density of (log) income does not vary across areas. In general, the survey sample sizes in any given area are too small to verify this assumption (recall, this is the basic reason why it is necessary to incorporate census data into the estimation at all).

I will not attempt to justify the use of any binary regressor over another. I aim merely to illustrate the fact that poverty maps can be estimated by this nonparametric technique, and with a very high degree of precision - given, of course, that the above assumptions are satisfied.

In Table 5.4 below, I present the estimated poverty headcounts for magisterial districts in the Free State using the same poverty line and data as in Alderman et al. [2002]. I include the ELL estimates (using the maximal specification, i.e. all available regressors), as well as two sets of nonparametric estimates. These estimates were calculated using as the (binary) predictor x the indicators for “formal dwelling” and “household head is female”. The estimated standard errors of each area-level headcount are included in parentheses. The ELL estimates were calculated by simulating the “location effect” at the region level, which is why their standard errors are so high. (See Chapter 3 for a full discussion of the computational implementation of the ELL technique.)

Magisterial District	$\hat{\mu}_a^{ELL}$	Formal Dwelling	Female Head
Boshof	0.5934 (0.1694)	0.5605 (0.0015)	0.4788 (0.0003)
Jacobsdal	0.5469 (0.1613)	0.4805 (0.0021)	0.4795 (0.0003)
Koffiefontein	0.3792 (0.1529)	0.4619 (0.0018)	0.4790 (0.0003)
Fauresmith	0.4263 (0.1753)	0.4230 (0.0017)	0.4788 (0.0003)
Petrusburg	0.5041 (0.1812)	0.5024 (0.0023)	0.4787 (0.0003)
Odendaalsrus	0.3674 (0.1345)	0.6357 (0.0009)	0.4786 (0.0003)
Virginia	0.2895 (0.1319)	0.6274 (0.0010)	0.4788 (0.0003)
Welkom	0.2975 (0.1176)	0.5515 (0.0006)	0.4788 (0.0003)

Continued on next page...

Table 5.4 (continued from previous page)

Magisterial District	$\hat{\mu}_a^{ELL}$	Formal Dwelling	Female Head
Bothaville	0.5013 (0.1685)	0.5958 (0.0011)	0.4790 (0.0003)
Bultfontein	0.5290 (0.1836)	0.6009 (0.0014)	0.4785 (0.0003)
Heilbron	0.5227 (0.1653)	0.6035 (0.0013)	0.4786 (0.0003)
Hennenman	0.3754 (0.1553)	0.5955 (0.0016)	0.4789 (0.0003)
Hoopstad	0.5166 (0.1824)	0.5557 (0.0017)	0.4786 (0.0003)
Koppies	0.6600 (0.1562)	0.6573 (0.0019)	0.4788 (0.0003)
Kroonstad	0.3417 (0.1469)	0.5085 (0.0008)	0.4785 (0.0003)
Parys	0.4340 (0.1762)	0.5822 (0.0012)	0.4785 (0.0003)
Theunissen	0.4842 (0.1652)	0.6339 (0.0015)	0.4787 (0.0003)
Ventersburg	0.4779 (0.1663)	0.6750 (0.0022)	0.4784 (0.0003)
Vredefort	0.5580 (0.1805)	0.6250 (0.0021)	0.4788 (0.0003)
Viljoenskroon	0.5263 (0.1669)	0.5870 (0.0013)	0.4790 (0.0003)
Wesselsbron	0.5277 (0.2042)	0.6540 (0.0014)	0.4788 (0.0003)
Bethlehem	0.4187 (0.1455)	0.5545 (0.0009)	0.4787 (0.0003)
Ficksburg	0.5775 (0.1589)	0.6681 (0.0012)	0.4782 (0.0003)
Fouriesburg	0.5264 (0.1718)	0.7014 (0.0019)	0.4782 (0.0003)
Frankfort	0.5018 (0.1600)	0.6548 (0.0012)	0.4787 (0.0003)
Harrismith	0.4335 (0.1758)	0.5702 (0.0010)	0.4787 (0.0003)
Lindley	0.5078 (0.1819)	0.6455 (0.0013)	0.4783 (0.0003)
Reitz	0.4900 (0.1736)	0.6525 (0.0015)	0.4789 (0.0003)
Senekal	0.5318 (0.1653)	0.6540 (0.0013)	0.4785 (0.0003)
Vrede	0.4896 (0.1830)	0.6464 (0.0014)	0.4785 (0.0003)
Brandfort	0.5089 (0.1616)	0.6008 (0.0017)	0.4786 (0.0003)
Clocolan	0.4911 (0.1875)	0.6363 (0.0018)	0.4785 (0.0003)
Dewetsdorp	0.4968 (0.1899)	0.5250 (0.0021)	0.4783 (0.0003)
Edenburg	0.3978 (0.1828)	0.4713 (0.0025)	0.4780 (0.0003)
Excelesior	0.4463 (0.1875)	0.6502 (0.0019)	0.4784 (0.0003)
Jagersfontein	0.4706 (0.1445)	0.4059 (0.0018)	0.4783 (0.0003)
Ladybrand	0.4221 (0.1531)	0.5670 (0.0015)	0.4784 (0.0003)
Marquard	0.5226 (0.1732)	0.6455 (0.0020)	0.4786 (0.0003)
Philippolis	0.4719 (0.1556)	0.4688 (0.0026)	0.4786 (0.0003)
Reddersburg	0.4203 (0.1666)	0.4717 (0.0027)	0.4791 (0.0003)
Trompsburg	0.4787 (0.1643)	0.4283 (0.0023)	0.4781 (0.0003)
Wepener	0.4506 (0.1811)	0.5331 (0.0021)	0.4784 (0.0003)
Winburg	0.4521 (0.1706)	0.5806 (0.0021)	0.4783 (0.0003)
Botshabelo	0.3646 (0.1767)	0.5568 (0.0006)	0.4779 (0.0003)

Continued on next page...

Table 5.4 (continued from previous page)

Magisterial District	$\widehat{\mu}_a^{ELL}$	Formal Dwelling	Female Head
Bloemfontein	0.2478 (0.1141)	0.5020 (0.0005)	0.4786 (0.0003)
Smithfield	0.5633 (0.1568)	0.4890 (0.0027)	0.4785 (0.0003)
Bethulie	0.3916 (0.1767)	0.4103 (0.0013)	0.4781 (0.0003)
Rouxville	0.4125 (0.1925)	0.4939 (0.0024)	0.4786 (0.0003)
Zastron	0.4717 (0.1984)	0.5686 (0.0019)	0.4782 (0.0003)
Sasolburg	0.2910 (0.1363)	0.5679 (0.0008)	0.4792 (0.0003)
Thaba'Nchu	0.3940 (0.1767)	0.5258 (0.0009)	0.4781 (0.0003)
Witsieshoek	0.5216 (0.2052)	0.6031 (0.0005)	0.4775 (0.0004)

Table 5.4: Estimated Headcounts Based on Alternative Predictors

From Table 5.4 two facts are immediately apparent. One is that the nonparametric estimates do not “agree” with one another, nor with the ELL estimates. Again, I emphasize that the importance of the specification choice (and by implication, the identifying assumptions) cannot be escaped. Different specifications will lead to different estimates at the area level. Indeed, when the districts are ranked in descending order by the estimated headcounts, the correlation between the ranks was not very high: the ranks according to the ELL estimates and the nonparametric estimates using the formal dwelling dummy, for example, is only 0.4282. The situation is even worse with the two nonparametric estimates: the implied ranks using the female-head dummy and the formal dwelling dummy have a correlation of just 0.008!

The second obvious feature of Table 5.4 is the compressed variation in the nonparametric estimates when the female-head dummy is used: the only differences between area estimates occur in the third decimal place. If these estimates are accurate, the largest possible pairwise difference in the headcount between areas in the Free State was 0.9% in 1995! The reason for this is clear from the conditional income distributions, illustrated in Figure 5.4 below: there is little difference in the probability mass below the poverty line, so $\mathbb{E}[w(y)|X = 0]$ will not be very different from $\mathbb{E}[w(y)|X = 1]$, even if intra-regional variation in p_a^x is high. It is easy to see why this is the case: if conditional homogeneity holds and x is independent of y , then in fact *unconditional* homogeneity holds, since then $f(y|x) = f(y)$ is constant across areas.

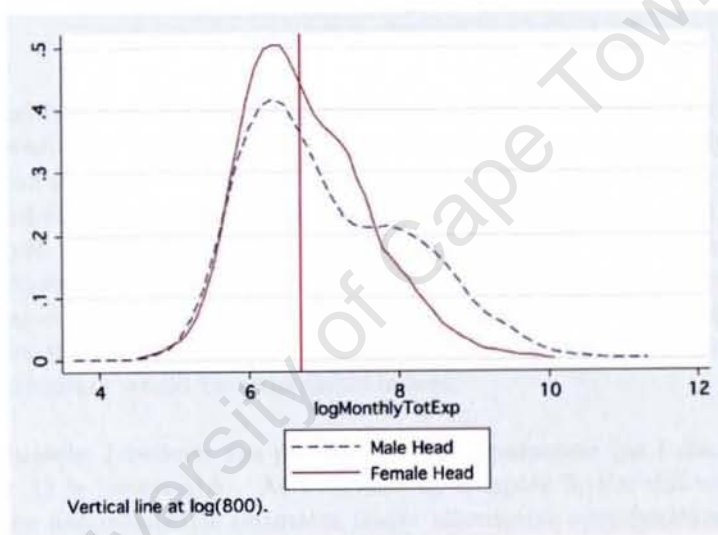


Figure 5.4: (log) Consumption, by Gender of Head

Chapter 6

Conclusions

I have exhibited, in Chapter 5, a numerically simple nonparametric estimator of small-area welfare functionals that is consistent under strictly weaker conditions than are needed to justify the use of ELL. I have shown that the estimated standard errors for this estimator are both accurate and small in magnitude. The primary drawback of this estimator, as I discuss below, is the requirement of a single discrete variable that is both relevant and is such that conditional homogeneity holds. Given the deficiencies of present models of the data-generating process for expenditure and income, finding such a covariate would be remarkable indeed.

Unfortunately, I believe, the problem of data sparseness (as I discussed in Chapter 1) is inescapable. As I showed in Chapter 5, the differences between the nonparametric estimates under alternative specifications (choice of x) can be significant. Furthermore, the nonparametric estimates can generate implausibly low intra-regional variability (under certain specification choices).

I think these results underline the difficulty of small-area estimation. For the nonparametric estimates to be consistent, we require conditional homogeneity - that but for the marginal distribution of the variable x , there are no differences between areas in their income distribution. It is very hard to believe that any single variable can capture all the relevant differences between areas: having a formal dwelling in an urban area like Bloemfontein, for example, is very likely less indicative of wealth than it is in a rural district like Koppies. Of course, the difficulty of finding a suitable "predictor" x is made even worse by the fact that the available discrete variables are

broad categories - for example, a “formal dwelling” can be anything from a simple brick hut to an extravagant mansion, so that the type of households that have formal dwellings in remote rural areas are systematically different to the type of households that have formal dwellings in wealthy suburbs - despite being identical on observables x , those two types of households will not have the same $\mathbb{E}[w(y)|X]$.

However, I claim that ELL also suffers from this weakness. The exogeneity requirement (which, as I have argued in Chapter 4, is needed for ELL to consistently estimate small-area welfare measures), is a stronger version of the conditional homogeneity assumption. Roughly speaking, we need to “control for” a number of variables in a model of consumption (or income), and linear regression (as in ELL) is a very imperfect way to do this. The nonparametric estimator I present here takes a more direct approach in that it assumes that we have a single variable that captures all of the relevant inter-area variation rather than a vector of them, but it forgoes the functional-form assumptions needed to accommodate the higher-dimensional nature of the covariates in ELL.

Absent more extensive high-quality data on consumption or income, estimating a convincing and accurate poverty map requires strong exogeneity (conditional homogeneity) assumptions. The scope for numerical techniques to reveal unobserved differences between areas is, I believe, quite limited.

Appendix A

Further Results

A.1 Summary Statistics: First-Stage R^2 Values

I have argued above that the poverty mapping literature imposes no restrictions on the specification of the first-stage model other than it have a high R^2 statistic. To show that I really have obeyed the methodological prescriptions of the literature, I document the R^2 values I obtained across all my estimations. Table A.1 summarises these values over the 459 estimations (9 provinces \times [50 random specifications +1 maximal model]) estimations from section 4.1.

Province	Mean	Std. Dev.	Min.	Max.
Western Cape	0.5578	0.036	0.4903	0.6175
Eastern Cape	0.5885	0.0271	0.5215	0.6289
Northern Cape	0.5926	0.0376	0.442	0.6429
Free State	0.6154	0.0254	0.5536	0.6518
KwaZulu-Natal	0.5334	0.024	0.469	0.5743
North West	0.5780	0.0179	0.5387	0.6319
Gauteng	0.5600	0.0213	0.5173	0.6031
Mpumalanga	0.5472	0.0203	0.499	0.5886
Limpopo	0.5221	0.0218	0.4626	0.5644

Table A.1: Summary Statistics: R^2 Values Over Alternative Specifications

In section 4.2 I calculated 3600 (9 provinces \times 200 bootstrap replications \times 2 specifications) ELL estimates. In Table A.2, I display the summary statistics

over those first-stage regressions, broken down by province. The results are encouraging (at least, by the standards of the poverty-mapping literature): all provinces have mean R^2 -values over 0.5, and *no* first-stage model obtains an R^2 lower than 0.47.

University of Cape Town

Province	Specification 1				Specification 2			
	Mean	Std. Dev.	Min.	Max.	Mean	Std. Dev.	Min.	Max.
W Cape	0.5748	0.0105	0.5497	0.6038	0.5377	0.0119	0.4982	0.5712
E Cape	0.5808	0.0086	0.5591	0.6064	0.5801	0.0087	0.547	0.6044
N Cape	0.6212	0.0161	0.5820	0.6688	0.5958	0.017	0.5506	0.6417
Free State	0.6255	0.0102	0.5900	0.6625	0.624	0.0102	0.5940	0.6571
KwaZulu-Natal	0.5453	0.0097	0.5161	0.5710	0.5235	0.0099	0.4978	0.5546
North West	0.5713	0.0159	0.5315	0.6171	0.6102	0.0141	0.5752	0.6506
Gauteng	0.5765	0.0102	0.5416	0.6049	0.5308	0.0116	0.4963	0.5603
Mpumalanga	0.5598	0.0128	0.5263	0.5933	0.5167	0.0138	0.4708	0.5577
Limpopo	0.5112	0.0135	0.4759	0.545	0.5115	0.0128	0.4757	0.5505

Table A.2: R^2 Over 200 Bootstrap Repetitions, by Province

A.2 Sensitivity: Area Headcount

Magisterial District	HC (maximal)	Mean	Min	Max	Range	IQR
Murraysburg	0.369	0.325	0.186	0.485	0.298	0.087
Uniondale	0.327	0.319	0.218	0.498	0.279	0.079
Calitzdorp	0.286	0.286	0.215	0.406	0.191	0.053
Prince Albert	0.275	0.275	0.188	0.428	0.240	0.054
Robertson	0.261	0.262	0.189	0.376	0.187	0.056
Swellendam	0.244	0.226	0.154	0.327	0.174	0.055
Van Rhynsdorp	0.236	0.210	0.139	0.289	0.150	0.072
Laingsburg	0.215	0.230	0.157	0.368	0.211	0.069
Worcester	0.198	0.190	0.120	0.290	0.170	0.045
Vredendal	0.191	0.212	0.142	0.370	0.228	0.047
Knysna	0.184	0.168	0.116	0.245	0.130	0.036
Moorreesburg	0.181	0.187	0.125	0.257	0.132	0.048
Hopefield	0.177	0.142	0.082	0.196	0.114	0.035
Montagu	0.174	0.175	0.115	0.278	0.163	0.049
Clanwilliam	0.169	0.193	0.124	0.310	0.186	0.051
Tulbagh	0.166	0.167	0.108	0.302	0.194	0.043
Riversdal	0.165	0.174	0.113	0.300	0.187	0.044
Caledon	0.163	0.175	0.114	0.256	0.143	0.049
Beaufort West	0.151	0.171	0.094	0.324	0.230	0.042
Ceres	0.149	0.174	0.113	0.299	0.186	0.051
Heidelberg	0.146	0.188	0.117	0.352	0.235	0.038
Oudtshoorn	0.146	0.137	0.072	0.242	0.171	0.038
Ladismith	0.138	0.170	0.093	0.287	0.194	0.062
George	0.134	0.135	0.093	0.209	0.116	0.042
Mitchellsplain	0.126	0.143	0.095	0.223	0.128	0.045
Piketberg	0.124	0.146	0.089	0.254	0.165	0.053
Mossel bay	0.122	0.126	0.071	0.177	0.106	0.037
Stellenbosch	0.119	0.110	0.058	0.169	0.111	0.032
Bredasdorp	0.114	0.121	0.065	0.178	0.113	0.038
Hermanus	0.112	0.122	0.075	0.175	0.101	0.030
Wellington	0.094	0.099	0.064	0.153	0.089	0.035
Paarl	0.092	0.114	0.067	0.190	0.123	0.041
Malmesbury	0.088	0.101	0.060	0.146	0.086	0.024
Strand	0.086	0.091	0.045	0.135	0.090	0.034
Goodwood	0.075	0.066	0.032	0.107	0.075	0.021

Continued on next page...

Table A.3 (continued from previous page)

Magisterial District	HC (maximal)	Mean	Min	Max	Range	IQR
Kuilsrivier	0.064	0.070	0.040	0.116	0.076	0.023
Vredenburg	0.063	0.082	0.042	0.137	0.094	0.025
Simonstown	0.053	0.063	0.039	0.098	0.059	0.023
Somerset West	0.052	0.055	0.022	0.083	0.061	0.022
Bellville	0.038	0.036	0.014	0.069	0.055	0.018
Cape	0.032	0.043	0.026	0.070	0.044	0.014
Wynberg	0.027	0.031	0.011	0.073	0.062	0.014

Table A.3: Estimates Over 50 Random Specifications, W
Cape

Magisterial District	HC (maximal)	Mean	Min	Max	Range	IQR
Mqanduli	0.656	0.609	0.509	0.696	0.187	0.049
Elliotdale	0.638	0.647	0.557	0.726	0.168	0.054
Tabankulu	0.636	0.608	0.547	0.686	0.139	0.044
Flagstaff	0.634	0.617	0.535	0.700	0.164	0.061
Kentani	0.625	0.609	0.541	0.685	0.144	0.044
Umzimkulu	0.614	0.585	0.509	0.715	0.206	0.047
Cala	0.605	0.581	0.521	0.677	0.156	0.047
Lusikisiki	0.602	0.594	0.531	0.676	0.145	0.051
Ngqueleni	0.599	0.608	0.526	0.692	0.167	0.050
Engcobo	0.594	0.591	0.536	0.669	0.133	0.039
Qumbu	0.584	0.583	0.504	0.676	0.172	0.051
Middeldrift	0.578	0.571	0.499	0.675	0.176	0.064
Tsomo	0.575	0.597	0.488	0.710	0.222	0.047
Mt Fletcher	0.574	0.598	0.512	0.689	0.177	0.049
Cofimvaba	0.573	0.574	0.516	0.643	0.127	0.039
Bizana	0.571	0.577	0.464	0.677	0.213	0.059
Libode	0.570	0.586	0.522	0.681	0.159	0.056
Mt Ayliff	0.565	0.581	0.522	0.720	0.197	0.062
Mt Frere	0.565	0.564	0.494	0.665	0.171	0.065
Maluti	0.565	0.573	0.497	0.682	0.185	0.061
Idutywa	0.563	0.573	0.485	0.664	0.178	0.056
Nqamakwe	0.557	0.559	0.466	0.684	0.218	0.064
Willowvale	0.557	0.582	0.516	0.670	0.154	0.062
Tsolo	0.553	0.566	0.485	0.640	0.155	0.052

Continued on next page...

Table A.4 (continued from previous page)

Magisterial District	HC (maximal)	Mean	Min	Max	Range	IQR
Pearston	0.550	0.517	0.359	0.611	0.252	0.051
Mpofu	0.546	0.524	0.397	0.686	0.290	0.114
Port St Johns	0.535	0.573	0.506	0.638	0.132	0.064
Lady Frere	0.503	0.508	0.444	0.558	0.114	0.037
Steytlerville	0.501	0.464	0.372	0.553	0.181	0.060
Umtata	0.489	0.477	0.430	0.538	0.108	0.036
Hofmeyer	0.484	0.490	0.384	0.584	0.200	0.067
Ntabathemba	0.475	0.452	0.376	0.575	0.199	0.050
Sterkspruit	0.471	0.509	0.424	0.594	0.171	0.053
Maclear	0.465	0.486	0.401	0.586	0.186	0.070
Bedford	0.458	0.461	0.365	0.577	0.212	0.050
Hankey	0.458	0.413	0.317	0.544	0.227	0.067
Wodehouse	0.457	0.454	0.380	0.528	0.148	0.057
Victoria East	0.457	0.439	0.380	0.534	0.154	0.056
Sterkstroom	0.455	0.451	0.353	0.533	0.180	0.050
Peddie	0.452	0.465	0.383	0.542	0.159	0.046
Keiskammahoek	0.443	0.451	0.335	0.601	0.265	0.088
Barkley-East	0.437	0.446	0.376	0.508	0.132	0.057
Steynsburg	0.434	0.442	0.362	0.546	0.185	0.061
Komga	0.434	0.499	0.366	0.632	0.266	0.078
Butterworth	0.427	0.426	0.348	0.515	0.167	0.059
Adelaide	0.427	0.411	0.310	0.510	0.200	0.063
Hewu	0.412	0.418	0.343	0.481	0.138	0.059
Jansenville	0.408	0.416	0.353	0.510	0.157	0.064
Lady Grey	0.403	0.450	0.391	0.533	0.142	0.029
Stutterheim	0.398	0.413	0.306	0.506	0.200	0.063
Willowmore	0.395	0.395	0.282	0.493	0.212	0.089
Zwelitsha	0.394	0.432	0.343	0.523	0.180	0.087
Alexandria	0.388	0.419	0.357	0.508	0.150	0.064
Somerset East	0.384	0.400	0.337	0.486	0.149	0.037
Bathurst	0.377	0.393	0.293	0.501	0.208	0.045
Kirkwood	0.373	0.398	0.326	0.502	0.176	0.065
Molteno	0.372	0.390	0.296	0.466	0.170	0.066
Tarka	0.365	0.409	0.305	0.503	0.198	0.050
Fort Beaufort	0.364	0.378	0.298	0.479	0.180	0.065
Cradock	0.358	0.362	0.298	0.417	0.120	0.047

Continued on next page...

Table A.4 (continued from previous page)

Magisterial District	HC (maximal)	Mean	Min	Max	Range	IQR
Albert	0.353	0.365	0.316	0.419	0.103	0.034
Cathcart	0.337	0.347	0.270	0.428	0.157	0.050
Indwe	0.335	0.367	0.285	0.452	0.168	0.044
Elliot	0.327	0.349	0.284	0.424	0.140	0.056
Venterstad	0.326	0.336	0.272	0.411	0.139	0.056
Aberdeen	0.325	0.329	0.229	0.418	0.189	0.057
Aliwal North	0.315	0.319	0.253	0.401	0.148	0.046
East-London	0.294	0.310	0.248	0.385	0.137	0.037
Mdantsane	0.292	0.303	0.249	0.369	0.120	0.035
Joubertina	0.286	0.353	0.250	0.471	0.221	0.065
Humansdorp	0.279	0.277	0.184	0.380	0.196	0.051
Queenstown	0.241	0.244	0.189	0.318	0.129	0.042
Albany	0.238	0.272	0.207	0.390	0.183	0.040
Middelburg	0.237	0.245	0.201	0.320	0.118	0.046
Graaff-Reinet	0.229	0.242	0.174	0.330	0.156	0.043
Uitenhage	0.203	0.233	0.181	0.300	0.119	0.031
Port Elizabeth	0.166	0.189	0.138	0.275	0.137	0.030
King William's Town	0.113	0.143	0.105	0.217	0.112	0.030

Table A.4: Estimates Over 50 Random Specifications, E
Cape

Magisterial District	HC (maximal)	Mean	Min	Max	Range	IQR
Noupoort	0.598	0.569	0.439	0.697	0.258	0.060
Warrenton	0.533	0.508	0.380	0.606	0.226	0.071
Hanover	0.516	0.467	0.259	0.647	0.388	0.082
Philipstown	0.472	0.431	0.299	0.522	0.223	0.062
Fraserburg	0.454	0.449	0.326	0.567	0.241	0.088
Richmond	0.444	0.467	0.395	0.537	0.142	0.058
Barkley-West	0.414	0.378	0.287	0.499	0.212	0.076
Prieska	0.400	0.380	0.294	0.462	0.168	0.053
Williston	0.376	0.358	0.275	0.472	0.197	0.054
Hartswater	0.369	0.403	0.323	0.528	0.204	0.043
Britstown	0.364	0.351	0.259	0.459	0.200	0.055
Calvinia	0.363	0.335	0.238	0.426	0.188	0.059
Hopetown	0.360	0.369	0.254	0.515	0.261	0.083

Continued on next page...

Table A.5 (continued from previous page)

Magisterial District	HC (maximal)	Mean	Min	Max	Range	IQR
Herbert	0.359	0.351	0.257	0.444	0.187	0.063
Sutherland	0.337	0.357	0.255	0.537	0.282	0.075
Kenhardt	0.287	0.302	0.197	0.415	0.218	0.076
Kuruman	0.287	0.331	0.252	0.458	0.206	0.069
Victoria-West	0.287	0.293	0.154	0.462	0.308	0.087
Carnarvon	0.287	0.314	0.222	0.440	0.218	0.081
Postmasburg	0.283	0.309	0.222	0.442	0.220	0.068
Hay	0.272	0.309	0.218	0.450	0.231	0.071
De Aar	0.268	0.281	0.184	0.403	0.219	0.056
Colesberg	0.246	0.299	0.202	0.414	0.212	0.048
Gordonia	0.241	0.256	0.177	0.391	0.214	0.052
Kimberley	0.203	0.233	0.174	0.338	0.164	0.044
Namakwaland	0.077	0.114	0.040	0.225	0.185	0.054

Table A.5: Estimates Over 50 Random Specifications, N
Cape

Magisterial District	HC (maximal)	Mean	Min	Max	Range	IQR
Koppies	0.622	0.604	0.500	0.691	0.192	0.053
Smithfield	0.568	0.550	0.458	0.604	0.147	0.040
Hoopstad	0.567	0.533	0.440	0.625	0.186	0.035
Vredefort	0.567	0.551	0.463	0.620	0.157	0.049
Boshof	0.565	0.552	0.473	0.634	0.161	0.048
Jacobsdal	0.564	0.533	0.467	0.607	0.140	0.055
Viljoenskroon	0.564	0.521	0.448	0.596	0.148	0.041
Ficksburg	0.545	0.578	0.456	0.715	0.258	0.090
Trompsburg	0.536	0.499	0.430	0.594	0.163	0.047
Dewetsdorp	0.524	0.504	0.417	0.609	0.192	0.042
Marquard	0.524	0.530	0.465	0.611	0.146	0.054
Clocolan	0.520	0.506	0.433	0.549	0.116	0.050
Wesselsbron	0.519	0.523	0.450	0.592	0.142	0.045
Reitz	0.519	0.526	0.464	0.608	0.144	0.051
Senekal	0.519	0.515	0.448	0.586	0.138	0.042
Heilbron	0.518	0.520	0.434	0.601	0.167	0.048
Frankfort	0.517	0.487	0.421	0.619	0.198	0.038
Bothaville	0.511	0.515	0.441	0.611	0.170	0.058

Continued on next page...

Table A.6 (continued from previous page)

Magisterial District	HC (maximal)	Mean	Min	Max	Range	IQR
Fouriesburg	0.509	0.528	0.448	0.636	0.189	0.062
Bultfontein	0.499	0.493	0.427	0.571	0.144	0.042
Theunissen	0.499	0.485	0.418	0.575	0.158	0.038
Ventersburg	0.491	0.476	0.405	0.543	0.138	0.047
Brandfort	0.484	0.491	0.408	0.549	0.141	0.048
Petrusburg	0.482	0.489	0.424	0.569	0.145	0.054
Vrede	0.481	0.527	0.432	0.648	0.216	0.052
Excelsior	0.477	0.471	0.402	0.592	0.190	0.051
Harrismith	0.476	0.471	0.392	0.539	0.147	0.046
Philippolis	0.476	0.472	0.391	0.538	0.148	0.038
Wepener	0.471	0.471	0.412	0.568	0.156	0.041
Fauresmith	0.469	0.484	0.373	0.567	0.194	0.059
Lindley	0.467	0.476	0.382	0.550	0.168	0.037
Zastron	0.457	0.473	0.378	0.579	0.201	0.060
Witsieshoek	0.454	0.508	0.396	0.633	0.238	0.065
Winburg	0.444	0.441	0.393	0.537	0.144	0.041
Jagersfontein	0.441	0.457	0.339	0.528	0.190	0.051
Ladybrand	0.434	0.442	0.392	0.551	0.159	0.043
Edenburg	0.434	0.434	0.378	0.527	0.148	0.045
Rouxville	0.429	0.470	0.387	0.560	0.174	0.058
Reddersburg	0.425	0.418	0.340	0.501	0.161	0.062
Bethlehem	0.422	0.410	0.349	0.488	0.139	0.038
Koffiefontein	0.411	0.398	0.342	0.494	0.152	0.048
Parys	0.409	0.406	0.324	0.512	0.188	0.068
Botshabelo	0.408	0.386	0.287	0.476	0.189	0.078
Thaba 'Nchu	0.395	0.431	0.336	0.534	0.198	0.062
Hennenman	0.386	0.370	0.304	0.451	0.147	0.044
Kroonstad	0.362	0.353	0.300	0.418	0.118	0.048
Odendaalsrus	0.359	0.370	0.295	0.474	0.179	0.045
Bethulie	0.354	0.384	0.312	0.441	0.128	0.032
Virginia	0.305	0.295	0.193	0.369	0.176	0.063
Welkom	0.291	0.280	0.218	0.344	0.126	0.045
Sasolburg	0.288	0.302	0.241	0.374	0.133	0.035
Bloemfontein	0.247	0.242	0.185	0.300	0.115	0.028

Table A.6: Estimates Over 50 Random Specifications, Free State

Magisterial District	HC (maximal)	Mean	Min	Max	Range	IQR
Weenen	0.600	0.541	0.423	0.620	0.197	0.079
Ngotshe	0.571	0.509	0.335	0.673	0.338	0.093
Underberg	0.482	0.454	0.340	0.591	0.251	0.033
Utrecht	0.446	0.420	0.324	0.626	0.302	0.072
Paulpietersburg	0.411	0.379	0.221	0.569	0.349	0.131
Kranskop	0.406	0.381	0.255	0.538	0.283	0.065
Mount Currie	0.383	0.382	0.316	0.483	0.167	0.042
New Hanover	0.383	0.363	0.261	0.461	0.200	0.049
Msinga	0.364	0.364	0.299	0.450	0.151	0.036
Polela	0.363	0.306	0.222	0.373	0.151	0.062
Mthonjaneni	0.353	0.309	0.242	0.474	0.232	0.058
Ixopo	0.351	0.351	0.256	0.445	0.190	0.048
Alfred	0.346	0.323	0.243	0.396	0.153	0.044
Nkandla	0.338	0.317	0.235	0.419	0.184	0.047
Mooi river	0.321	0.323	0.217	0.465	0.249	0.081
Umvoti	0.319	0.327	0.269	0.380	0.111	0.038
Richmond	0.319	0.326	0.220	0.430	0.210	0.081
Babanango	0.305	0.299	0.221	0.410	0.189	0.052
Umzinto	0.301	0.292	0.237	0.371	0.134	0.036
Lower Tugela	0.293	0.282	0.215	0.343	0.128	0.043
Simdlangentsha	0.286	0.266	0.196	0.371	0.175	0.042
Nongoma	0.285	0.265	0.189	0.358	0.168	0.051
Vryheid	0.283	0.272	0.213	0.357	0.144	0.058
Nqutu	0.282	0.265	0.213	0.345	0.132	0.032
Mhlabathini	0.266	0.258	0.178	0.343	0.165	0.046
Mapumulo	0.263	0.257	0.214	0.324	0.110	0.041
Bergville	0.244	0.235	0.179	0.306	0.127	0.033
Estcourt	0.243	0.222	0.159	0.305	0.146	0.038
Ubombo	0.236	0.261	0.198	0.401	0.203	0.037
Eshowe	0.232	0.227	0.187	0.267	0.081	0.027
Kliprivier	0.231	0.246	0.202	0.315	0.113	0.038
Impendle	0.225	0.232	0.174	0.326	0.152	0.052
Dundee	0.218	0.227	0.176	0.286	0.110	0.038
Dannhauser	0.217	0.209	0.141	0.296	0.155	0.044
Hlabisa	0.208	0.231	0.182	0.319	0.137	0.042
Ndwedwe	0.198	0.197	0.144	0.280	0.136	0.042
Ingwavuma	0.192	0.234	0.169	0.328	0.159	0.043

Continued on next page...

Table A.7 (continued from previous page)

Magisterial District	HC (maximal)	Mean	Min	Max	Range	IQR
Umbumbulu	0.187	0.188	0.134	0.275	0.141	0.050
Port Shepstone	0.181	0.182	0.136	0.240	0.104	0.032
Inanda	0.173	0.171	0.125	0.243	0.118	0.038
Mtunzini	0.167	0.176	0.130	0.243	0.114	0.026
Glencoe	0.165	0.184	0.122	0.270	0.148	0.040
Lions River	0.160	0.168	0.115	0.239	0.124	0.031
Lower Umfolozi	0.158	0.157	0.116	0.238	0.122	0.035
Newcastle	0.157	0.149	0.096	0.237	0.141	0.028
Umlazi	0.154	0.137	0.079	0.193	0.115	0.033
Camperdown	0.145	0.166	0.117	0.235	0.117	0.037
Pietermaritzburg	0.136	0.133	0.086	0.182	0.096	0.028
Pinetown	0.096	0.108	0.078	0.146	0.067	0.021
Durban	0.069	0.062	0.043	0.081	0.038	0.014
Chatswoth	0.061	0.068	0.042	0.098	0.057	0.019

Table A.7: Estimates Over 50 Random Specifications,
KwaZulu-Natal

Magisterial District	HC (maximal)	Mean	Min	Max	Range	IQR
Bronkhorstspuit	0.320	0.305	0.221	0.376	0.155	0.045
Nigel	0.212	0.185	0.099	0.247	0.149	0.033
Cullinan	0.193	0.208	0.103	0.326	0.223	0.056
Oberholzer	0.186	0.176	0.085	0.253	0.168	0.052
Westonaria	0.173	0.160	0.074	0.217	0.143	0.058
Heidelberg	0.172	0.189	0.123	0.297	0.175	0.046
Vanderbijlpark	0.150	0.127	0.082	0.190	0.108	0.024
Randfontein	0.140	0.156	0.092	0.213	0.121	0.031
Vereeniging	0.133	0.126	0.079	0.186	0.107	0.029
Brakpan	0.131	0.126	0.076	0.175	0.099	0.026
Benoni	0.129	0.125	0.089	0.162	0.073	0.029
Randburg	0.119	0.107	0.079	0.138	0.058	0.016
Soshanguve	0.119	0.132	0.063	0.215	0.153	0.038
Alberton	0.113	0.104	0.074	0.157	0.083	0.022
Kempton Park	0.111	0.108	0.073	0.146	0.072	0.027
Krugersdorp	0.102	0.107	0.070	0.146	0.075	0.015
Boksburg	0.089	0.079	0.051	0.104	0.054	0.022

Continued on next page...

Table A.8 (continued from previous page)

Magisterial District	HC (maximal)	Mean	Min	Max	Range	IQR
Springs	0.084	0.068	0.049	0.129	0.080	0.016
Wonderboom	0.081	0.083	0.063	0.117	0.054	0.015
Roodepoort	0.079	0.072	0.044	0.094	0.050	0.017
Johannesburg	0.060	0.055	0.029	0.080	0.051	0.008
Soweto	0.048	0.065	0.025	0.120	0.095	0.030
Germiston	0.047	0.051	0.022	0.083	0.061	0.012
Pretoria	0.047	0.035	0.022	0.046	0.024	0.008

Table A.8: Estimates Over 50 Random Specifications, Gauteng

Magisterial District	HC (maximal)	Mean	Min	Max	Range	IQR
Carolina	0.651	0.601	0.390	0.723	0.334	0.097
Eerstehoek	0.597	0.526	0.360	0.641	0.282	0.081
Bethal	0.580	0.514	0.340	0.658	0.318	0.103
Waterval Boven	0.507	0.478	0.345	0.543	0.198	0.062
Amersfoort	0.433	0.402	0.242	0.591	0.349	0.103
Ermelo	0.403	0.397	0.227	0.528	0.301	0.078
Balfour	0.347	0.305	0.190	0.412	0.222	0.072
Belfast	0.330	0.347	0.187	0.499	0.313	0.109
Standerton	0.323	0.339	0.232	0.437	0.205	0.078
Nkomazi	0.307	0.315	0.160	0.453	0.294	0.097
Moretele	0.299	0.273	0.148	0.415	0.266	0.083
Volksrust	0.294	0.293	0.224	0.367	0.143	0.048
Wakkerstroom	0.282	0.291	0.191	0.462	0.271	0.091
Lydenburg	0.281	0.306	0.213	0.473	0.260	0.062
Piet Retief	0.272	0.334	0.213	0.523	0.310	0.095
Pelgrimsrust	0.258	0.270	0.183	0.392	0.209	0.069
Nsikazi	0.252	0.231	0.147	0.298	0.152	0.043
Groblersdal	0.251	0.226	0.143	0.308	0.165	0.050
Barberton	0.235	0.269	0.157	0.394	0.236	0.081
Witbank	0.187	0.192	0.127	0.278	0.151	0.051
Middelburg	0.181	0.167	0.109	0.241	0.132	0.039
Hoveldrif	0.171	0.205	0.123	0.297	0.174	0.041
Witrivier	0.159	0.168	0.107	0.253	0.146	0.060
Kwamhlanga	0.153	0.163	0.068	0.352	0.284	0.153

Continued on next page...

Table A.9 (continued from previous page)

Magisterial District	HC (maximal)	Mean	Min	Max	Range	IQR
Delmas	0.153	0.189	0.105	0.294	0.190	0.057
Nelspruit	0.146	0.151	0.087	0.224	0.138	0.040
Kriel	0.117	0.085	0.007	0.228	0.221	0.087
Mbibana	0.111	0.101	0.025	0.247	0.223	0.065
Mdutjana	0.089	0.114	0.055	0.191	0.136	0.041
Mkobola	0.070	0.066	0.029	0.159	0.131	0.038
Moutse	0.039	0.083	0.025	0.252	0.227	0.065

Table A.9: Estimates Over 50 Random Specifications,
Mpumalanga

Magisterial District	HC (maximal)	Mean	Min	Max	Range	IQR
Letaba	0.641	0.596	0.450	0.741	0.292	0.073
Messina	0.567	0.470	0.363	0.600	0.237	0.085
Mhala	0.541	0.468	0.343	0.559	0.215	0.058
Bolobedu	0.455	0.385	0.260	0.514	0.254	0.113
Sekhukhuleni	0.442	0.402	0.265	0.495	0.230	0.058
Mapulaneng	0.433	0.415	0.262	0.524	0.262	0.086
Bochum	0.409	0.376	0.233	0.510	0.277	0.101
Mokerong	0.397	0.379	0.277	0.471	0.194	0.062
Seshego	0.359	0.335	0.228	0.409	0.181	0.064
Thabamooop	0.334	0.300	0.179	0.415	0.237	0.060
Nebo	0.330	0.341	0.238	0.439	0.202	0.039
Sekgose	0.321	0.351	0.215	0.541	0.327	0.084
Soutpansberg	0.316	0.315	0.205	0.426	0.221	0.072
Mutali	0.311	0.347	0.193	0.494	0.301	0.114
Dzanani	0.280	0.311	0.148	0.478	0.329	0.109
Phalaborwa	0.279	0.258	0.159	0.337	0.178	0.062
Warmbad	0.273	0.323	0.212	0.472	0.260	0.099
Ritavi	0.261	0.263	0.181	0.333	0.152	0.061
Thabazimbi	0.232	0.334	0.244	0.485	0.241	0.084
Vuwani	0.232	0.251	0.164	0.337	0.173	0.040
Malamulela	0.230	0.261	0.161	0.435	0.275	0.102
Hlanganani	0.221	0.231	0.134	0.359	0.225	0.080
Potgietersrus	0.214	0.308	0.201	0.482	0.281	0.104
Namakgale	0.205	0.213	0.127	0.338	0.211	0.062

Continued on next page...

Table A.10 (continued from previous page)

Magisterial District	HC (maximal)	Mean	Min	Max	Range	IQR
Waterberg	0.190	0.262	0.109	0.520	0.410	0.113
Lulekani	0.189	0.234	0.131	0.333	0.202	0.073
Thohoyandou	0.180	0.198	0.115	0.284	0.169	0.055
Naphuno	0.175	0.207	0.126	0.338	0.212	0.063
Pietersburg	0.163	0.172	0.106	0.251	0.145	0.043
Giyani	0.136	0.172	0.070	0.347	0.277	0.083
Ellisras	0.089	0.152	0.079	0.287	0.208	0.068

Table A.10: Estimates Over 50 Random Specifications,
Limpopo

A.3 Sensitivity: Rankings

Magisterial District	Rank (maximal)	Mean	Min	Max	Range	IQR
Murraysburg	1	2.4	1	9	8	2
Uniondale	2	2.4	1	7	6	2
Calitzdorp	3	3.5	1	7	6	2
Prince Albert	4	4.8	1	17	16	2
Robertson	5	4.9	1	12	11	3
Swellendam	6	7.5	1	14	13	3
Van Rhynsdorp	7	10.2	4	19	15	6
Laingsburg	8	8.9	1	22	21	7
Worcester	9	12.9	5	22	17	7
Vredendal	10	9.6	3	21	18	4
Knysna	11	17.2	10	26	16	5
Moorreesburg	12	13.6	5	30	25	7
Hopefield	13	23.0	12	34	22	6
Montagu	14	15.8	7	25	18	7
Clanwilliam	15	12.1	4	20	16	4
Tulbagh	16	17.8	8	32	24	8
Riversdal	17	15.8	8	24	16	7
Caledon	18	15.8	8	23	15	7
Beaufort West	19	17.2	7	30	23	9
Ceres	20	16.4	6	28	22	7
Heidelberg	21	13.9	4	28	24	8
Oudtshoorn	22	24.6	14	32	18	6

Continued on next page...

Table A.11 (continued from previous page)

Magisterial District	Rank (maximal)	Mean	Min	Max	Range	IQR
Ladismith	23	17.6	5	29	24	12
George	24	24.5	18	30	12	4
Mitchellsplain	25	23.1	14	31	17	7
Piketberg	26	22.4	13	31	18	6
Mossel bay	27	26.6	18	33	15	5
Stellenbosch	28	30.0	22	36	14	3
Bredasdorp	29	27.6	18	34	16	5
Hermanus	30	27.4	20	34	14	5
Wellington	31	32.2	23	37	14	3
Paarl	32	29.5	22	34	12	3
Malmesbury	33	31.8	25	36	11	3
Strand	34	33.6	28	37	9	2
Goodwood	35	36.7	32	39	7	2
Kuilsrivier	36	36.3	32	40	8	1
Vredenburg	37	35.0	30	39	9	2
Simonstown	38	37.4	35	40	5	1
Somerset West	39	38.5	35	41	6	1
Bellville	40	41.1	40	42	2	1
Cape	41	40.1	39	41	2	0
Wynberg	42	41.6	35	42	7	1

Table A.11: Within-Province Rankings Over 50 Random Specifications, W Cape

Magisterial District	Rank (maximal)	Mean	Min	Max	Range	IQR
Mqanduli	1	8.6	1	28	27	9
Elliotdale	2	3.0	1	16	15	2
Tabankulu	3	7.7	2	20	18	6
Flagstaff	4	6.4	1	19	18	5
Kentani	5	8.2	1	22	21	8
Umzimkulu	6	14.1	1	29	28	10
Cala	7	15.1	1	30	29	10
Lusikisiki	8	11.6	1	24	23	9
Ngqueleni	9	8.6	1	24	23	10
Engcobo	10	12.5	2	26	24	11
Qumbu	11	14.3	3	28	25	11

Continued on next page...

Table A.12 (continued from previous page)

Magisterial District	Rank (maximal)	Mean	Min	Max	Range	IQR
Middeldrift	12	17.5	2	29	27	12
Tsomo	13	11.0	1	32	31	11
Mt Fletcher	14	10.7	1	26	25	9
Cofimvaba	15	17.2	3	28	25	10
Bizana	16	15.9	2	35	33	11
Libode	17	13.8	3	24	21	10
Mt Ayliff	18	15.7	1	26	25	9
Mt Frere	19	20.1	6	30	24	8
Maluti	20	17.7	3	32	29	9
Idutywa	21	17.6	4	31	27	10
Nqamakwe	22	20.7	4	37	33	10
Willowvale	23	15.2	2	26	24	9
Tsolo	24	18.9	5	29	24	9
Pearston	25	28.0	11	58	47	6
Mpofu	26	25.9	1	54	53	18
Port St Johns	27	17.6	5	29	24	9
Lady Frere	28	29.7	18	48	30	6
Steytlerville	29	37.6	27	53	26	10
Umtata	30	34.6	29	42	13	5
Hofmeyer	31	32.4	18	51	33	7
Ntabathemba	32	41.1	12	57	45	16
Sterkspruit	33	29.4	9	42	33	5
Maclear	34	32.9	8	56	48	8
Bedford	35	38.0	27	57	30	11
Hankey	36	49.5	30	66	36	12
Wodehouse	37	40.0	27	56	29	13
Victoria East	38	43.8	28	57	29	12
Sterkstroom	39	40.6	20	59	39	13
Peddie	40	37.9	27	58	31	8
Kciskammahoek	41	42.0	22	65	43	20
Barkley-East	42	41.3	30	58	28	7
Steynsburg	43	43.4	26	61	35	17
Komga	44	31.0	13	56	43	10
Butterworth	45	47.1	32	62	30	8
Adelaide	46	49.6	34	64	30	10
Hewu	47	48.8	31	63	32	11

Continued on next page...

Table A.12 (continued from previous page)

Magisterial District	Rank (maximal)	Mean	Min	Max	Range	IQR
Jansenville	48	48.7	34	67	33	11
Lady Grey	49	41.0	32	54	22	9
Stutterheim	50	49.4	35	68	33	15
Willowmore	51	53.2	37	69	32	12
Zwelitsha	52	45.1	30	61	31	12
Alexandria	53	47.4	34	60	26	12
Somerset East	54	52.4	40	64	24	9
Bathurst	55	53.8	35	73	38	11
Kirkwood	56	53.0	38	65	27	10
Molteno	57	54.0	36	70	34	10
Tarka	58	50.1	37	70	33	9
Fort Beaufort	59	57.2	36	72	36	8
Cradock	60	60.2	45	70	25	5
Albert	61	59.8	46	70	24	6
Cathcart	62	62.5	41	71	30	5
Indwe	63	59.2	46	69	23	7
Elliot	64	62.3	49	70	21	5
Venterstad	65	64.4	52	72	20	6
Aberdeen	66	64.8	44	73	29	6
Aliwal North	67	66.8	54	73	19	4
East-London	68	68.0	62	72	10	3
Mdantsane	69	68.8	62	75	13	3
Joubertina	70	61.2	46	72	26	8
Humansdorp	71	71.1	63	76	13	3
Queenstown	72	74.0	71	77	6	2
Albany	73	71.4	66	76	10	3
Middelburg	74	73.7	68	76	8	2
Graaff-Reinet	75	73.8	70	77	7	3
Uitenhage	76	74.7	72	76	4	2
Port Elizabeth	77	76.9	75	78	3	0
King William's Town	78	78.0	77	78	1	0

Table A.12: Within-Province Rankings Over 50 Random Specifications, E Cape

Magisterial District	Rank (maximal)	Mean	Min	Max	Range	IQR
Noupoort	1	1.4	1	4	3	1
Warrenton	2	2.9	1	9	8	2
Hanover	3	5.2	1	23	22	3
Philipstown	4	6.9	2	20	18	4
Fraserburg	5	5.6	2	15	13	4
Richmond	6	4.2	2	7	5	2
Barkley-West	7	10.9	4	24	20	6
Prieska	8	10.7	4	20	16	5
Williston	9	12.8	6	21	15	5
Hartswater	10	8.2	3	14	11	3
Britstown	11	13.3	5	23	18	6
Calvinia	12	15.0	6	24	18	6
Hopetown	13	12.3	2	23	21	7
Herbert	14	13.1	5	25	20	4
Sutherland	15	12.8	2	24	22	5
Kenhardt	16	18.2	7	25	18	9
Kuruman	17	15.4	7	22	15	7
Victoria-West	18	18.9	4	26	22	8
Carnarvon	19	17.3	3	24	21	7
Postmasburg	20	17.2	9	24	15	6
Hay	21	17.5	5	24	19	5
De Aar	22	20.4	14	25	11	5
Colesberg	23	18.6	10	24	14	5
Gordonia	24	22.4	16	26	10	3
Kimberley	25	23.9	16	25	9	2
Namakwaland	26	26.0	25	26	1	0

Table A.13: Within-Province Rankings Over 50 Random Specifications, N Cape

Magisterial District	Rank (maximal)	Mean	Min	Max	Range	IQR
Koppies	1	1.9	1	6	5	1
Smithfield	2	7.5	2	25	23	7
Hoopstad	3	11.1	1	30	29	8
Vredefort	4	7.3	1	29	28	5
Boshof	5	7.1	2	18	16	5
Jacobsdal	6	11.8	1	29	28	12

Continued on next page...

Table A.14 (continued from previous page)

Magisterial District	Rank (maximal)	Mean	Min	Max	Range	IQR
Viljoenskroon	7	14.0	3	35	32	10
Ficksburg	8	6.7	1	33	32	10
Trompsburg	9	20.4	5	36	31	14
Dewetsdorp	10	18.8	2	38	36	13
Marquard	11	11.7	2	33	31	9
Clocolan	12	18.3	5	33	28	12
Wesselsbron	13	13.7	2	32	30	8
Reitz	14	13.0	3	30	27	9
Senekal	15	15.6	3	31	28	10
Heilbron	16	14.4	4	34	30	9
Frankfort	17	23.9	4	35	31	11
Bothaville	18	16.3	4	36	32	15
Fouriesburg	19	13.0	1	28	27	12
Bultfontein	20	22.1	9	35	26	9
Theunissen	21	24.6	9	36	27	9
Ventersburg	22	26.5	8	43	35	11
Brandfort	23	22.1	6	37	31	12
Petrusburg	24	23.2	6	38	32	13
Vrede	25	12.9	1	35	34	9
Excelsior	26	28.8	8	42	34	10
Harrismith	27	28.7	8	40	32	11
Philippolis	28	28.1	12	40	28	12
Wepener	29	28.2	10	41	31	9
Fauresmith	30	23.8	5	44	39	16
Lindley	31	27.1	13	42	29	9
Zastron	32	27.7	6	43	37	14
Witsieshoek	33	18.3	1	38	37	21
Winburg	34	35.2	19	44	25	9
Jagersfontein	35	31.0	12	46	34	13
Ladybrand	36	35.2	13	43	30	8
Edenburg	37	36.6	19	46	27	6
Rouxville	38	27.7	7	40	33	17
Reddersburg	39	38.9	27	48	21	8
Bethlehem	40	40.5	26	46	20	6
Koffiefontein	41	41.6	20	48	28	5
Parys	42	40.8	33	48	15	6

Continued on next page...

Table A.14 (continued from previous page)

Magisterial District	Rank (maximal)	Mean	Min	Max	Range	IQR
Botshabelo	43	43.2	29	50	21	6
Thaba 'Nchu	44	36.3	6	48	42	8
Hennenman	45	45.2	40	49	9	3
Kroonstad	46	46.6	36	49	13	3
Odendaalsrus	47	45.0	38	48	10	4
Bethulie	48	43.7	39	50	11	4
Virginia	49	49.7	46	52	6	2
Welkom	50	50.6	48	52	4	1
Sasolburg	51	49.5	46	52	6	1
Bloemfontein	52	51.8	50	52	2	0

Table A.14: Within-Province Rankings Over 50 Random Specifications, Free State

Magisterial District	Rank (maximal)	Mean	Min	Max	Range	IQR
Weenen	1	1.5	1	5	4	1
Ngotshe	2	2.1	1	11	10	0
Underberg	3	3.4	1	6	5	1
Utrecht	4	5.5	2	14	12	4
Paulpietersburg	5	9.5	2	30	28	10
Kranskop	6	8.2	1	25	24	6
Mount Currie	7	7.1	2	16	14	3
New Hanover	8	9.2	4	21	17	3
Msinga	9	8.7	2	18	16	3
Polela	10	16.6	8	31	23	9
Mthonjaneni	11	15.7	6	25	19	8
Ixopo	12	10.3	4	22	18	4
Alfred	13	14.6	7	28	21	8
Nkandla	14	14.8	6	38	32	6
Mooi river	15	14.7	3	35	32	9
Umvoti	16	13.1	7	27	20	3
Richmond	17	13.9	4	32	28	10
Babanango	18	17.4	8	25	17	5
Umzinto	19	18.9	11	29	18	6
Lower Tugela	20	20.5	12	34	22	10
Simdlangentsha	21	23.2	8	35	27	5

Continued on next page...

Table A.15 (continued from previous page)

Magisterial District	Rank (maximal)	Mean	Min	Max	Range	IQR
Nongoma	22	23.6	11	33	22	8
Vryheid	23	22.4	13	35	22	9
Nqutu	24	23.5	15	34	19	6
Mhlabathini	25	25.4	11	38	27	8
Mapumulo	26	25.2	10	34	24	7
Bergville	27	30.2	17	41	24	7
Estcourt	28	32.7	18	42	24	5
Ubombo	29	24.4	11	37	26	10
Eshowe	30	31.8	23	39	16	4
Kliprivier	31	27.5	18	37	19	8
Impendle	32	30.3	16	41	25	8
Dundee	33	31.6	24	39	15	8
Dannhauser	34	34.9	19	46	27	5
Hlabisa	35	30.7	20	42	22	8
Ndwedwe	36	37.4	27	45	18	4
Ingwavuma	37	30.1	12	41	29	8
Umbumbulu	38	38.6	27	46	19	5
Port Shepstone	39	39.9	33	46	13	6
Inanda	40	41.6	34	48	14	4
Mtunzini	41	40.9	35	47	12	4
Glencoe	42	39.1	30	48	18	6
Lions River	43	41.9	27	48	21	5
Lower Umfolozi	44	44.2	38	48	10	4
Newcastle	45	44.9	36	49	13	4
Umlazi	46	45.9	36	49	13	3
Camperdown	47	42.1	28	48	20	5
Pietermaritzburg	48	46.6	42	50	8	3
Pinetown	49	48.6	46	49	3	1
Durban	50	50.6	50	51	1	1
Chatswoth	51	50.3	49	51	2	1

Table A.15: Within-Province Rankings Over 50 Random Specifications, KwaZulu-Natal

Magisterial District	Rank (maximal)	Mean	Min	Max	Range	IQR
Bronkhorstspuit	1	1.0	1	1	0	0

Continued on next page...

Table A.16 (continued from previous page)

Magisterial District	Rank (maximal)	Mean	Min	Max	Range	IQR
Nigel	2	4.2	2	7	5	2
Cullinan	3	3.1	2	7	5	2
Oberholzer	4	5.0	2	15	13	3
Westonaria	5	7.0	3	16	13	3
Heidelberg	6	4.2	2	15	13	2
Vanderbijlpark	7	10.2	5	16	11	5
Randfontein	8	6.7	4	11	7	1
Vereeniging	9	10.4	5	15	10	4
Brakpan	10	10.4	4	17	13	3
Benoni	11	10.5	4	15	11	4
Randburg	12	13.5	8	18	10	3
Soshanguve	13	9.7	4	18	14	4
Alberton	14	14.0	6	18	12	3
Kempton Park	15	13.6	8	19	11	3
Krugersdorp	16	13.7	9	17	8	3
Boksburg	17	18.2	16	22	6	2
Springs	18	20.0	13	22	9	2
Wonderboom	19	17.7	13	21	8	2
Roodepoort	20	19.4	16	22	6	1
Johannesburg	21	21.7	18	24	6	1
Soweto	22	19.9	7	24	17	5
Germiston	23	22.2	18	24	6	1
Pretoria	24	23.8	22	24	2	0

Table A.16: Within-Province Rankings Over 50 Random Specifications, Gauteng

Magisterial District	Rank (maximal)	Mean	Min	Max	Range	IQR
Carolina	1	1.3	1	6	5	0
Eerstehoek	2	3.0	1	7	6	1
Bethal	3	3.4	1	7	6	2
Waterval Boven	4	4.0	1	7	6	0
Amersfoort	5	6.6	1	17	16	4
Ermelo	6	7.0	3	19	16	2
Balfour	7	11.8	7	20	13	5
Belfast	8	9.3	2	20	18	6

Continued on next page...

Table A.17 (continued from previous page)

Magisterial District	Rank (maximal)	Mean	Min	Max	Range	IQR
Standerton	9	9.9	5	19	14	5
Nkomazi	10	11.9	3	27	24	9
Moretele	11	14.9	6	25	19	8
Volksrust	12	12.6	7	20	13	3
Wakkerstroom	13	12.8	5	19	14	6
Lydenburg	14	11.8	2	20	18	5
Piet Retief	15	10.5	4	20	16	5
Pelgrimsrust	16	14.6	6	22	16	7
Nsikazi	17	17.9	11	26	15	4
Groblersdal	18	18.2	10	24	14	6
Barberton	19	14.6	7	22	15	7
Witbank	20	21.5	16	27	11	3
Middelburg	21	23.7	19	28	9	4
Hoveldrif	22	20.3	11	25	14	4
Witrivier	23	23.4	16	29	13	4
Kwamhlanga	24	22.9	10	29	19	10
Delmas	25	21.8	14	29	15	3
Nelspruit	26	24.8	20	29	9	2
Kriel	27	28.1	19	31	12	5
Mbibana	28	27.7	18	31	13	3
Mdutjana	29	27.3	22	30	8	3
Mkobola	30	29.9	28	31	3	0
Moutse	31	28.8	18	31	13	3

Table A.17: Within-Province Rankings Over 50 Random Specifications, Mpumalanga

Magisterial District	Rank (maximal)	Mean	Min	Max	Range	IQR
Lctaba	1	1.2	1	6	5	0
Messina	2	4.0	2	9	7	3
Mhala	3	3.8	1	13	12	2
Bolobedu	4	8.9	2	19	17	9
Sekhukhuleni	5	7.5	2	16	14	4
Mapulaneng	6	6.6	2	17	15	4
Bochum	7	9.9	2	22	20	8
Mokerong	8	9.2	4	20	16	4

Continued on next page...

Table A.18 (continued from previous page)

Magisterial District	Rank (maximal)	Mean	Min	Max	Range	IQR
Seshego	9	12.8	7	24	17	6
Thabamooopo	10	16.5	6	30	24	8
Nebo	11	12.2	2	24	22	4
Sekgoscse	12	11.8	1	26	25	8
Soutpansberg	13	14.1	5	26	21	6
Mutali	14	12.1	2	25	23	11
Dzanani	15	15.4	3	31	28	11
Phalaborwa	16	20.4	12	31	19	7
Warmbad	17	14.0	2	26	24	9
Ritavi	18	20.0	14	31	17	7
Thabazimbi	19	13.1	2	22	20	8
Vuwani	20	21.1	10	29	19	4
Malamulela	21	19.7	3	28	25	6
Hlanganani	22	22.8	9	31	22	7
Potgietersrus	23	15.2	2	25	23	10
Namakgale	24	24.3	10	30	20	6
Waterberg	25	20.4	2	30	28	11
Lulekani	26	22.8	11	30	19	7
Thohoyandou	27	26.3	19	31	12	5
Naphuno	28	25.2	15	31	16	6
Pietersburg	29	28.1	25	31	6	2
Giyani	30	27.6	9	31	22	5
Ellisras	31	28.9	12	31	19	2

Table A.18: Within-Province Rankings Over 50 Random Specifications, Limpopo

Appendix B

Descriptive Statistics

Below, I report some basic descriptive statistics for each of the nine provinces, broken down by the data source (IES/OHS or Census Data). Descriptive statistics by province for the dependent variable, the logarithm of total monthly household consumption, appear in section B.1. Next, I report the statistics for the household-level controls in section B.2, while the descriptive statistics for the area-level controls are tabulated in section B.3. All statistics are individual-level estimates, i.e. having been weighted by household size and sampling weights (in the case of the IES) or post-stratification weights (in the case of the census).

B.1 Consumption Data

Province	Mean	(Std. Dev.)	Min.	Max.	N
W Cape	7.7952	(0.9362)	4.3737	11.9343	3 860 967
E Cape	6.9501	(0.9846)	3.8833	12.0237	6 059 647
N Cape	7.1605	(1.0041)	4.0999	11.1425	811 126
Free State	6.9565	(1.0359)	3.8677	11.1075	2 448 094
KwaZulu-Natal	7.5049	(0.9295)	4.6883	12.6373	7 786 987
North West	7.1572	(1.0335)	4.6883	12.932	2 014 530
Gauteng	8.1135	(0.9600)	4.8122	12.0034	6 562 701
Mpumalanga	7.3076	(0.9082)	4.7362	11.0076	2 645 663
Limpopo	7.1985	(1.0646)	4.2195	11.9525	4 773 999

Table B.1: Summary Statistics - Log Monthly Total Expenditure, by Province (IES Data)

B.2 Household-Level Covariates

Table B.2: Summary Statistics - HH Controls, W Cape (Census Data)

Variable	Mean	(Std. Dev.)	Min.	Max.
(mean) logHHsize	1.499	(0.549)	0	3.434
africanHH	0.2074	(0.4054)	0	1
whiteHH	0.1924	(0.3942)	0	1
formalDwelling	0.8024	(0.3982)	0	1
roomsPerPerson	1.0842	(0.8985)	0	65
sanitationFacilities	0.8689	(0.3375)	0	1
electricLighting	0.8711	(0.3351)	0	1
refuseCollection	0.8579	(0.3491)	0	1
hasTelephone	0.5318	(0.499)	0	1
numPrimaryEd	3.5084	(2.021)	0	24
numProfessional	0.2258	(0.5219)	0	5
numSkilled	0.3917	(0.6726)	0	14
femaleHeadedHH	0.2562	(0.4365)	0	1
(mean) farm	0.1015	(0.302)	0	1
(mean) urban	0.8912	(0.3114)	0	1
N	3803234			

Table B.3: Summary Statistics - HH Controls, W Cape (IES Data)

Variable	Mean	(Std. Dev.)	Min.	Max.
(mean) logHHsize	1.5108	(0.5273)	0	3.2958
africanHH	0.1918	(0.3937)	0	1
whiteHH	0.2123	(0.4089)	0	1
formalDwelling	0.8656	(0.3411)	0	1
roomsPerPerson	1.1112	(0.9380)	0.1111	11
sanitationFacilities	0.1716	(0.377)	0	1
electricLighting	0.8921	(0.3103)	0	1
refuseCollection	0.8485	(0.3586)	0	1
hasTelephone	0.4719	(0.4992)	0	1
numPrimaryEd	2.6687	(1.7495)	0	11
numProfessional	0.1487	(0.4172)	0	3
numSkilled	0.4256	(0.7196)	0	5
femaleHeadedHH	0.2263	(0.4184)	0	1
(mean) farm	0.1431	(0.3501)	0	1
(mean) urban	0.8390	(0.3675)	0	1
N	3860967			

Table B.4: Summary Statistics: HH Controls, E Cape (Census Data)

Variable	Mean	(Std. Dev.)	Min.	Max.
(mean) logHHsize	1.6758	(0.5634)	0	3.3673
africanHH	0.8673	(0.3393)	0	1
whiteHH	0.0486	(0.2149)	0	1
formalDwelling	0.4184	(0.4933)	0	1
roomsPerPerson	0.7678	(0.7314)	0	23
sanitationFacilities	0.2692	(0.4435)	0	1
electricLighting	0.2919	(0.4547)	0	1
refuseCollection	0.3195	(0.4663)	0	1
hasTelephone	0.1363	(0.3431)	0	1
numPrimaryEd	3.2585	(2.2136)	0	24
numProfessional	0.1068	(0.3792)	0	8

Continued on next page...

... table B.4 continued

Variable	Mean	(Std. Dev.)	Min.	Max.
numSkilled	0.1436	(0.4337)	0	7
femaleHeadedHH	0.5027	(0.5)	0	1
(mean) farm	0.0351	(0.1841)	0	1
(mean) urban	0.3593	(0.4798)	0	1
(mean) tribal	0.5892	(0.4920)	0	1
N	616770			

University of Cape Town

Table B.5: Summary Statistics: HH Controls, E Cape (IES Data)

Variable	Mean	(Std. Dev.)	Min.	Max.
(mean) logHHsize	1.7037	(0.5049)	0	3.0445
africanHH	0.8649	(0.3418)	0	1
whiteHH	0.0485	(0.2148)	0	1
formalDwelling	0.5259	(0.4993)	0	1
roomsPerPerson	0.8581	(0.7869)	0.0833	13
sanitationFacilities	0.1748	(0.3798)	0	1
electricLighting	0.3184	(0.4658)	0	1
refuseCollection	0.3374	(0.4728)	0	1
hasTelephone	0.127	(0.3329)	0	1
numPrimaryEd	2.523	(1.8507)	0	10
numProfessional	0.1218	(0.391)	0	4
numSkilled	0.1293	(0.3762)	0	4
femaleHeadedHH	0.4487	(0.4974)	0	1
(mean) farm	0.0882	(0.2836)	0	1
(mean) urban	0.353	(0.4779)	0	1
(mean) tribal	0.4616	(0.4985)	0	1
N	6059647			

Table B.6: Summary Statistics - HH Controls, N Cape 1

Variable	Mean	(Std. Dev.)	Min.	Max.
(mean) logHHsize	1.5935	(0.5698)	0	3.7377
africanHH	0.3142	(0.4642)	0	1
whiteHH	0.1202	(0.3252)	0	1
formalDwelling	0.7823	(0.4127)	0	1
roomsPerPerson	0.9054	(0.8580)	0	14
sanitationFacilities	0.5890	(0.4920)	0	1
electricLighting	0.7297	(0.4441)	0	1
refuseCollection	0.7315	(0.4432)	0	1
hasTelephone	0.2833	(0.4506)	0	1
numPrimaryEd	3.162	(2.1628)	0	13
numProfessional	0.1239	(0.3881)	0	4

Continued on next page...

... table B.6 continued

Variable	Mean	(Std. Dev.)	Min.	Max.
numSkilled	0.3087	(0.6457)	0	10
femaleHeadedHH	0.2978	(0.4573)	0	1
(mean) farm	0.2389	(0.4264)	0	1
(mean) urban	0.7049	(0.4561)	0	1
N	802263			

Table B.7: Summary Statistics - HH Controls, N Cape (IES Data)

Variable	Mean	(Std. Dev.)	Min.	Max.
(mean) logHHsize	1.5028	(0.5666)	0	2.6391
africanHH	0.3094	(0.4622)	0	1
whiteHH	0.1344	(0.3411)	0	1
formalDwelling	0.8426	(0.3642)	0	1
roomsPerPerson	0.9641	(0.8947)	0.1429	11
sanitationFacilities	0.2344	(0.4236)	0	1
electricLighting	0.7688	(0.4216)	0	1
refuseCollection	0.7361	(0.4407)	0	1
hasTelephone	0.2544	(0.4355)	0	1
numPrimaryEd	1.9612	(1.6547)	0	9
numProfessional	0.0718	(0.2877)	0	2
numSkilled	0.2364	(0.531)	0	4
femaleHeadedHH	0.2572	(0.4371)	0	1
(mean) farm	0.2551	(0.4359)	0	1
(mean) urban	0.7000	(0.4583)	0	1
N	811126			

Table B.8: Summary Statistics - HH Controls, Free State (Census Data)

Variable	Mean	(Std. Dev.)	Min.	Max.
(mean) logHHsize	1.5224	(0.5524)	0	3.1355

Continued on next page...

... table B.8 continued

Variable	Mean	(Std. Dev.)	Min.	Max.
africanHH	0.8410	(0.3656)	0	1
whiteHH	0.1129	(0.3164)	0	1
formalDwelling	0.5826	(0.4931)	0	1
roomsPerPerson	0.9340	(0.8438)	0	20.75
sanitationFacilities	0.4087	(0.4916)	0	1
electricLighting	0.5671	(0.4955)	0	1
refuseCollection	0.6341	(0.4817)	0	1
hasTelephone	0.2057	(0.4042)	0	1
numPrimaryEd	3.0211	(1.8649)	0	15
numProfessional	0.1153	(0.3781)	0	7
numSkilled	0.3028	(0.5476)	0	10
femaleHeadedHH	0.3364	(0.4725)	0	1
(mean) farm	0.166	(0.3721)	0	1
(mean) urban	0.7094	(0.4541)	0	1
N	2473262			

Table B.9: Summary Statistics - HH Controls, Free State (IES Data)

Variable	Mean	(Std. Dev.)	Min.	Max.
(mean) logHHsize	1.4835	(0.5276)	0	2.7081
africanHH	0.8480	(0.359)	0	1
whiteHH	0.1127	(0.3162)	0	1
formalDwelling	0.727	(0.4455)	0	1
roomsPerPerson	1.0643	(0.904)	0.1111	11
sanitationFacilities	0.2027	(0.402)	0	1
electricLighting	0.6830	(0.4653)	0	1
refuseCollection	0.6088	(0.488)	0	1
hasTelephone	0.2075	(0.4055)	0	1
numPrimaryEd	2.1238	(1.7011)	0	10
numProfessional	0.1244	(0.4067)	0	3
numSkilled	0.2042	(0.4486)	0	3
femaleHeadedHH	0.2627	(0.4401)	0	1
(mean) farm	0.331	(0.4706)	0	1

Continued on next page...

... table B.9 continued

Variable	Mean	(Std. Dev.)	Min.	Max.
(mean) urban	0.6245	(0.4843)	0	1
N	2448094			

Table B.10: Summary Statistics - HH Controls, KZN (Census Data)

Variable	Mean	(Std. Dev.)	Min.	Max.
(mean) logHHsize	1.7444	(0.5995)	0	3.912
africanHH	0.8147	(0.3885)	0	1
whiteHH	0.061	(0.2393)	0	1
formalDwelling	0.4647	(0.4988)	0	1
roomsPerPerson	0.8505	(0.7242)	0	43
sanitationFacilities	0.3374	(0.4728)	0	1
electricLighting	0.4852	(0.4998)	0	1
refuseCollection	0.3538	(0.4782)	0	1
hasTelephone	0.2218	(0.4155)	0	1
numPrimaryEd	3.6594	(2.5185)	0	28
numProfessional	0.1371	(0.4292)	0	15
numSkilled	0.2383	(0.603)	0	32
femaleHeadedHH	0.4065	(0.4912)	0	1
(mean) farm	0.0544	(0.2268)	0	1
(mean) urban	0.4266	(0.4946)	0	1
(mean) tribal	0.4832	(0.4997)	0	1
N	8097994			

Table B.11: Summary Statistics - HH Controls, KZN (IES Data)

Variable	Mean	(Std. Dev.)	Min.	Max.
(mean) logHHsize	1.7758	(0.5128)	0	3.434
africanHH	0.8196	(0.3845)	0	1

Continued on next page...

... table B.11 continued

Variable	Mean	(Std. Dev.)	Min.	Max.
whiteHH	0.0606	(0.2386)	0	1
formalDwelling	0.5396	(0.4984)	0	1
roomsPerPerson	0.9253	(0.7196)	0.0769	14
sanitationFacilities	0.1766	(0.3813)	0	1
electricLighting	0.5185	(0.4997)	0	1
refuseCollection	0.3996	(0.4898)	0	1
hasTelephone	0.2201	(0.4143)	0	1
numPrimaryEd	2.9497	(2.0222)	0	12
numProfessional	0.1609	(0.4638)	0	4
numSkilled	0.3138	(0.5648)	0	4
femaleHeadcdHH	0.3444	(0.4752)	0	1
(mean) farm	0.1076	(0.3099)	0	1
(mean) urban	0.4184	(0.4933)	0	1
(mean) tribal	0.4384	(0.4962)	0	1
N	7786987			

Table B.12: Summary Statistics - HH Controls, North West (Census Data)

Variable	Mean	(Std. Dev.)	Min.	Max.
(mean) logHHsize	1.6526	(0.5883)	0	3.5264
africanHH	0.9128	(0.2821)	0	1
whiteHH	0.0616	(0.2405)	0	1
formalDwelling	0.6769	(0.4677)	0	1
roomsPerPerson	0.8963	(0.7695)	0	28
sanitationFacilities	0.2733	(0.4456)	0	1
electricLighting	0.4237	(0.4941)	0	1
refuseCollection	0.3252	(0.4684)	0	1
hasTelephone	0.1464	(0.3535)	0	1
numPrimaryEd	3.3128	(2.1729)	0	15
numProfessional	0.1124	(0.3682)	0	5
numSkilled	0.3031	(0.5689)	0	8
fcmaleHeadcdHH	0.3962	(0.4891)	0	1
(mean) farm	0.0913	(0.2881)	0	1

Continued on next page...

... table B.12 continued

Variable	Mean	(Std. Dev.)	Min.	Max.
(mean) urban	0.3514	(0.4774)	0	1
(mean) tribal	0.4639	(0.4987)	0	1
N	3216039			

Table B.13: Summary Statistics - HH Controls, North West (IES Data)

Variable	Mean	(Std. Dev.)	Min.	Max.
(mean) logHHsize	1.5784	(0.5848)	0	3.1355
africanHH	0.8554	(0.3517)	0	1
whiteHH	0.1058	(0.3075)	0	1
formalDwelling	0.7972	(0.4021)	0	1
roomsPerPerson	0.9756	(0.7848)	0.125	12
sanitationFacilities	0.1753	(0.3802)	0	1
electricLighting	0.5349	(0.4988)	0	1
refuseCollection	0.4276	(0.4947)	0	1
hasTelephone	0.1663	(0.3723)	0	1
numPrimaryEd	2.3662	(1.777)	0	9
numProfessional	0.1055	(0.3615)	0	3
numSkilled	0.2805	(0.503)	0	3
femaleHeadedHH	0.2462	(0.4308)	0	1
(mean) farm	0.2453	(0.4303)	0	1
(mean) urban	0.4196	(0.4935)	0	1
(mean) tribal	0.1632	(0.3695)	0	1
N	2014530			

Table B.14: Summary Statistics - HH Controls, Gauteng (Census)

Variable	Mean	(Std. Dev.)	Min.	Max.
(mean) logHHsize	1.4239	(0.5929)	0	3.912

Continued on next page...

... table B.14 continued

Variable	Mean	(Std. Dev.)	Min.	Max.
africanHH	0.6858	(0.4642)	0	1
whiteHH	0.2133	(0.4096)	0	1
formalDwelling	0.6951	(0.4603)	0	1
roomsPerPerson	1.0976	(0.9295)	0	30
sanitationFacilities	0.8360	(0.3703)	0	1
electricLighting	0.8191	(0.3849)	0	1
refuseCollection	0.8577	(0.3494)	0	1
hasTelephone	0.4637	(0.4987)	0	1
numPrimaryEd	3.3858	(2.1076)	0	29
numProfessional	0.2495	(0.5508)	0	9
numSkilled	0.354	(0.6017)	0	9
femaleHeadedHH	0.2811	(0.4495)	0	1
(mean) farm	0.0258	(0.1584)	0	1
(mean) urban	0.9713	(0.167)	0	1
N	6890762			

Table B.15: Summary Statistics - HH Controls, Gauteng (IES Data)

Variable	Mean	(Std. Dev.)	Min.	Max.
(mean) logHHsize	1.4621	(0.5475)	0	3.091
africanHH	0.6854	(0.4644)	0	1
whiteHH	0.2349	(0.424)	0	1
formalDwelling	0.8343	(0.3718)	0	1
roomsPerPerson	1.2618	(0.9467)	0.1111	11
sanitationFacilities	0.3191	(0.4661)	0	1
electricLighting	0.9241	(0.2649)	0	1
refuseCollection	0.8808	(0.3241)	0	1
hasTelephone	0.4425	(0.4967)	0	1
numPrimaryEd	3.0178	(1.7779)	0	13
numProfessional	0.2172	(0.4993)	0	4
numSkilled	0.388	(0.6184)	0	5
femaleHeadedHH	0.1969	(0.3976)	0	1
(mean) farm	0.0448	(0.2069)	0	1

Continued on next page...

... table B.15 continued

Variable	Mean	(Std. Dev.)	Min.	Max.
(mean) urban	0.9272	(0.2598)	0	1
N	6562701			

Table B.16: Summary Statistics - HH Controls, Mpumalanga (Census Data)

Variable	Mean	(Std. Dev.)	Min.	Max.
(mean) logHHsize	1.6604	(0.5730)	0	3.4657
africanHH	0.8938	(0.3081)	0	1
whiteHH	0.0812	(0.2732)	0	1
formalDwelling	0.6202	(0.4853)	0	1
roomsPerPerson	0.9447	(0.7644)	0	28
sanitationFacilities	0.3274	(0.4693)	0	1
electricLighting	0.5712	(0.4949)	0	1
refuseCollection	0.3533	(0.478)	0	1
hasTelephone	0.1568	(0.3636)	0	1
numPrimaryEd	3.2396	(2.118)	0	18
numProfessional	0.1108	(0.3744)	0	5
numSkilled	0.3463	(0.6113)	0	10
femaleHeadedHH	0.3787	(0.4851)	0	1
(mean) farm	0.1218	(0.3271)	0	1
(mean) urban	0.3875	(0.4872)	0	1
(mean) tribal	0.4683	(0.499)	0	1
N	2775474			

Table B.17: Summary Statistics - HH Controls, Mpumalanga (IES Data)

Variable	Mean	(Std. Dev.)	Min.	Max.
(mean) logHHsize	1.7633	(0.4621)	0	2.9957
africanHH	0.9052	(0.2929)	0	1

Continued on next page...

... table B.17 continued

Variable	Mean	(Std. Dev.)	Min.	Max.
whiteHH	0.0777	(0.2677)	0	1
formalDwelling	0.5962	(0.4906)	0	1
roomsPerPerson	0.9917	(0.7212)	0.125	11
sanitationFacilities	0.3572	(0.4792)	0	1
electricLighting	0.5574	(0.4967)	0	1
refuseCollection	0.3393	(0.4735)	0	1
hasTelephone	0.1401	(0.3471)	0	1
numPrimaryEd	2.5433	(1.7665)	0	12
numProfessional	0.0827	(0.3127)	0	2
numSkilled	0.398	(0.5640)	0	4
femaleHeadedHH	0.2326	(0.4225)	0	1
(mean) farm	0.1885	(0.3911)	0	1
(mean) urban	0.2613	(0.4394)	0	1
(mean) tribal	0.3343	(0.4718)	0	1
N	2645663			

Table B.18: Summary Statistics - HH Controls, Limpopo (Census Data)

Variable	Mean	(Std. Dev.)	Min.	Max.
(mean) logHHsize	1.7065	(0.5141)	0	3.5835
africanHH	0.9701	(0.1703)	0	1
whiteHH	0.0226	(0.1485)	0	1
formalDwelling	0.6027	(0.4893)	0	1
roomsPerPerson	0.8420	(0.7029)	0	24
sanitationFacilities	0.1019	(0.3025)	0	1
electricLighting	0.3577	(0.4793)	0	1
refuseCollection	0.0975	(0.2966)	0	1
hasTelephone	0.0614	(0.2401)	0	1
numPrimaryEd	3.2415	(1.9918)	0	20
numProfessional	0.0972	(0.3601)	0	9
numSkilled	0.1561	(0.4432)	0	10
femaleHeadedHH	0.5274	(0.4992)	0	1
(mean) farm	0.043	(0.2029)	0	1

Continued on next page...

... table B.18 continued

Variable	Mean	(Std. Dev.)	Min.	Max.
(mean) urban	0.1047	(0.3062)	0	1
(mean) tribal	0.8470	(0.36)	0	1
N	4738988			

Table B.19: Summary Statistics - HH Controls, Limpopo (IES Data)

Variable	Mean	(Std. Dev.)	Min.	Max.
(mean) logHHsize	1.749	(0.4793)	0	2.9444
africanHH	0.9687	(0.1741)	0	1
whiteHH	0.0239	(0.1529)	0	1
formalDwelling	0.5999	(0.4899)	0	1
roomsPerPerson	0.9266	(0.6494)	0.1111	13
sanitationFacilities	0.2391	(0.4265)	0	1
electricLighting	0.3381	(0.4731)	0	1
refuseCollection	0.1365	(0.3434)	0	1
hasTelephone	0.0911	(0.2878)	0	1
numPrimaryEd	2.6196	(1.7836)	0	9
numProfessional	0.1696	(0.4645)	0	4
numSkilled	0.1367	(0.396)	0	4
femaleHeadcdHH	0.4374	(0.4961)	0	1
(mean) farm	0.0165	(0.1272)	0	1
(mean) urban	0.1096	(0.3123)	0	1
(mean) tribal	0.7686	(0.4217)	0	1
N	4773999			

B.3 Area-Level Controls (Census Covariates)

Table B.20: Summary Statistics - Area Controls, W Cape
(Census Data)

Variable	Mean	(Std. Dev.)	Min.	Max.
logHHsize	1.1543	(0.1018)	0.8254	1.3458
africanHH	0.2256	(0.2377)	0	0.6971
whiteHH	0.2652	(0.1692)	0.0002	0.5972
formalDwelling	0.7738	(0.1622)	0.4632	0.9615
roomsPerPerson	1.4784	(0.3072)	1.0294	2.1681
sanitationFacilities	0.8516	(0.1127)	0.232	0.9795
electricLighting	0.8473	(0.1)	0.665	0.9773
refuseCollection	0.8416	(0.1335)	0.3	0.9753
hasTelephone	0.5366	(0.185)	0.2525	0.8184
numPrimaryEd	2.692	(0.2331)	1.9412	2.9899
numProfessional	0.2193	(0.1045)	0.0368	0.3865
numSkilled	0.3057	(0.0496)	0.1714	0.4930
femaleHeadedHH	0.2743	(0.0511)	0.1429	0.3382
tribal	0.0002	(0.0014)	0	0.0127
urban	0.8846	(0.1645)	0.3268	1
farm	0.1085	(0.1564)	0	0.5631
waterServices	0.8913	(0.0855)	0.6634	0.9833
propertyOwnedByHH	0.6870	(0.1501)	0.2313	0.9106
N	3803234			

Table B.21: Summary Statistics - Area Controls, W Cape
(IES Data)

Variable	Mean	(Std. Dev.)	Min.	Max.
logHHsize	1.1535	(0.1019)	0.8254	1.3458
africanHH	0.1627	(0.1678)	0	0.6971
whiteHH	0.2858	(0.1387)	0.0002	0.5972
formalDwelling	0.8087	(0.1182)	0.4632	0.9615
roomsPerPerson	1.5021	(0.2548)	1.0294	2.1681

Continued on next page...

... table B.21 continued

Variable	Mean	(Std. Dev.)	Min.	Max.
sanitationFacilities	0.8157	(0.1343)	0.232	0.9795
electricLighting	0.8357	(0.0858)	0.665	0.9773
refuseCollection	0.7812	(0.1535)	0.3	0.9753
hasTelephone	0.4971	(0.1588)	0.2525	0.8184
numPrimaryEd	2.5847	(0.2631)	1.9412	2.9899
numProfessional	0.1901	(0.0936)	0.0368	0.3865
numSkilled	0.3	(0.0572)	0.1714	0.4930
femaleHeadedHH	0.2539	(0.0481)	0.1429	0.3382
tribal	0.0003	(0.0019)	0	0.0127
urban	0.7967	(0.187)	0.3268	1
farm	0.1929	(0.1773)	0	0.5631
waterServices	0.896	(0.0695)	0.6634	0.9833
propertyOwnedByHH	0.6306	(0.1437)	0.2313	0.9106
N	3860967			

Table B.22: Summary Statistics - Area Controls, E Cape
(Census Data)

Variable	Mean	(Std. Dev.)	Min.	Max.
logHHsize	1.2796	(0.1144)	1.0634	1.4955
africanHH	0.8597	(0.2117)	0.0446	0.9981
whiteHH	0.0694	(0.1056)	0	0.3178
formalDwelling	0.4172	(0.2344)	0.0731	0.9582
roomsPerPerson	1.1074	(0.2241)	0.7807	1.7982
sanitationFacilities	0.2798	(0.3247)	0.0019	0.8423
electricLighting	0.2935	(0.2721)	0.0108	0.864
refuseCollection	0.3254	(0.3581)	0.0005	0.9289
hasTelephone	0.141	(0.1735)	0.001	0.5183
numPrimaryEd	2.3946	(0.3059)	1.4545	2.9073
numProfessional	0.1023	(0.0607)	0.0266	0.2922
numSkilled	0.1218	(0.0917)	0.0219	0.4232
femaleHeadedHH	0.4976	(0.1596)	0.1369	0.6934
tribal	0.5711	(0.4256)	0	1
urban	0.3723	(0.3827)	0	0.9730

Continued on next page...

... table B.22 continued

Variable	Mean	(Std. Dev.)	Min.	Max.
farm	0.0398	(0.0899)	0	0.6082
waterServices	0.3248	(0.3268)	0.0073	0.9036
propertyOwnedByHH	0.8609	(0.0951)	0.3185	0.9595
N	6167770			

Table B.23: Summary Statistics - Area Controls, E Cape (IES Data)

Variable	Mean	(Std. Dev.)	Min.	Max.
logHHsize	1.2803	(0.1048)	1.0634	1.4955
africanHH	0.8619	(0.2145)	0.0446	0.9981
whiteHH	0.0586	(0.0884)	0	0.3178
formalDwelling	0.4513	(0.2525)	0.0731	0.9582
roomsPerPerson	1.1105	(0.2183)	0.7807	1.7982
sanitationFacilities	0.2288	(0.2719)	0.0019	0.8423
electricLighting	0.2982	(0.2682)	0.0108	0.864
refuseCollection	0.3091	(0.3247)	0.0005	0.9289
hasTelephone	0.1248	(0.1499)	0.001	0.5183
numPrimaryEd	2.3217	(0.3058)	1.4545	2.9073
numProfessional	0.0919	(0.0531)	0.0266	0.2922
numSkilled	0.1237	(0.0929)	0.0219	0.4232
femaleHeadedHH	0.4776	(0.1721)	0.1369	0.6934
tribal	0.5398	(0.4337)	0	1
urban	0.3585	(0.3506)	0	0.9730
farm	0.0762	(0.1271)	0	0.6082
waterServices	0.3178	(0.3104)	0.0073	0.9036
propertyOwnedByHH	0.8355	(0.1218)	0.3185	0.9595
N	6059647			

Table B.24: Summary Statistics - Area Controls, N Cape
(Census Data)

Variable	Mean	(Std. Dev.)	Min.	Max.
logHHsize	1.204	(0.0771)	0.9277	1.3082
africanHH	0.3161	(0.2003)	0	0.7010
whiteHH	0.1815	(0.0439)	0.0971	0.3131
formalDwelling	0.7646	(0.0864)	0.5773	0.9907
roomsPerPerson	1.3143	(0.123)	1.0438	1.7407
sanitationFacilities	0.6063	(0.1904)	0.1702	0.8961
electricLighting	0.7183	(0.0836)	0.4947	0.8314
refuseCollection	0.7120	(0.1746)	0.3912	0.9351
hasTelephone	0.3094	(0.0743)	0.1583	0.3915
numPrimaryEd	2.3663	(0.409)	1.4255	2.8106
numProfessional	0.1267	(0.0547)	0.0319	0.2107
numSkilled	0.2658	(0.0764)	0.0909	0.7365
femaleHeadedHH	0.2947	(0.0371)	0.2039	0.3896
urban	0.6987	(0.2259)	0	0.9679
farm	0.2483	(0.2074)	0.0321	1
waterServices	0.8360	(0.0936)	0.506	0.9739
propertyOwnedByHH	0.6654	(0.1443)	0.4211	0.8461
N	802263			

Table B.25: Summary Statistics - Area Controls, N Cape
(IES Data)

Variable	Mean	(Std. Dev.)	Min.	Max.
logHHsize	1.1901	(0.0859)	0.9277	1.3082
africanHH	0.2924	(0.2018)	0	0.7010
whiteHH	0.1756	(0.0522)	0.0971	0.3131
formalDwelling	0.7806	(0.1031)	0.5773	0.9907
roomsPerPerson	1.3232	(0.1541)	1.0438	1.7407
sanitationFacilities	0.5177	(0.1884)	0.1702	0.8961
electricLighting	0.6969	(0.0812)	0.4947	0.8314
refuseCollection	0.6543	(0.1576)	0.3912	0.9351
hasTelephone	0.2829	(0.0722)	0.1583	0.3915

Continued on next page...

... table B.25 continued

Variable	Mean	(Std. Dev.)	Min.	Max.
numPrimaryEd	2.1632	(0.3849)	1.4255	2.8106
numProfessional	0.102	(0.0464)	0.0319	0.2107
numSkilled	0.2624	(0.0891)	0.0909	0.7365
femaleHeadedHH	0.2931	(0.0407)	0.2039	0.3896
urban	0.6416	(0.2237)	0	0.9679
farm	0.3203	(0.2172)	0.0321	1
waterServices	0.8120	(0.1115)	0.506	0.9739
propertyOwnedByHH	0.6354	(0.1331)	0.4211	0.8461
N	811126			

Table B.26: Summary Statistics - Area Controls, Free State
(Census Data)

Variable	Mean	(Std. Dev.)	Min.	Max.
logHHsize	1.169	(0.0964)	0.8955	1.3205
africanHH	0.8170	(0.1351)	0.4333	0.9952
whiteHH	0.1436	(0.1009)	0.0003	0.3073
formalDwelling	0.5794	(0.1013)	0.3578	0.9055
roomsPerPerson	1.2717	(0.1624)	0.9422	1.7691
sanitationFacilities	0.4292	(0.2674)	0.0723	0.8269
electricLighting	0.5601	(0.204)	0.1675	0.8021
refuseCollection	0.6372	(0.2304)	0.1182	0.9489
hasTelephone	0.2175	(0.1218)	0.0506	0.4252
numPrimaryEd	2.3395	(0.2052)	1.5055	2.6163
numProfessional	0.1137	(0.0537)	0.032	0.2192
numSkilled	0.2661	(0.0598)	0.0947	0.4792
femaleHeadedHH	0.3449	(0.075)	0.1395	0.498
tribal	0.1179	(0.2843)	0	0.8472
urban	0.7152	(0.2785)	0	1
farm	0.1621	(0.1874)	0	1
waterServices	0.6936	(0.2036)	0.2991	0.9349
propertyOwnedByHH	0.7564	(0.1196)	0.4659	0.9581
N	2473262			

Table B.27: Summary Statistics - Arca Controls, Free State (IES Data)

Variable	Mean	(Std. Dev.)	Min.	Max.
logHHsize	1.1751	(0.0868)	0.8955	1.3205
africanHH	0.8094	(0.1226)	0.4333	0.9952
whiteHH	0.1409	(0.0775)	0.0003	0.3073
formalDwelling	0.5812	(0.121)	0.3578	0.9055
roomsPerPerson	1.2819	(0.1591)	0.9422	1.7691
sanitationFacilities	0.4058	(0.2445)	0.0723	0.8269
electricLighting	0.5917	(0.1724)	0.1675	0.8021
refuseCollection	0.6419	(0.1796)	0.1182	0.9489
hasTelephone	0.2075	(0.095)	0.0506	0.4252
numPrimaryEd	2.2476	(0.2561)	1.5055	2.6163
numProfessional	0.0967	(0.046)	0.032	0.2192
numSkilled	0.2761	(0.0616)	0.0947	0.4792
femaleHeadedHH	0.327	(0.0652)	0.1395	0.498
tribal	0.0597	(0.2069)	0	0.8472
urban	0.6969	(0.2211)	0	1
farm	0.2392	(0.187)	0	1
waterServices	0.731	(0.1685)	0.2991	0.9349
propertyOwnedByHH	0.7161	(0.1149)	0.4659	0.9581
N	2448094			

Table B.28: Summary Statistics: Area Controls, KZN (Census Data)

Variable	Mean	(Std. Dev.)	Min.	Max.
logHHsize	1.3546	(0.2029)	0.9524	1.7222
africanHH	0.794	(0.2207)	0.0955	0.9985
whiteHH	0.0842	(0.1062)	0	0.3836
formalDwelling	0.4705	(0.1978)	0.0383	0.7904
roomsPerPerson	1.1515	(0.2159)	0.7859	1.6816
sanitationFacilities	0.3634	(0.2669)	0.0044	0.8988
electricLighting	0.4893	(0.2701)	0.0104	0.9148
refuseCollection	0.3765	(0.2861)	0.0004	0.9064

Continued on next page...

... table B.28 continued

Variable	Mean	(Std. Dev.)	Min.	Max.
hasTelephone	0.2293	(0.1937)	0.004	0.6708
numPrimaryEd	2.6706	(0.359)	1.6042	3.3218
numProfessional	0.1261	(0.0742)	0.0303	0.3061
numSkilled	0.2014	(0.0905)	0.0244	0.4917
femaleHeadedHH	0.4087	(0.1056)	0.2298	0.6637
tribal	0.4466	(0.3749)	0	1
urban	0.4534	(0.3668)	0	1
farm	0.0612	(0.109)	0	0.7826
waterServices	0.4276	(0.2691)	0.0127	0.9302
propertyOwnedByHH	0.8133	(0.1132)	0.3669	0.9705
N	8097994			

Table B.29: Summary Statistics - Area Controls, KZN (IES Data)

Variable	Mean	(Std. Dev.)	Min.	Max.
logHHsize	1.3919	(0.1881)	0.9524	1.7222
africanHH	0.8310	(0.1831)	0.0955	0.9985
whiteHH	0.0745	(0.0874)	0	0.3836
formalDwelling	0.4569	(0.187)	0.0383	0.7904
roomsPerPerson	1.1267	(0.1957)	0.7859	1.6816
sanitationFacilities	0.3222	(0.2397)	0.0044	0.8988
electricLighting	0.4635	(0.2554)	0.0104	0.9148
refuseCollection	0.3304	(0.2573)	0.0004	0.9064
hasTelephone	0.1944	(0.1646)	0.004	0.6708
numPrimaryEd	2.6932	(0.3603)	1.6042	3.3218
numProfessional	0.1136	(0.0625)	0.0303	0.3061
numSkilled	0.2014	(0.0904)	0.0244	0.4917
femaleHeadedHH	0.4152	(0.0977)	0.2298	0.6637
tribal	0.4945	(0.3563)	0	1
urban	0.3898	(0.3313)	0	1
farm	0.0776	(0.1327)	0	0.7826
waterServices	0.3947	(0.2462)	0.0127	0.9302
propertyOwnedByHH	0.8238	(0.1085)	0.3669	0.9705

Continued on next page...

... table B.29 continued

Variable	Mean	(Std. Dev.)	Min.	Max.
N	7786987			

Table B.30: Summary Statistics - Area Controls, North West
(Census Data)

Variable	Mean	(Std. Dev.)	Min.	Max.
logHHsize	1.2439	(0.1343)	0.9847	1.4504
africanHH	0.8947	(0.1105)	0.638	0.9954
whiteHH	0.0815	(0.0922)	0	0.298
formalDwelling	0.6564	(0.0963)	0.5092	0.8036
roomsPerPerson	1.2528	(0.1171)	0.9199	1.5269
sanitationFacilities	0.2966	(0.1981)	0.0367	0.7219
electricLighting	0.4285	(0.1759)	0.1409	0.7129
refuseCollection	0.3353	(0.2473)	0.0327	0.8632
hasTelephone	0.1566	(0.0993)	0.0219	0.38
numPrimaryEd	2.4456	(0.2956)	1.8208	2.9428
numProfessional	0.1118	(0.029)	0.0657	0.1726
numSkilled	0.2657	(0.0875)	0.1258	0.3797
femaleHeadedHH	0.382	(0.0899)	0.2433	0.5469
tribal	0.443	(0.3202)	0	0.9582
urban	0.3609	(0.2711)	0	0.8954
farm	0.1038	(0.1171)	0	0.482
waterServices	0.475	(0.2339)	0.1015	0.8688
propertyOwnedByHH	0.8098	(0.1196)	0.4467	0.9580
N	3216039			

Table B.31: Summary Statistics - Area Controls, North West
(IES Data)

Variable	Mean	(Std. Dev.)	Min.	Max.
logHHsize	1.2399	(0.1442)	0.9847	1.4504

Continued on next page...

... table B.31 continued

Variable	Mean	(Std. Dev.)	Min.	Max.
africanHH	0.8946	(0.1005)	0.638	0.9954
whiteHH	0.0801	(0.0785)	0	0.298
formalDwelling	0.6541	(0.0912)	0.5092	0.8036
roomsPerPerson	1.2438	(0.0953)	0.9199	1.5269
sanitationFacilities	0.2624	(0.1652)	0.0367	0.7219
electricLighting	0.4101	(0.1672)	0.1409	0.7129
refuseCollection	0.3042	(0.2116)	0.0327	0.8632
hasTelephone	0.1444	(0.0816)	0.0219	0.38
numPrimaryEd	2.389	(0.3263)	1.8208	2.9428
numProfessional	0.1025	(0.0242)	0.0657	0.1726
numSkilled	0.2657	(0.0839)	0.1258	0.3797
femaleHeadedHH	0.3789	(0.0957)	0.2433	0.5469
tribal	0.4255	(0.3141)	0	0.9582
urban	0.3272	(0.2331)	0	0.8954
farm	0.1565	(0.1489)	0	0.482
waterServices	0.4501	(0.2004)	0.1015	0.8688
propertyOwnedByHH	0.7835	(0.1372)	0.4467	0.9580
N	2014530			

Table B.32: Summary Statistics - Area Controls, Gauteng
(Census Data)

Variable	Mean	(Std. Dev.)	Min.	Max.
logHHsize	1.0405	(0.1184)	0.7981	1.3089
africanHH	0.6793	(0.2167)	0.2586	0.9962
whiteHH	0.2454	(0.1689)	0.0001	0.6123
formalDwelling	0.6307	(0.1197)	0.2709	0.8067
roomsPerPerson	1.4128	(0.2938)	0.9911	1.935
sanitationFacilities	0.8207	(0.1405)	0.3909	0.9578
electricLighting	0.7896	(0.1282)	0.3636	0.9339
refuseCollection	0.8441	(0.1458)	0.1769	0.9425
hasTelephone	0.4369	(0.1493)	0.1312	0.7038
numPrimaryEd	2.4827	(0.2716)	1.9664	3.0063
numProfessional	0.2173	(0.0902)	0.0875	0.4018

Continued on next page...

... table B.32 continued

Variable	Mean	(Std. Dev.)	Min.	Max.
numSkilled	0.3071	(0.0595)	0.1779	0.4343
femaleHeadedHH	0.2871	(0.0352)	0.2243	0.3679
urban	0.9668	(0.0522)	0.6009	1
farm	0.0301	(0.0518)	0	0.3991
waterServices	0.8473	(0.0881)	0.5404	0.9354
propertyOwnedByHH	0.7462	(0.0945)	0.5638	0.9652
N	6890762			

Table B.33: Summary Statistics - Area Controls, Gauteng (IES Data)

Variable	Mean	(Std. Dev.)	Min.	Max.
logHHsize	1.0342	(0.1071)	0.7981	1.3089
africanHH	0.6573	(0.1688)	0.2586	0.9962
whiteHH	0.2657	(0.1343)	0.0002	0.6123
formalDwelling	0.612	(0.1234)	0.2709	0.8067
roomsPerPerson	1.4352	(0.2389)	1.0932	1.935
sanitationFacilities	0.7892	(0.1389)	0.3909	0.9223
electricLighting	0.7565	(0.1285)	0.3636	0.9339
refuseCollection	0.8170	(0.1443)	0.1769	0.9404
hasTelephone	0.4258	(0.1459)	0.1312	0.7038
numPrimaryEd	2.4152	(0.1974)	1.9664	2.89
numProfessional	0.2111	(0.0845)	0.0875	0.4018
numSkilled	0.3192	(0.0599)	0.1779	0.4343
femaleHeadedHH	0.2739	(0.0251)	0.2243	0.3679
urban	0.9509	(0.0679)	0.6009	1
farm	0.0466	(0.0683)	0	0.3991
waterServices	0.8228	(0.0903)	0.5404	0.9254
propertyOwnedByHH	0.7433	(0.091)	0.5638	0.9652
N	6562701			

Table B.34: Summary Statistics - Area Controls, Mpumalanga (Census Data)

Variable	Mean	(Std. Dev.)	Min.	Max.
logHHsize	1.2763	(0.1695)	0.7241	1.47
africanHH	0.8868	(0.1282)	0.5485	0.9989
whiteHH	0.0924	(0.1096)	0	0.3866
formalDwelling	0.6092	(0.1262)	0.2238	0.8394
roomsPerPerson	1.2683	(0.1625)	1.0105	1.6376
sanitationFacilities	0.3324	(0.2854)	0.0067	0.8066
electricLighting	0.5553	(0.2103)	0.227	0.8781
refuseCollection	0.3545	(0.2731)	0.0034	0.8463
hasTelephone	0.1544	(0.1347)	0.0086	0.4275
numPrimaryEd	2.4051	(0.2835)	1.4859	2.8969
numProfessional	0.101	(0.0418)	0.0386	0.2582
numSkilled	0.302	(0.1053)	0.1165	0.525
femaleHeadedHH	0.3778	(0.125)	0.1036	0.5654
tribal	0.4588	(0.4165)	0	1
urban	0.3924	(0.3039)	0	0.9521
farm	0.1241	(0.1673)	0	0.5725
waterServices	0.6094	(0.1764)	0.115	0.9009
propertyOwnedByHH	0.8373	(0.1503)	0.3076	0.9734
N	2775474			

Table B.35: Summary Statistics - Area Controls, Mpumalanga (IES Data)

Variable	Mean	(Std. Dev.)	Min.	Max.
logHHsize	1.2443	(0.1987)	0.7241	1.47
africanHH	0.8687	(0.1283)	0.5485	0.9989
whiteHH	0.1092	(0.1104)	0	0.3866
formalDwelling	0.5982	(0.1422)	0.2238	0.8341
roomsPerPerson	1.3046	(0.16)	1.0105	1.5908
sanitationFacilities	0.3552	(0.2776)	0.0067	0.7482
electricLighting	0.5658	(0.2013)	0.227	0.8781
refuseCollection	0.3712	(0.2653)	0.0034	0.8109

Continued on next page...

... table B.35 continued

Variable	Mean	(Std. Dev.)	Min.	Max.
hasTelephone	0.1664	(0.1341)	0.0086	0.4275
numPrimaryEd	2.3302	(0.3473)	1.4859	2.8969
numProfessional	0.0971	(0.0464)	0.0386	0.2582
numSkilled	0.303	(0.1034)	0.1165	0.5214
femaleHeadedHH	0.3654	(0.1266)	0.1889	0.5654
tribal	0.3864	(0.4133)	0	1
urban	0.3993	(0.2905)	0	0.9521
farm	0.1694	(0.1834)	0	0.5725
waterServices	0.6024	(0.1773)	0.115	0.8201
propertyOwnedByHH	0.7993	(0.1613)	0.3221	0.9734
N	2645663			

Table B.36: Summary Statistics - Area Controls, Limpopo
(Census Data)

Variable	Mean	(Std. Dev.)	Min.	Max.
logHHsize	1.3825	(0.1359)	0.7629	1.4913
africanHH	0.9678	(0.1017)	0.382	0.999
whiteHH	0.0251	(0.0892)	0	0.498
formalDwelling	0.5876	(0.1521)	0.2007	0.8118
roomsPerPerson	1.1283	(0.1587)	0.9226	1.8452
sanitationFacilities	0.1095	(0.1621)	0.0068	0.8354
electricLighting	0.3553	(0.153)	0.088	0.8435
refuseCollection	0.106	(0.1314)	0.0048	0.6918
hasTelephone	0.0625	(0.0936)	0.0073	0.5831
numPrimaryEd	2.5192	(0.2821)	1.3906	2.8952
numProfessional	0.0921	(0.0362)	0.0528	0.3455
numSkilled	0.1334	(0.08)	0.0568	0.4594
femaleHeadedHH	0.5356	(0.0938)	0.1571	0.6547
tribal	0.8406	(0.2485)	0	1
urban	0.115	(0.1451)	0	0.6788
farm	0.039	(0.1258)	0	0.5769
waterServices	0.3328	(0.1649)	0.1387	0.8954
propertyOwnedByHH	0.9202	(0.1323)	0.2934	0.9854

Continued on next page...

... table B.36 continued

Variable	Mean	(Std. Dev.)	Min.	Max.
N	4738988			

Table B.37: Summary Statistics - Area Controls, Limpopo (IES Data)

Variable	Mean	(Std. Dev.)	Min.	Max.
logHHsize	1.3597	(0.1629)	0.7629	1.4913
africanHH	0.9540	(0.126)	0.382	0.999
whiteHH	0.0374	(0.1095)	0	0.498
formalDwelling	0.5768	(0.1561)	0.2007	0.8118
roomsPerPerson	1.1436	(0.1876)	0.9226	1.8452
sanitationFacilities	0.1328	(0.1932)	0.0068	0.8354
electricLighting	0.3804	(0.1689)	0.088	0.8435
refuseCollection	0.1236	(0.1569)	0.0048	0.6918
hasTelephone	0.0751	(0.1137)	0.0073	0.5831
numPrimaryEd	2.4655	(0.3107)	1.3906	2.8952
numProfessional	0.0932	(0.0408)	0.0528	0.3455
numSkilled	0.1441	(0.0963)	0.0568	0.4594
femaleHeadedHH	0.5248	(0.1125)	0.1571	0.6547
tribal	0.8025	(0.297)	0	1
urban	0.1348	(0.1737)	0	0.6788
farm	0.0576	(0.1545)	0	0.5769
waterServices	0.3503	(0.1874)	0.1387	0.8954
propertyOwnedByHH	0.9014	(0.1606)	0.2934	0.9854
N	4773999			

Bibliography

- H. Alderman, M. Babita, J. Lanjouw, P. Lanjouw, N. Makhatha, A. Mohamed, B. Özler, and O. Qaba. Combining Census and Survey data to Construct a Poverty Map of South Africa. In R. Hirschowitz, A. Head, and S. Africa, editors, *Measuring Poverty in South Africa*, chapter 2, pages 5 – 52. Statistics South Africa, Pretoria, South Africa, 2000.
- H. Alderman, M. Babita, G. Demombynes, N. Makhatha, and B. Özler. How Low Can You Go? Combining Census and Survey Data for Mapping Poverty in South Africa. *Journal of African Economies*, 11(2):169, 2002.
- A. Banerjee and E. Duflo. Inequality and Growth: What Can the Data Say? *Journal of Economic Growth*, 8(3):267–299, 2003.
- A. Baschieri and J. Falkingham. Developing a Poverty Map of Tajikistan: a Technical Note. Applications & Policy Working Paper A05-11, Southampton Statistical Sciences Research Institute, 2005.
- A. Baschieri, J. Falkingham, D. Hornby, and C. Hutton. Creating a Poverty Map for Azerbaijan. Policy Research Working Paper 3793, World Bank, 2005.
- C. Carletton, A. Dabalen, and A. Moubayed. Constructing and Using Poverty Maps for Policy Making: The Experience in Albania. In T. Bedi, A. Coudouel, and K. Simler, editors, *More Than a Pretty Picture: Using Poverty Maps to Design Better Policies and Interventions*, chapter 3, pages 53 – 66. World Bank, Washington, D.C., 2007.
- A. Deaton. *The Analysis of Household Surveys: A Microeconomic Approach to Development Policy*. Johns Hopkins University Press, 1997.
- G. Demombynes and B. Özler. Crime and Local Inequality in South Africa. *Journal of Development Economics*, 76(2):265–292, 2005.

- Department of Local Government, Western Cape Provincial Administration. Submission on Municipal Equitable Share Formula. Technical Report, Cape Town, 2002.
- C. Dube. Census Geography of South Africa. Technical report, Statistics South Africa, 2005. URL <http://www.statssa.gov.za/africagis2005/presentations/OralColemanDube.pdf>. Presented at the 2005 AfricaGIS Conference.
- C. Elbers, J. Lanjouw, and P. Lanjouw. Micro-Level Estimation of Welfare. Policy Research Working Paper 2911, World Bank Development Research Group, 2002.
- C. Elbers, J. Lanjouw, and P. Lanjouw. Micro-Level Estimation of Poverty and Inequality. *Econometrica*, 71(1):355–364, 2003a.
- C. Elbers, P. Lanjouw, J. Mistiaen, B. Özler, and K. Simler. Are Neighbours Equal?: Estimating Local Inequality in Three Developing Countries. Discussion Paper, Food Consumption and Nutrition Division 147, International Food Policy Research Institute, 2003b. URL <http://www.ifpri.org/divs/fcnd/dp/papers/fcndp147.pdf>.
- C. Elbers, T. Fujii, P. Lanjouw, B. Özler, and W. Yin. Poverty Alleviation through Geographic Targeting: How Much Does Disaggregation Help? *Journal of Development Economics*, 83(1):198–213, 2007.
- C. Elbers, P. Lanjouw, and P. G. Leite. Brazil Within Brazil: Testing the Poverty Map Methodology in Minas Gerais. Policy Research Working Paper 4513, World Bank Development Research Group, February 2008.
- T. Fujii. Commune-Level Estimation of Poverty Measures and Its Application in Cambodia. Available online at <http://siteresources.worldbank.org/INTPGI/Resources/342674-1092157888460/Fujii.Commune-LevelCambodia.pdf>, June 2003.
- J. Gibson, G. Datt, B. Allen, V. Hwang, R. Bourke, and D. Parajuli. Mapping Poverty in Rural Papua New Guinea. *Pacific Economic Bulletin*, 19(4):14–29, 2005.
- A. J. Healy, S. Jitsuchon, and Y. Vajaragupta. Spatially Disaggregated Estimates of Poverty and Inequality in Thailand. Unpublished Technical Report, World Bank, World Bank, September 2003. Available online at <http://siteresources.worldbank.org/INTPGI/Resources/342674-1092157888460/Healy.DisaggregatedThailand.pdf>.

- N. Henninger and M. Snel. Where Are the Poor? Experiences with the Development and Use of Poverty Maps. Technical report, World Resources Institute and UNEP/GRID, 2002. Available online at http://www.povertymap.net/publications/wherethepoor/where_are_the_poor.pdf.
- J. Hentschel and P. Lanjouw. Using Disaggregated Poverty Maps to Plan Sectoral Investments. PREMnotes 5, World Bank, Washington, D.C., May 1998.
- J. Hentschel, J. Lanjouw, P. Lanjouw, and J. Poggi. Combining Census and Survey Data to Trace the Spatial Dimensions of Poverty: A Case Study of Ecuador. *The World Bank Economic Review*, 14(1):147–165, 2000. URL <http://wber.oxfordjournals.org/cgi/content/abstract/14/1/147>.
- S. Jitsuchon and K. Richter. Thailand's Poverty Maps: From Construction to Application. In T. Bedi, A. Coudouel, and K. Simler, editors, *More Than a Pretty Picture: Using Poverty Maps to Design Better Policies and Interventions*, chapter 13, pages 241 – 260. World Bank, Washington, D.C., 2007.
- P. Lanjouw. The Geography of Poverty in Morocco: Micro-Level Estimates of Poverty and Inequality from Combined Census and Household Survey Data. Unpublished Technical Report, World Bank, World Bank, February 2004. Available online at <http://siteresources.worldbank.org/INTPGI/Resources/342674-1092157888460/Lanjouw.GeographyPovertyMorocco.pdf>.
- P. Lanjouw and M. Ravallion. Response to the Evaluation Panel's Critique of Poverty Mapping. Technical report, World Bank, October 2006.
- P. Lanjouw, J. Lanjouw, C. Elbers, and G. Demombynes. How Good a Map? Putting Small Area Estimation to the Test. Policy Research Working Paper 4155, World Bank, March 2007.
- J. Litvack. The Poverty Mapping Application in Morocco. In T. Bedi, A. Coudouel, and K. Simler, editors, *More Than a Pretty Picture: Using Poverty Maps to Design Better Policies and Interventions*, chapter 11, pages 208 – 224. World Bank, 2007.
- L. J. Loots. Equity and the Local Government Equitable Share in South Africa. In *10 Years of the FFC: Consolidation for Greater Equity, Conference*, Cape Town, August 2004. Financial

and Fiscal Commission of South Africa. Available online at <http://www.ffc.co.za/conf/papers/lesequity.pdf>.

N. Minot and B. Baulch. Spatial Patterns of Poverty in Vietnam and Their Implications for Policy. *Food Policy*, 30(5-6):461–475, 2005.

National Treasury, Republic of South Africa. *2009 National Budget Review*, chapter 8 - Division of Revenue and Intergovernmental Transfers. Pretoria, March 2009a.

National Treasury, Republic of South Africa. *2009 National Budget Review, Annexure W1: Explanatory Memorandum to the Division of Revenue*. Pretoria, March 2009b.

G. Ndeng'e, C. Opiyo, J. Mistiaen, and P. Kristjanson. Geographic Dimensions of Well-Being in Kenya: Where Are The Poor? From Districts to Locations. Technical Report, Central Bureau of Statistics and Ministry of Planning and National Development, Kenya, Nairobi, 2003. Available online at <http://go.worldbank.org/OX3RULGI00>.

Parliament of the Republic of South Africa. *Constitution of the Republic of South Africa*, chapter 13 - Finance. Cape Town, December 1996. Act Number 108 of 1996. Date of Commencement: 4 February, 1997.

Parliament of the Republic of South Africa. *Division of Revenue Act, 2009*. Cape Town, April 2009. Act Number 12 of 2009.

K. Simler and V. Nhate. Poverty, Inequality and Geographic Targeting: Evidence from Small-Area Estimates in Mozambique. In *Northeast Universities Development Consortium Conference*, pages 25–27. Williams College, October 2002.

A. Tarozzi. Can Census Data Alone Signal Heterogeneity in the Estimation of Poverty Maps? Working Paper, Duke University, 2008.

T. Vishwanath and N. Yoshida. Poverty Maps in Sri Lanka: Policy Impacts and Lessons. In T. Bedi, A. Coudouel, and K. Simler, editors, *More Than a Pretty Picture: Using Poverty Maps to Design Better Policies and Interventions*, chapter 12, pages 225 – 240. World Bank, Washington, D.C., 2007.

World Bank. *More Than a Pretty Picture: Using Poverty Maps to Design Better Policies and Interventions*. World Bank, Washington, D.C., 2007. Bedi, Tara and Coudouel, Aline and Simler, Kenneth, eds.