Populating a Simulation Model of the Food System in Zimbabwe with Representative Household Data

SW Gundry, JA Wright, Institute of Ecology and Resource Management, University of Edinburgh
E Austin, Department of Psychology, University of Edinburgh
EJ Worrall, Institute of Ecology and Resource Management, University of Edinburgh

Abstract

As part of a simulation model of the food system of a district in Zimbabwe, a database of households has been created to mimic the extant population. The methodology adopted combines census data with survey data from a 1% sample of households in selected wards. In many developing countries disaggregated data at a local scale are rarely available from government statistical services. Typically, published census reports summarise data collected at district level, with only limited demographic information for the constituent wards. As dynamic simulation models are developed to represent local socio-economies, such as the current model for the district of Buhera in Zimbabwe, they require to be populated with households fitting as closely as possible the profile of the 'real' households, taking proper account of spatial variability across the district. This paper reports the development of a methodology that uses various statistical techniques to extrapolate household attributes from a small sample, using the demographics of the census as the target numbers of households for each ward within the district. The wards were characterised based on agricultural and census data. Standardised multi-dimensional distances were calculated using these characteristics and used to identify the survey ward that was most similar to each of the unsurveyed wards. A population was then extrapolated on a ward by ward basis from the households in the matched survey ward. Because few if any of the household attributes are independent, this extrapolation characterises household data as a series of multivariate normal distributions for each survey ward, from which data for the required number of households can be drawn. The methodology, including both the statistical techniques used and computer algorithms employed are described. The various sources of data, the census and the household survey, are also outlined.

1 INTRODUCTION

1.1 Purpose

This paper is concerned with the setting up of a database of approximately 40,000 households and their related agricultural resources and socio-economic data. The database forms part of an agent-based model that simulates the effects of changes in the food system upon nutritional status. The model gives a view of food security at a local level, for a single district of Zimbabwe.

1.2 Food Security

Food security in developing countries is recognised as the result of the interaction of various economic, environmental and health factors, coupled with the complex social relationships within and between households and communities. Hay [1994] suggests that food availability, both in terms of quantity and quality, is mediated by a *hierarchy of relationships* – households within communities, communities within countries, countries within the world at large. Tomkins and Watson [1989] emphasise that the underlying cause of poor nutritional status is the interaction of disease and poor care with the effects of reduced food access at household level. Tomkins and Watson describe this as the *malnutrition-infection complex*. Thus, variability in the spatial and temporal levels of malnutrition requires an examination of the factors that influence food access and health at household level, rather than more limited analysis of supply side variables. Maxwell and Frankenburger [1992] observe that by the end of the 1980's the focus had shifted from national and global food supplies to questions of *access to food at household and individual levels*.

1.3 Simulation Model of the Food and Nutrition System

This interaction of reduced food access with disease and poor care is the fundamental mechanism embodied within a simulation model of the food and nutrition system developed, as part of a collaborative African-European research project, for the Buhera District of Zimbabwe. This concept of *nutritional security* is similar to that attributed to UNICEF in Maxwell & Frankenberger [*ibid*]. Data collection for the research project comprised a field survey and the extraction of numerous secondary data sets. The household resources and characteristics are derived from a combination of the field survey data and census, agricultural and health statistics. The field survey comprised multi-round anthropometry and household questionnaires. The development of the simulation model was reported previously by Gundry et al. [1997]. In brief, however, the model comprises a database that mimics the population of the study district as closely as possible, with one record for each of approximately 40,000 households thus treating each household as an individual 'agent'. Each simulation processes all 40,000 household records in ten-day time steps, throughout an agricultural year. Localised rainfall is simulated from meteorological data and input to a simple crop growth component of the model. Household response to these changes is modelled from a rule base that takes account of location and season. Changes in levels of grain, cash, other assets and disease initiate interactions with other households. Imbalances of demand and supply at ward level are satisfied through an inter-ward trading mechanism, transferring grain across the district.

2. Data

2.1 Buhera District

Buhera District lies two hundred kilometres south of Harare in the province of Manicaland. It straddles the three least productive of the five agroecological zones of Zimbabwe. The southern part of Buhera falls under zone V where rainfall levels are low and the lands unproductive. The quality of land improves northwards so that the northern third of the district, falling under zone III, is generally food self-sufficient in aggregate. The interior of the district is served by a poorly maintained road system. The 1992 census published by the Central Statistics Office (CSO) [1994] indicated that 62% of the district's predominantly rural population were living in traditional housing and that 57% were using unsafe water sources. These indicators put Buhera amongst the poorest districts in Zimbabwe.

2.2 Household Survey

As part of a survey of Buhera District, 354 households in ten wards were interviewed during the 1994/1995 agricultural season. Figure 1 shows the ward boundaries for Buhera, overlaid with the agroecological zones. For more details about the sampling plan adopted, see Wright [1998].

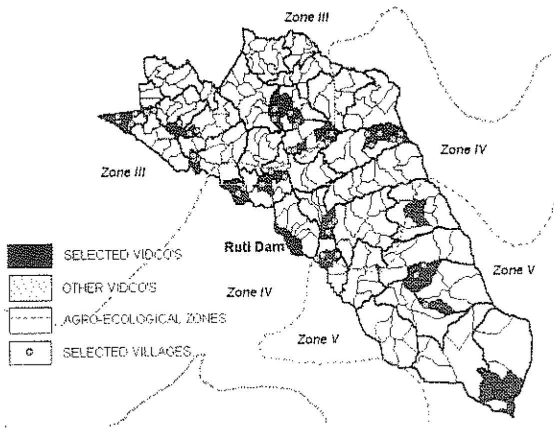

Figure 1 Buhera District, Zimbabwe – household survey

Households in each of the ten wards were interviewed at three points during the year, covering the following subject areas: location, demography, environment, agriculture, food and income/expenditure. All members of the household were measured for height and weight. Data for the sample, including anthropometric data and the socio-economic attributes and resources of the households have been extracted and grouped under appropriate headings.

2.3 Secondary data

Secondary data collection comprised ward and district level data covering administrative boundaries, environmental variables (rainfall, temperature), natural resources (forests, rivers, lakes, terrain), infrastructure (roads, transport availability, markets, potable water supplies), agriculture (crops and livestock) and health statistics (morbidity, mortality, growth monitoring data, healthcare availability). The most recent census is for 1992, with national summaries and more detailed provincial profiles published by CSO [1992]. The provincial profiles disaggregate data to ward level, within each district of the province. Household level data is not made available for reasons of confidentiality.

3. Methods

3.1 Overview

The flow diagram in Figure 2 shows the principal elements of the procedure employed.

3.2 Characterising Wards

Seventeen variables were used characterise all the wards in Buhera. These variables were chosen because they represented the best proxies from secondary data for the more detailed household variables required. The values for each ward are shown in table 1 (see Appendix A). Two wards, numbered 35 and 36, are the urban wards Murambinda and Dorowa. Their characteristics are noticeably different to the remainder of the wards, all of which are rural. These two urban wards were therefore excluded from the matching process and from the first implementation of the simulation model.

3.3 Cluster Analysis

The variables have been used in a cluster analysis to group together the unsurveyed wards with those survey wards 'most closely matching'. This closeness of match is the squared Euclidean distance between wards in n-dimensional space, for the given n characteristics. It is calculated as the sum of the squared differences between the values for each of the n characteristics for the two wards whose closeness is being measured. It is thus important for each variable to have a common scale, so that artificial weights are not introduced when the closeness is calculated. Each of the seventeen variables was therefore standardised prior to clustering. The statistical procedure adopted was k-means clustering, a subset of cluster analysis, which allows the user to specify not only the number of clusters required from the analysis, but also the initial locations of the cluster centres. The locations in seventeen dimensions of each of the 34 wards, unsurveyed and surveyed, are then input to the procedure. The algorithm then matches each ward to the closest cluster centre. At this point each cluster has at least one ward (the surveyed ward used as its centre), and any unsurveyed wards located nearest thereto. The algorithm then recomputes the cluster centres, which will have moved from the initial position specified to the centroid of the locations of all the wards comprising that cluster. It then repeats the matching part of the algorithm and recomputes the cluster centres, which will only change if wards are now closer to a revised cluster centre than to the original position. This iteration continues until a stable cluster set is achieved.
Once a stable cluster set is achieved, the cluster membership is reported.

3.4 Data Extrapolation

The model is populated on a one-for-one basis for each level of the hierarchy: district, agroecological zones, wards and households. Thus, the number of households in each ward, as per the 1992 census, is assumed to be the required number of simulated households. The household database tables are: *Socio-economic* static variables and five tables of dynamic variables *People, Livestock, Crops, Food and consumption* and *Cash.*

Before the data were extrapolated, it was necessary to analyse the household data and compute the mean and standard deviation for each of the 51 variables, for each of the 10 surveyed wards. Independent extrapolation from each variable was not appropriate as the variables are known to show substantial correlation. In addition, therefore, a covariance matrix was created for each surveyed ward. A Fortran computer program was written to extrapolate the required number of households' data, from an input file of descriptors: correlations, standard deviations and means. The extrapolation program produces an output file in text format, which was then converted to the database format used in the simulation model (Microsoft Access).

4. Results

4.1 Matching Surveyed and Unsurveyed Wards

Using cluster analysis to match the surveyed and unsurveyed wards, required that the 34 wards were grouped into ten clusters, one cluster per surveyed ward. To achieve this, the locations of the initial cluster centres were set to correspond to the values of the standardised characteristic variables for each of the surveyed wards. The clustering algorithm then iterates until a stable cluster set is reached. Table 2 shows the ten final clusters and the wards included within each cluster. Each cluster has one surveyed ward and between zero and six additional, unsurveyed wards. Thus, each unsurveyed ward is effectively 'matched' to a surveyed ward by cluster membership. It will be noted in table 2 that the reported clusters appear to be well defined by the agroecological zone variable, despite its exclusion as a characterising variable in the cluster analysis. This lends support to the use of agroecological zone in the sample stratification and the use of these zones as separate hierarchical levels in the simulation model.

The statistical package used by the authors also provides a post-processing ANOVA report of the clusters. This is shown in table 3. It is a one way analysis of variance which shows the between-cluster mean square in column 2 and the within-cluster mean square in column 4. The ratio of these two mean squares is the usual ANOVA F statistic. (The significance levels reported in column 7 should be ignored.) It can be seen from table 3 that the characteristics of DEPENDCY, SEXRATIO and SANITIND differ most across the ten clusters, with small differences for PRWATER, AGEHH, ACCWATER and SIZEHHLD. The other ten characteristics show intermediate levels of difference across clusters.

Cluster		Surveyed	Unsurveyed						TOTAL
1	Wards	2	16						2
	AgZones	*3*	*3*						*1*
2	Wards	5	3	11					3
	AgZones	*3*	*3*	*3*					*1*
3	Wards	15	4	7	9	12			5
	AgZones	*3*	*3*	*3*	*3*	*3*			*1*
4	Wards	17	13	14	21	22	23	25	7
	AgZones	*3*	*4*	*4*	*4*	*4*	*4*	*4*	*2*
5	Wards	18	1	19	29				4
	AgZones	*4*	*3*	*4*	*4*				*2*
6	Wards	20	6	8	10				4
	AgZones	*4*	*3*	*3*	*3*				*2*
7	Wards	24	26	31					3
	AgZones	*4*	*4*	*5*					*2*
8	Wards	27							1
	AgZones	*5*							*1*
9	Wards	32	28	30	33				4
	AgZones	*5*	*5*	*5*	*5*				*1*
10	Wards	34							1
	AgZones	*4*							*1*

Table 2. Cluster Membership

4.2 Extrapolating Household Data

The extrapolation program was run for each of the 34 wards, using the appropriate source data file i.e. the three descriptors for each variable in the household sample data of the matched surveyed ward.

The results reported herein have assumed that ward characteristics are normally distributed and are unweighted by the sample frame used. It seems clear that neither of these assumptions are likely to be valid, although the implications for changing the weights used in the cluster analysis are currently being investigated. Possibly of greater concern is the assumption of normality of the household survey data used in the extrapolation. Some of these data are known to exhibit marked skewness and the categorical and boolean variables are obviously not normally distributed. Possible procedures such as transformation of the distributions prior to extrapolation could be used to mitigate the problem. In particular, for the boolean variables the following method may offer an improvement: (1) Tabulate all possible combinations of 0/1 for the boolean variables and from the data calculate the probability of their occurring for each ward; (2) For each combination, calculate the parameters of the multivariate normal distribution for the other variables for each ward (this is the distribution of all the other variables conditional on a given combination of the binary variables); (3) Simulate by generating a set of values for the binary variables using the relevant probabilities, then generate the other variables from the relevant conditional distribution. This procedure could easily be extended to variables with more than two categories and is currently being evaluated, together with various transformation routines for the non-boolean, non-categoric data.

5. DISCUSSION AND CONCLUSIONS

Where government resources are constrained, developing countries' statistical services are often less robust than those of the OECD countries. It is suggested that the lack of availability of localised household data, as experienced during this research in Zimbabwe, is likely to be repeated in many countries in Africa and other disadvantaged regions. Socio-economic, environmental and agricultural data are often restricted either to small field surveys with different sample frames or to secondary data that has been aggregated at levels above that which is being studied.

Variable	Cluster		Error		F	Sig.
	Mean Sq	df	Mean Sq	df		
ACCHLTH	2.341	9	.497	24	4.709	.001
ACCWATR	2.017	9	.619	24	3.262	.010
AGEHH	1.955	9	.642	24	3.047	.014
AGINC	2.461	9	.452	24	5.441	.000
DEPENDCY	3.009	9	.247	24	12.207	.000
DISTMKT	2.571	9	.411	24	6.259	.000
FEMALEHH	2.668	9	.375	24	7.123	.000
FUELWOOD	2.694	9	.365	24	7.388	.000
HOUSGIND	2.767	9	.338	24	8.196	.000
MAIZEPPN	2.798	9	.326	24	8.584	.000
NONAGINC	2.432	9	.463	24	5.256	.001
POPNDENS	2.549	9	.419	24	6.085	.000
PRWATER	1.682	9	.744	24	2.260	.054
SANITIND	2.898	9	.288	24	10.049	.000
SEPRMS	2.490	9	.441	24	5.644	.000
SEXRATIO	2.960	9	.265	24	11.178	.000
SIZEHHLD	2.035	9	.612	24	3.326	.009

Table 3. Showing ANOVA for Cluster Analysis

Simulation modelling has historically used aggregate data or small representative samples to enable results to be produced within realistic processing times. With the increasing power of personal computers, the data used in simulations can now be increased and modelling low level processes directly can incorporate the random effects seen in complex systems. This can be achieved by several methods. The authors have collaborated to produce a simulation model that uses expert rules to execute behaviour at various levels in the socio-economic hierarchy. Other authors report the development of agent based modelling approaches, the most well known of which, in academic areas, is Swarm – see Minar, Burkhart, Langton, & Askenazi [1996].

The procedure outlined herein offers a systematic solution to this difficulty by using secondary data to show how a small number of surveyed areas can be related to unsurveyed areas by a process of modified cluster analysis. From this matching process straightforward statistical techniques can be employed to extrapolate the sample data for the matched, unsurveyed wards.

6. ACKNOWLEDGEMENTS

This research was funded by the Commission of the European Communities DGXII, Science and Technology for Developing Countries Programme, contract reference TS3*-CT92-0048.

7. REFERENCES

CSO, Census 1992 - Zimbabwe Preliminary Report, Central Statistics Office, Harare, Zimbabwe, 1992.

CSO, Census 1992 - Manicaland Provincial Profile, Central Statistics Office, Harare, Zimbabwe, 1994.

Gundry, S. W., Wright, J. A., & Ferro-Luzzi, A. Simulating the food and nutrition system in rural Zimbabwe to support targetting of emergency aid, paper presented at the ModSim 97, the International Congress on Modelling and Simulation, Hobart, Tasmania, 8-11 December 1997.

Hay, R., Nutritional defects and public expenditure priorities. Food Policy, 19(1), 5-8, 1994.

Maxwell, S., & Frankenberger, T. R., Household food security: concepts, indicators, measurement - a technical review, UNICEF and IFAD, New York, 1992.

Minar, N., Burkhart, R., Langton, C., & Askenazi, M., The SWARM simulation system: a toolkit for building multi-agent simulations, Working paper 96-06-042, Santa Fe Institute, New Mexico, 1996.

Tomkins, A., Watson, F., & Scrimshaw, N. S., Malnutrition and Infection: A Review, State-of-the-Art series, number 5, United Nations Administrative Committee on Coordination - Subcommittee on Nutrition (ACC/SCN), London, 1989.

Wright, J. A., Information Systems for Child Nutrition in Zimbabwe, Ph.D. thesis, University of Edinburgh, 1998.

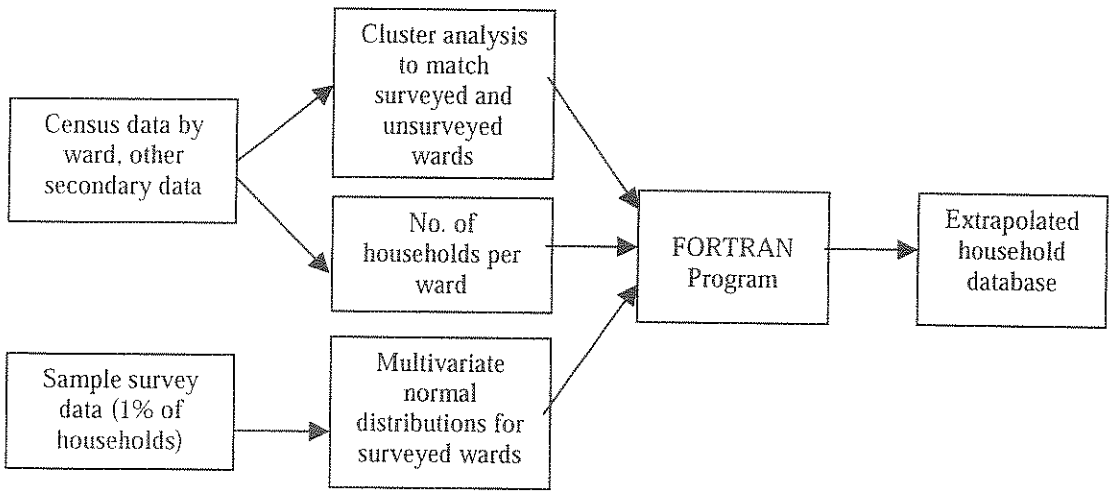

Figure 2. Outline flow diagram of method

Appendix A. Table 1 showing the 17 characteristic variables of 36 wards in Buhera District, Zimbabwe

Ward No	Ward Name	In field survey?	AgZoneID	ACCHLTH	ACCWATER	AGEHH	AGINC	DEPENDCY	DISTMKT	FEMALEHH	FUELWOOD	HOUSGIND	MAIZEPPN	NONAGINC	POPNDENS	PRWATER	SANITIND	SEPRMS	SEXRATIO	SIZEHHLD
1	Mudzamiri	No	3	-5.09	631	46.7	0.810	0.431	5.79	0.502	0.988	0.767	0.957	0.055	41.50	0.587	0.830	0.042	1.44	4.95
2	Chapwanya	Yes	3	6.97	666	46.1	0.837	0.433	8.64	0.505	0.998	0.851	0.907	0.059	57.36	0.323	0.868	0.042	1.42	5.21
3	Garamwera	No	3	10.64	480	48.2	0.689	0.450	5.63	0.484	0.969	0.743	1.000	0.078	38.25	0.591	0.658	0.033	1.25	4.85
4	Nerutanga	No	3	13.38	734	45.7	0.846	0.418	5.20	0.496	0.997	0.740	0.933	0.040	43.28	0.235	0.727	0.038	1.34	5.12
5	Marume	Yes	3	18.67	544	43.4	0.689	0.494	7.00	0.453	0.961	0.584	0.975	0.097	46.04	0.513	0.611	0.067	1.14	4.72
6	Makumbe	No	3	14.25	708	47.5	0.679	0.459	5.18	0.492	0.969	0.748	0.953	0.067	41.48	0.445	0.840	0.028	1.20	5.28
7	Munyira	No	3	7.01	799	44.7	0.744	0.459	4.89	0.521	0.984	0.694	0.905	0.073	52.16	0.590	0.731	0.031	1.33	5.39
8	Mombeyarara A	No	3	21.94	578	45.0	0.762	0.436	3.92	0.517	0.979	0.760	0.724	0.075	44.92	0.729	0.746	0.036	1.31	5.13
9	Mombeyarara B	No	3	21.33	587	47.2	0.829	0.418	4.25	0.530	0.996	0.811	0.776	0.050	55.06	0.586	0.766	0.025	1.40	5.02
10	Chitsunge	No	3	5.61	634	45.0	0.785	0.422	4.88	0.480	0.991	0.759	0.881	0.067	39.81	0.483	0.830	0.021	1.32	4.70
11	Chimombe A	No	3	1.75	525	44.9	0.588	0.444	4.59	0.495	0.941	0.562	0.893	0.100	64.90	0.527	0.598	0.073	1.41	5.14
12	Chimombe B	No	3	6.96	742	45.9	0.580	0.435	4.92	0.507	0.975	0.661	0.750	0.114	52.50	0.553	0.688	0.040	1.36	5.08
13	Chimombe 'B	No	4	6.44	697	44.7	0.705	0.393	6.79	0.569	0.991	0.787	0.690	0.056	35.94	0.691	0.887	0.032	1.55	5.29
14	Chiweshe	No	4	4.19	802	43.3	0.709	0.394	5.66	0.549	0.990	0.764	0.373	0.051	33.27	0.530	0.796	0.027	1.50	5.16
15	Neshava	Yes	3	12.15	814	44.2	0.783	0.427	4.58	0.531	0.994	0.776	0.654	0.054	49.93	0.420	0.859	0.022	1.32	5.15
16	Murambinda	No	3	21.26	536	44.1	0.824	0.418	5.64	0.505	0.994	0.797	0.875	0.059	60.50	0.388	0.825	0.021	1.42	5.44
17	Mudinzwa A	Yes	3	4.61	754	44.2	0.823	0.388	7.21	0.558	0.995	0.839	0.492	0.044	39.59	0.622	0.848	0.026	1.53	5.22
18	Nechavava	Yes	4	3.78	592	44.4	0.775	0.400	8.17	0.533	0.994	0.808	0.656	0.056	38.95	0.566	0.913	0.025	1.45	5.18
19	Mudinzwa B	No	4	3.11	702	44.2	0.728	0.415	5.15	0.516	0.994	0.767	0.620	0.057	44.60	0.594	0.794	0.029	1.40	5.37
20	Mabvuregudo	Yes	4	11.67	711	45.7	0.876	0.414	4.92	0.493	0.989	0.824	0.691	0.038	30.39	0.612	0.853	0.024	1.36	5.23
21	Chikuwa	No	4	8.85	840	44.6	0.791	0.394	5.13	0.535	0.988	0.820	0.738	0.053	43.48	0.503	0.869	0.019	1.45	5.25
22	Makuvise	No	4	-1.48	654	43.7	0.831	0.390	5.04	0.530	0.984	0.837	0.140	0.037	49.23	0.657	0.892	0.021	1.51	5.58
23	Betera	No	4	0.61	661	43.4	0.788	0.388	6.55	0.533	0.990	0.780	0.529	0.049	23.12	0.603	0.850	0.029	1.59	4.96
24	Murwira	Yes	4	3.90	831	44.1	0.545	0.384	9.95	0.543	0.995	0.839	0.465	0.055	38.03	0.646	0.847	0.019	1.61	5.30
25	Mawire	No	4	-2.45	702	43.6	0.589	0.394	4.17	0.507	0.984	0.825	0.208	0.040	41.32	0.548	0.843	0.016	1.46	5.31
26	Chirozva A	No	4	-2.12	782	43.0	0.199	0.384	6.57	0.513	0.991	0.849	0.220	0.024	33.58	0.538	0.910	0.027	1.58	5.35
27	Chirozva B	Yes	5	-8.58	593	43.4	0.507	0.373	4.80	0.568	0.991	0.815	0.244	0.044	34.27	0.297	0.803	0.011	1.70	5.59
28	Mushumba East	No	5	-3.45	644	43.0	0.810	0.376	8.24	0.561	0.996	0.871	0.145	0.040	30.11	0.379	0.878	0.019	1.58	5.47
29	Mushumba West	No	4	0.33	521	43.1	0.722	0.381	9.73	0.528	0.988	0.818	0.198	0.049	43.73	0.545	0.825	0.016	1.61	5.19
30	Chimombe West	No	5	-0.02	885	42.8	0.829	0.377	10.23	0.542	0.991	0.873	0.140	0.040	30.36	0.747	0.823	0.020	1.42	5.44
31	Mutiusinazita	No	5	5.24	607	43.1	0.552	0.389	11.81	0.563	0.986	0.862	0.011	0.036	30.82	0.879	0.823	0.022	1.56	5.47
32	Chimombe East	Yes	5	-6.81	667	44.3	0.760	0.383	9.33	0.532	0.996	0.921	0.211	0.043	21.88	0.582	0.932	0.024	1.61	5.34
33	Chimombe West	No	5	1.90	842	42.6	0.830	0.389	13.07	0.565	0.995	0.876	0.140	0.038	25.90	0.719	0.890	0.018	1.49	5.43
34	Chimutsa East	Yes	5	-0.46	836	41.4	0.547	0.466	10.31	0.472	0.973	0.717	0.082	0.115	30.64	0.702	0.661	0.062	1.14	5.54
35	Dorowa	*Exclude*	*3*	*3.28*	*30*	*35.2*	*0.107*	*0.608*	*0.58*	*0.102*	*0.009*	*0.000*	*0.000*	*0.335*	*850.00*	*1.000*	*0.002*	*0.049*	*0.80*	*3.74*
36	Murambinda (u)	*Exclude*	*3*	*716.63*	*102*	*33.7*	*0.157*	*0.665*	*1.52*	*0.332*	*0.539*	*0.031*	*0.000*	*0.366*	*608.40*	*0.971*	*0.175*	*0.102*	*1.31*	*3.94*
		Mean		5.47	685	44.44	0.717	0.412	6.70	0.521	0.986	0.787	0.564	0.057	40.79	0.557	0.809	0.030	1.43	5.23
		Std Dev		7.93	108	1.50	0.138	0.030	2.40	0.029	0.012	0.078	0.323	0.022	10.40	0.135	0.086	0.014	0.14	0.22

NB It should be noted that some duplication exists in the ward names shown above, which are taken from the published census document CSO [1994].