The applicability of the South African Census 2011 data for evidence-based urban planning

Sibusisiwe Khuluse-Makhanya (Corresponding author)

Statistician

Researcher at the CSIR & PhD student at the University of Twente

Postal address: CSIR Built Environment, PO Box 395, Pretoria, 0001, South Africa

E-mail: SMakhanya@csir.co.za

Nontembeko Dudeni-Tlhone

Statistician

Researcher at the CSIR & PhD student at the University of KwaZulu-Natal

Postal address: CSIR Built Environment, PO Box 395, Pretoria, 0001, South Africa

E-mail: NDudeni-Tlhone@csir.co.za

Jennifer Holloway

Statistician

Senior Researcher at the CSIR

Postal address: CSIR Built Environment, PO Box 395, Pretoria, 0001, South Africa

E-mail: JHollowa@csir.co.za

Dr Peter Schmitz

GIS professional

Principal Researcher at the CSIR

Postal address: CSIR Built Environment, PO Box 395, Pretoria, 0001, South Africa

E-mail: PSchmitz@csir.co.za

Dr Louis Waldeck

Civil Engineer

Principal Researcher at the CSIR

Postal address: CSIR Built Environment, PO Box 395, Pretoria, 0001, South Africa

E-mail: LWaldeck@csir.co.za

Prof. Alfred Stein

Professor of Spatial Statistics and Image Analysis

Postal address: Faculty of Geo-Information and Earth Sciences (ITC), University of Twente, PO Box 217, 7500 AE Enschede, Netherlands

E-mail: a.stein@utwente.nl

Prof. Pravesh Debba

Statistician

Competence Area Manager at the CSIR & Visiting Professor at the School of Statistics and Actuarial Sciences, University of the Witwatersrand

Postal address: CSIR Built Environment, PO Box 395, Pretoria, 0001, South Africa

E-mail: PDebba@csir.co.za

Theo Stylianides

Operations research

Principal Researcher at the CSIR

Postal address: CSIR Built Environment, PO Box 395, Pretoria, 0001, South Africa

E-mail: TStylian@csir.co.za

Pierre du Plessis

GIS professional

Senior Researcher at the CSIR

Postal address: CSIR Built Environment, PO Box 395, Pretoria, 0001, South Africa

E-mail: PduPlessis@csir.co.za

Antony Cooper

GIS professional

Principal Researcher at the CSIR

Postal address: CSIR Built Environment, PO Box 395, Pretoria, 0001, South Africa

E-mail: ACooper@csir.co.za

Ethel Baloyi

Trainee GIS professional

Intern at the CSIR

Postal address: CSIR Built Environment, PO Box 395, Pretoria, 0001, South Africa

E-mail: EBaloyi1@csir.co.za

ABSTRACT

In urban planning, it is important to understand settlements in terms of demographics, socio-economic and location characteristics. A key dataset used for such studies is a national census small area dataset. This paper provides three cases studies using the 2011 South African census small area data to highlight the value added as well as challenges identified in using these data. The data were used to study patterns in housing conditions for urban growth simulation, to determine new school locations in feeder catchments, and to aid air quality mapping. The methods undertaken for these case studies were an adaptation of *k*-means cluster analysis for distinguishing housing and household patterns in order to identify homogenous areas with similar demands for infrastructure and services; catchment analysis using Flowmap to determine new school locations; and Kriging with external drift to map an indicator of air quality. In all case studies, the census small area data were critical and the information derived is deemed useful for various urban planning decisions, including planning for the provision of essential services, planning for air pollution control at locations where regulatory

levels of pollutants are exceeded frequently, and planning for the development of social infrastructure in areas that are under-serviced.

Key words:

Urban planning, spatial data quality, multi-stage cluster analysis, air quality mapping and urban simulation.

1. Introduction

In urban planning, it is important to understand settlements in terms of demographic, socio-economic, physical and political environmental characteristics. The objective is often to support municipalities, regional and national governments with long-term planning tools for the development of infrastructure, facilities and services. In a municipal context, planning relies on spatially explicit estimates of the future demand for services, which depend largely on where households will live, where they will work and how they will commute using the transportation networks available to them (Waddell, 2005), as depicted in Figure 1.

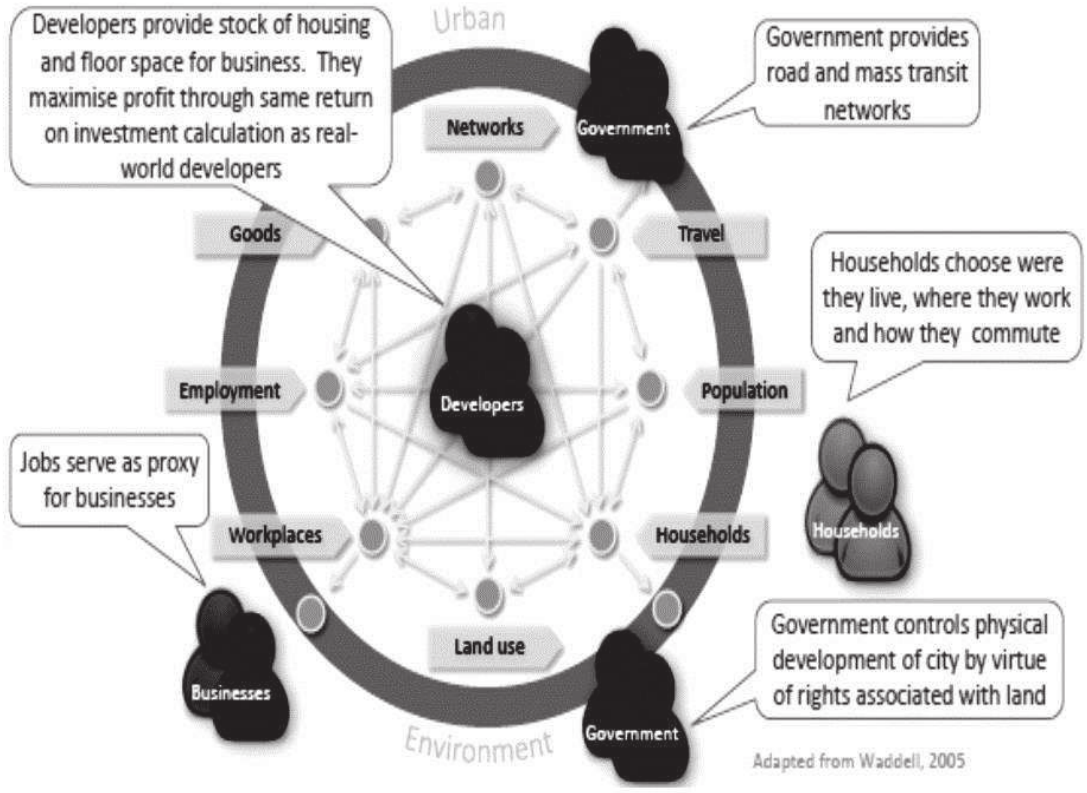

Figure 1: An illustration showing main actors and interactions within the urban environment. Source: Waldeck (2013)

Given the complexity of large cities as a system, modelling and simulation tools are valuable for municipal decisions regarding the likes of environmental quality and safety in settlements and stimulation of economic productivity through provision of infrastructure. Models cannot capture such a complex system entirely; however, they do allow integration of data indicative of factors shown in Figure 1 and

their dynamics simplistically, for consumption by the relevant stakeholders. That is, they provide evidence upon which spatially equitable and sustainable planning of urban areas can be based (Cilliers et al., 2014).

A crucial dataset for evidence-based planning is a national census. With the release of the South African Census 2011 small area data in 2012 (Statistics South Africa, 2012a & 2012b), this paper reflects on research undertaken where an important input was this dataset. Specifically, we discuss how the census data were used in urban growth simulation, which is relevant for municipal services demand planning, catchment analyses to support municipal infrastructure planning, and urban air quality mapping to support planning for municipal and regional air pollution control. Given the intention to highlight the value added by the Census 2011 data for different case studies, an in-depth exposition of the different methods and results is not pursued in this paper. References for each case study are given in the appropriate sections. The paper proceeds with the materials section, Section 2, which introduces each case study and the variables that were selected from the census dataset. A summary of methods

undertaken for each study is presented in Section 3, with findings following in Section 4. A discussion of what was achieved and challenges experienced with the Census 2011 data ensues in Section 5, and conclusions are given in Section 6.

2. Materials

The background of each case study is given here, with details of the applicable census data. The three case studies use the census small area data and are of the metropolitan areas of the Gauteng province. The urban growth simulation study focused on the Ekurhuleni Metropolitan Municipality due to the increasing growth and development observed and projected for this area. The Tshwane Metropolitan Municipality was of interest in the catchment analysis, given continued settlement growth in this city, especially of young families requiring education facilities. In mapping air quality, the whole Gauteng province was considered to provide a sufficient sample size for characterising the spatial distribution of the chosen air quality statistic. Generally, municipalities have less than eight air quality monitoring stations each, so we

needed to pool stations over a wider region to make statistical mapping possible. From a regional air quality map, results for Ekurhuleni were extracted and discussed further with reference to housing types and location classes, determined from the household classification activity for urban simulations.

2.1 Urban growth simulations and household characteristics data

UrbanSim is a numerical modelling and simulation platform used at the Council for Scientific and Industrial Research (CSIR) to plan for urban growth (30-year planning horizon), with focus on demand for services within municipalities (Borning et al., 2008). Changes in demand for municipal services have implications for policy and infrastructure investment decisions. UrbanSim was created for and used in developed countries, and hence it was necessary to adapt it for use in South Africa, given the lack of spatial equity within our urban areas.

UrbanSim is an agent-based model where choices made by agents (households, property developers, businesses and government) in the urban system as shown in Figure 1 are maximised, based on the associated utility. For instance, the choice of a household to buy or rent a specific property depends upon household income, place of work and the presence of school-age children. Such demographic and socio-economic information is the basis upon which property demand can be evaluated. Expected or even assumed changes in demographic profiles, as translated from demographic and economic projections and policy changes where applicable, are then used to simulate growth in residential property demand. This demand output provides insight into which settlements are expected to grow significantly and therefore need attention regarding municipal services provision.

Historically, the demographic input for UrbanSim was the sub-place level ClusterPlus geo-demographic clusters dataset bought from a company called Knowledge Factory. Following discontinuation of ClusterPlus, the CSIR developed a method for classifying households at a small spatial scale using Census 2011 data. The small area layer (SAL) was used to

determine homogenous groups of households that are expected to have similar demands for infrastructure and services.

Table 2: Final list of variables used in the household segmentation analysis

Broad variable classes (Factors)	**Key variables**
Dwelling location characteristics and density	Enumeration area type Density of dwellings within small areas
Dwelling type and conditions	Type of dwelling Household size Number of rooms
Socio-economic	Weighted average annual household income Employment status of head of household Highest education level of the head of household
Life cycle stage and household (family) structure	Marital status Age group Relationship structure within households
Demographics	Population group Gender of household head Gender of persons in households Property ownership of households

For the Ekurhuleni Metropolitan Municipality, 4,610 small areas from Census 2011 were considered. Key variables were identified and grouped as factors associated with socio-

economic conditions, and demographic and housing characteristics, as shown in Table 1. All variables were categorical, with the exception of household income and density of dwellings. Further refinement of the categories using other government publications was performed to simplify the analysis and ensure the results fit the context. Although the data appear to be about either the households or the members thereof, they were aggregated and summarised at the level of small areas and not individual households.

2.2 Census school-age population data as a proxy for demand for schools

Infrastructure is necessary for stimulating economic growth and improving the quality of life of citizens by enabling the provision of municipal services (CSIR, 2011). Catchment analysis in urban planning is used to assess spatially whether current infrastructure is sufficient for associated services. Sufficiency is analysed in terms of the number and accessibility of service points (Green and Argue, 2012). Therefore, population information plays a critical role in such analyses,

together with transportation networks and facilities inventories. Flowmap is specialised software enabling catchment analysis through an optimal assessment of flows of goods, services and people, using standards and other constraints to delineate areas that are under- or over-supplied (De Jong and Van der Vaart, 2013).

The Tshwane Metropolitan Municipality lies north of Ekurhuleni and it is currently experiencing settlement growth in various locations, especially in the northern, eastern and south-eastern areas. The residents in these areas are young working-class families who require educational institutions for their children. For public schools, the Department of Basic Education has standards concerning the number of learners per class and the size of the catchment that the school may service. These standards were set to ensure better quality of education through small learner-to-teacher ratios and better well-being of learners by minimising travel times to school. In this context, the research question was, which locations within Tshwane needed to be prioritised for investment into new schools? In response, a case study assessed the optimal location of new schools within Tshwane. The focus here is on

the value added by the Census 2011 small area data. Reference is made to the impact that a faulty inventory of schools had on the location-allocation decisions.

The feeder catchments for schools within Tshwane were determined using Census 2011 data for 4,524 small areas (Schmitz and Eksteen, 2014). The population variable of interest was the number of school-age individuals per small area, namely those aged between 6 and 18 years. This group was divided into the pre-school age group of less than 6 years old, the primary school age group of between 6 and 12 years old, and the secondary school age group of between 13 and 18 years old. Land-use data from GeoTerra Image was also used.

In processing the population small area data, the municipal area was tessellated in Flowmap, using hexagons of 350 m per side. These were smaller than the smallest small area polygon. To redistribute the number of learners from the Census 2011 small areas into hexagons, land-use and the school-age population data were used to re-allocate proportionately the counts of learners into the hexagons

(Schmitz and Eksteen, 2014), resulting in a finer scale spatial distribution of the learners. This is shown in Figure 2, where higher concentrations in school-age children are observed along the north-west to south-east directions from the CBD.

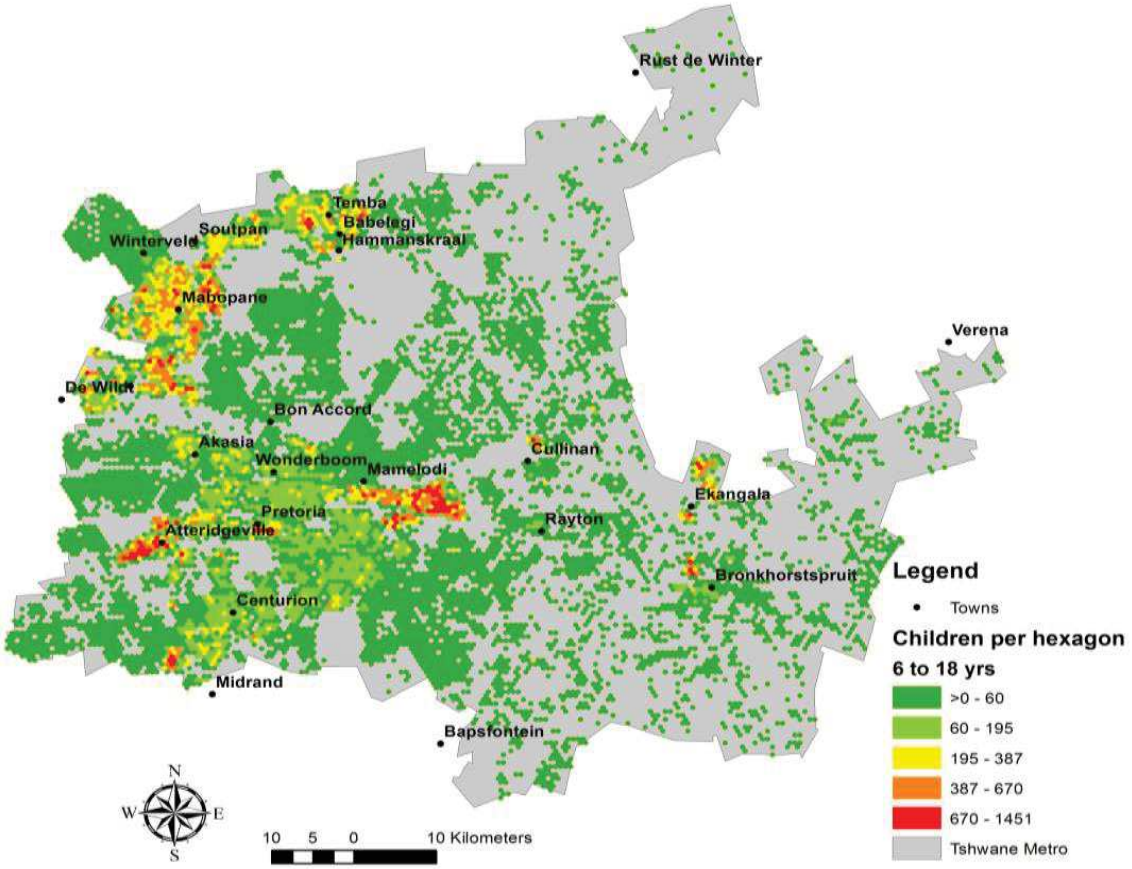

Figure 2: Spatial distribution (using hexagon tessellation) of the school-age population (children aged 6–18 years) from the SA Census 2011 data for the City of Tshwane (Schmitz and Eksteen, 2014)

2.3 Urban air quality mapping with proxy variables for domestic particulate matter emissions from the census dataset

A study was undertaken to map how often the South African air quality standard for PM_{10} was exceeded in the Highveld region, between September 2009 and August 2012. The study region included parts of Mpumalanga bordering Gauteng, to have an adequate sample size (number of air quality monitoring stations) to develop valid statistical models mapping PM_{10} exceedances. Key anthropogenic sources of PM_{10} in the study area include industries, domestic combustion of alternative energy fuels and vehicles, and dust from unpaved roads and mine dumps. Domestic combustion of alternative energy fuels has been related to heavy haze events, mainly in less affluent areas (informal settlements) of the study area (Piketh et al., 2004; Norman et al., 2007; Wright et al., 2011). Therefore, in mapping PM_{10} concentrations for this region, we considered household information such as dwelling type and energy use, obtained from the Census 2011 small area data.

A total of 24,584 small areas covering parts of Gauteng and Mpumalanga were considered. From the output, results for Ekurhuleni were extracted and are discussed in Section 4. Census variables of interest, shown for Gauteng in Figure 3, were the percentage of households residing in informal dwellings (Figure 3(a)) and the percentage of households using alternative energy sources for heating (Figure 3(b)). These were used as explanatory variables in mapping the annual exceedance rate of the RSA PM_{10} standard. The total number of dwellings per small area (a proxy for dwelling density) and alternative energy used for cooking were the other two variables considered.

White spaces in Figures 3(a) and 3(b) are non-residential small areas, typically industrial areas including mines and quarries. Informal settlements (those with more than 80% of the small area being informal dwellings) are seen in Figure 3(a) as slivers of red throughout the province, typically on the periphery of industrial or mining areas. On the southern boundary of Ekurhuleni, a moderate to high proportion of informality is observed, attributable to a mixture of informal settlements and formal residences (townships such as

Vosloorus, Thokoza, Katlehong and Tsakane) with backyard shack dwellings. Figure 3(b) shows that household use of biomasss for heating is more prevalent in the same areas where high proportions of informality were observed.

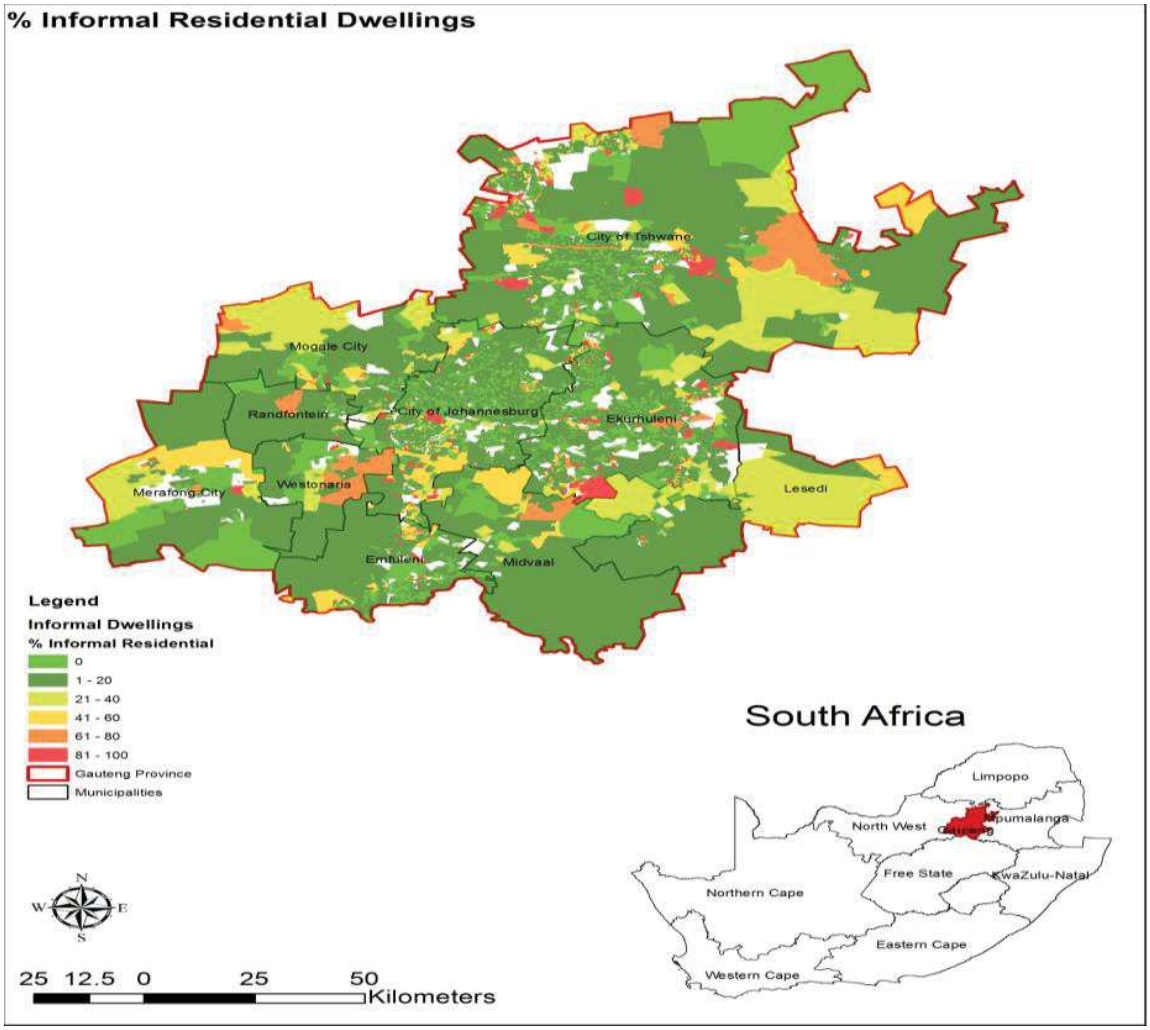

Figure 3(a): Percentage of dwellings in each small area categorised as informal from the SA Census 2011 small area layer for Gauteng. These data were used as covariate information in mapping PM_{10} exceedance rates

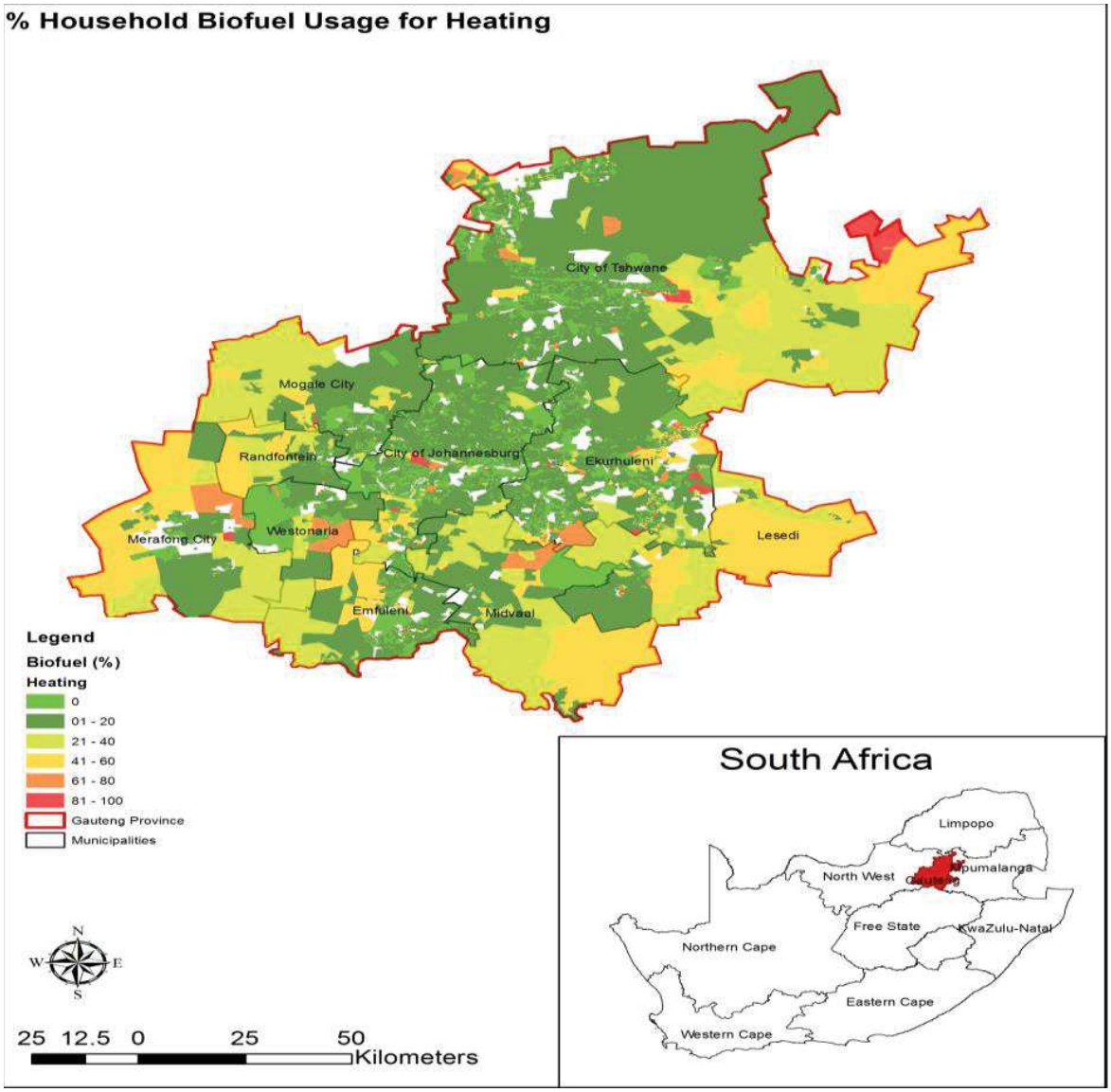

Figure 3(b): Percentage of dwellings in each small area categorised as using biomass for their heating needs. These data extracted from the SA Census 2011 small area layer for Gauteng were used as covariate information in mapping PM_{10} exceedance rates

3. Methods

In the previous section, the contexts and variables of interest for each case study were discussed. In this section, key methods related to using census data are given for each case study, but for more details, see Dudeni-Tlhone et al. (2013); Khuluse and Stein (2013); and Schmitz and Eksteen (2014).

3.1 Household classification for urban growth simulation

Modelling urban growth in UrbanSim requires as input classified household information for the municipal area of interest, as explained in Section 2.1. To make these typologies readily available for any South African municipal area that may be of interest for urban simulations, a generic method for classifying households using census small area data had to be developed. Internationally, geo-demographic classification for public policy-related applications is an active area of research, with clustering techniques having been applied successfully to small area census data to identify homogenous groups of households (Vickers and Rees, 2007; Adnan et al., 2010; Ojo et

al., 2013; Dudeni-Tlhone et al., 2013). The most popular technique used is the *k*-means algorithm because of its efficiency in minimising the intra-cluster variances and efficiency in processing large numeric data, particularly with many clusters (MacQueen et al., 1967; Huang, 1998; Jain et al., 1999; Adnan et al., 2010). The *k*-means algorithm partitions *n*-dimensional datasets into k clusters, where k is the number of cluster centroids $m_j,\ \ j = 1,2, \dots, k$. The objective is to minimise the average squared distance $\left(x_i^{(j)} - m_j\right)^2$ between observations and their cluster centroids. Therefore, the objective function is a measure of how well the centroids represent observations in the relevant clusters, and is given by:

$$O = \sum_{j=1}^{k} \sum_{i=1}^{n_j} \left(x_i^{(j)} - m_j\right)^2$$

Running the *k*-means algorithm on the observations as described, with all variables considered at once, can be defined as a single-stage approach. The multi-stage *k*-means clustering approach can be defined as grouping observations by considering subsets of variables sequentially, until the

desired number of clusters is obtained and all variables have been considered. The multi-stage approach was developed to minimise the incidence of outliers, which is a common problem with single-stage *k*-means clustering.

A multi-stage *k*-means clustering method was implemented in Ekurhuleni to detect groups of small areas exhibiting similar characteristics concerning housing, demographics and socio-economic conditions. Two iterations were performed. The first iteration was to identify problems with the data and to gain insight on how best to set up the various stages of the algorithm. Firstly, the dwelling density variable was removed and the first stage of the clustering was done purely on the enumeration area (EA) type variable. An EA is synonymous with a census tract and the EA type encodes the dominant land-use within the EA. There were cases where small areas belonged to more than one EA type (mixed) and some that had no EA type assigned. These atypical small areas formed only 2% of the total, but non-homogeneity in clusters was observed where these atypical small areas were placed. To overcome this, atypical small areas were reallocated into

existing land-use categories, using an independent building-based land-use dataset.

In the final set-up of the algorithm, the initial stage involved implementing the *k*-means clustering technique on all small areas, separately for each of the five factors given in Table 1. Each factor is a group of census variables relevant for the household characteristic the factor represents. Therefore, the initial stage resulted in five groups of clusters, each group corresponding to a factor as shown in Figure 4. The subsequent stages consisted of consolidating the groups of clusters from the initial stage, by applying the *k*-means algorithm on clusters from the different groups, sequentially until all the groups had been considered and a final set of clusters was obtained. To avoid outlying clusters in this final set-up, consolidation of a subset of clusters within each group rather than the whole group was performed.

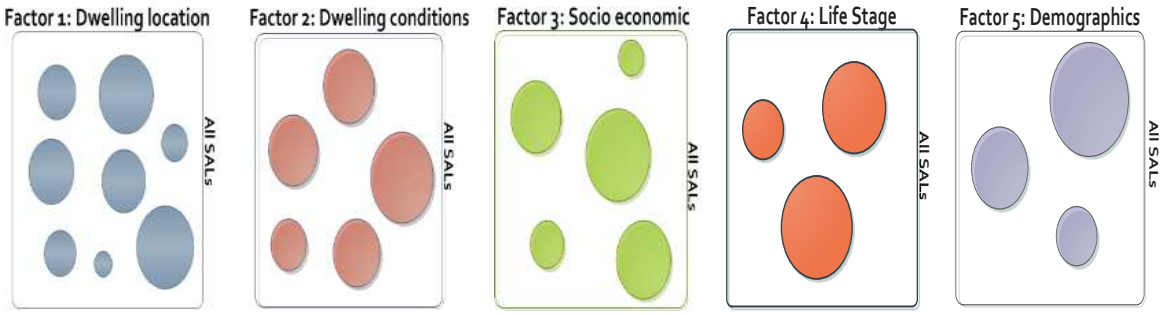

Figure 4: Illustrating the formation of k-means clusters for each factor in the first stage of our initial multi-stage approach. These clusters are then subjected to further clustering, where the clusters from the different factors are consolidated sequentially

The order of consolidation, shown in Figure 4, was based on the importance associated with each factor in differentiating between households and land-use types. The small areas were first split according to the clusters of the "Dwelling location" factor, which was defined as the first stage group. One characteristic of *k*-means clustering is its sensitivity to outliers, and this can be used in clustering out the outliers (Yoon et al., 2007). This characteristic was exploited to remove the small groups from clustering that differed from the predominantly residential population, such as clusters dominated by farms. In the second stage of the clustering, the

residential clusters (low-density formal, high-density formal and informal) were split further by the clusters from the dwelling type and conditions factor. This helped separate the small areas further according to dwelling type, dwelling size and household size. The resulting clusters, excepting small clusters with fewer than ten small areas, were then clustered with those corresponding to the "socio-economic" factor. No further clustering was done because the lack of data on family composition per household made it difficult to describe life-stage and family structures within small areas. Therefore the remaining two factors, life-stage and demographics, were used to analyse descriptively each cluster that resulted from the integration of clusters formed when the dwelling location, dwelling condition and socio-economic condition factors were considered. This post-processing of clusters helped identify whether there were large intra-cluster variations necessitating further separation.

3.2 Delineating catchments for schools in Flowmap

Flowmap is used to derive optimal locations, based on distance and accessibility from the target population. For

schools, the school-age population determines the demand. For Tshwane, the school-age population was derived at a finer spatial scale using the Census 2011 data, as discussed in Section 2.2. For Flowmap to assess flows between demand and target locations, a roads dataset from the municipality was used. Two different school inventories were considered for comparison.

Two iterations of catchment analysis using Flowmap were performed; the first using an unverified school inventory (containing inaccurate locations) and the second using a verified inventory. In Flowmap, schools were the destinations and the residential areas of the school-age population were the origins. School capacities and a maximum travel distance of 5 km were criteria used to assign learners to schools. If more than one school was located in the same hexagon, the capacity was computed as the sum of each school's capacity (Schmitz and Eksteen, 2014).

Figure 5 shows the resulting catchments for north-western Tshwane, based on the number of learners that can be accommodated by each school and/or are within 5 km of

the school, when the verified inventory was used. The grey areas in Figure 5 were not allocated to a school and were subsequently used by Flowmap to determine potential locations of new schools to accommodate the population in these areas. Finally, twenty new school allocations (10 primary and 10 high schools), obtained using the verified schools inventory, were compared to those obtained using the unverified inventory. The differences were used to deriving an estimate of the financial impact of poor quality spatial data by considering the cost of building a new school (Schmitz and Eksteen, 2014).

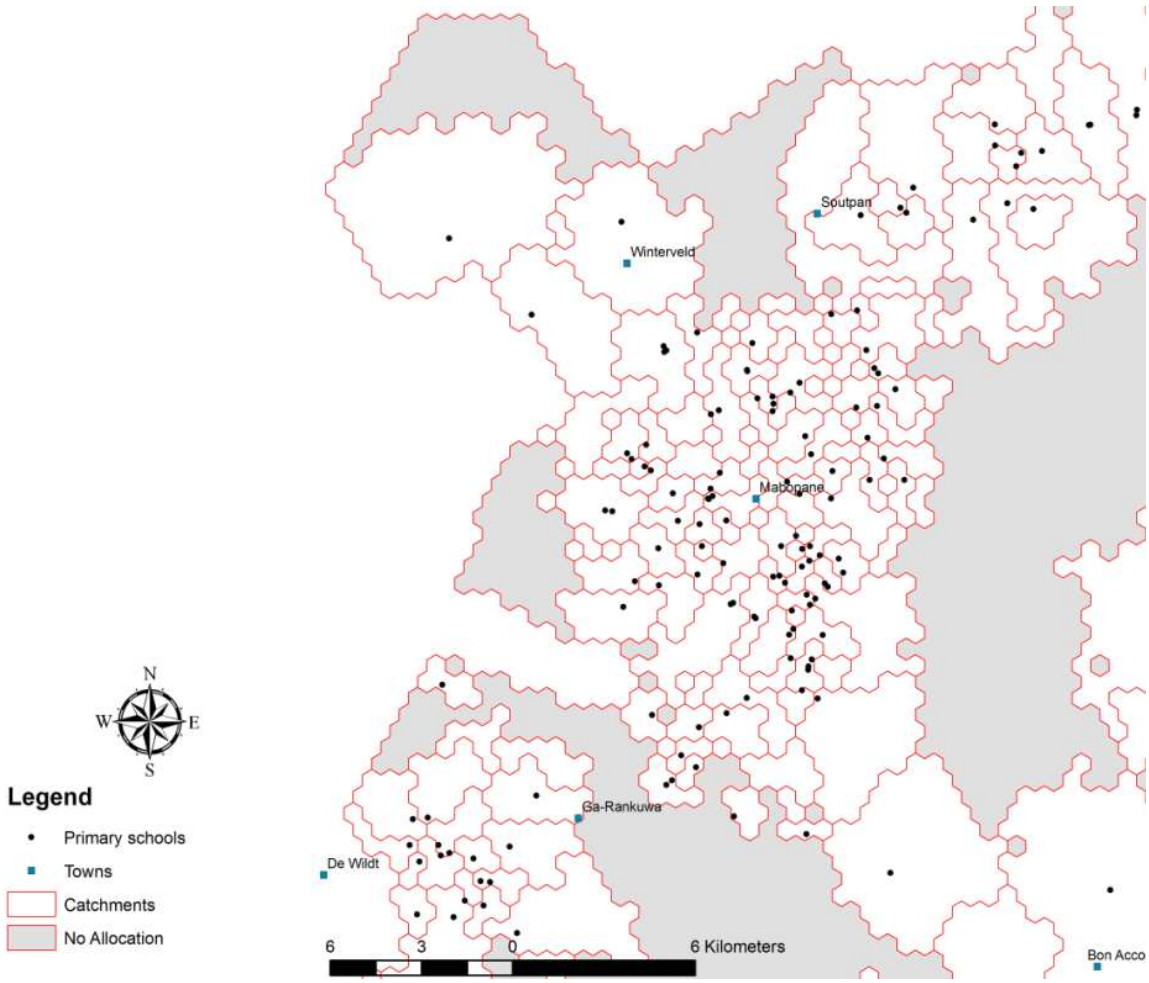

Figure 5: School catchment areas, where the grey spaces represent areas where there was demand for new schools

3.3 Urban air quality mapping

The South African PM_{10} air quality standard stipulates that the daily average PM_{10} concentrations in an area should be below 120 µg/m^3 (RSA *Government Gazette*, 2009). Air

quality monitors are installed with the purpose of monitoring compliance with air quality standards. However, these are stationed in specific locations, which is a challenge given that the need is for a pollutant concentration surface over the whole region of interest. Spatial interpolation techniques enable estimation of regional pollution surfaces from in situ air quality monitoring data. Kriging is a spatial interpolation technique, a distance-weighted average, where the weights are assigned to the observations modelled by a function of the correlation between pairs of points as a function of their separation (Bivand et al., 2008). Points in this case are the locations of air quality monitors in the study area. Distances between pairs of locations are considered in this case study and two variants, ordinary Kriging and Kriging with external drift, were explored. In ordinary Kriging, only the spatial variation of the response variable is considered. In Kriging with external drift, an assumption is of a spatial trend in the target or response variable that is explained by relating the response variable with variables capturing that underlying trend. Fluctuations remaining after removing the spatial trend are the local spatial variation and are modelled by means of a spatial correlation function.

4. Findings

In this section, findings from the three case studies are highlighted, focusing on the use of the census data and the outcomes thereof.

4.1 Findings in urban growth simulation

4.1.1 Household classification results

The final set of clusters for Ekurhuleni was a homogeneous group of small areas, where household characteristics and housing types met expectations for residential areas in this municipality. For instance, townships and informal settlements cover a large area in Ekurhuleni. Typical township formal houses have four rooms, inclusive of the living room, kitchen and two bedrooms. This corresponds with the substantial percentage (21%) of small areas characterised by four-roomed houses, occupied by six or more people with annual average income below the overall average for the municipality, and low education levels and employment amongst household heads.

To establish how well final classifications based on census data can be sufficiently accurate for urban modelling applications dependent on land-use characteristics, it was necessary to compare the census enumeration area (EA) type (the dominant land-use and the base for clustering) with an independent land-use dataset, the building-based land-use dataset of GeoTerra Image (GTI). The GTI dataset is derived from very high-resolution satellite images and aerial photography and municipal cadastral information (GeoTerra Image, 2012). The two datasets were overlaid on a map, so that small area boundaries could be superimposed on the GTI spatial point dataset, as shown in Figure 6. There were differences in definition of dwelling types between the two datasets. For informal dwellings, the census had two categories, differentiating free-hold informal dwellings from informal backyard structures. GTI also had these categories with an additional "transitional informal" dwellings category, which defines dwellings that cannot be classified as formal, free-hold or backyard informal structures. Due to these differences, complete agreement between the two datasets was not expected, but moderate to high agreement (more than 60%) was anticipated for most categories.

Figure 6: An overlay of small area boundaries on GTI building-based land-use point data for validation of Census 2011 dwelling type data

An interrater agreement statistic, namely the proportion of the agreement statistic (Cicchetti and Feinstein, 1990), was used for comparing the census and the GTI land-use information. Overall, the consensus between the two datasets as shown in Table 2 was high, with the highest level of agreement observed for formal residences, which includes detached, semi-detached and cluster housing. Moderate

agreement was observed for backyard informal dwellings, which could be attributed to the GTI data having an additional transitional informal category, which was absent from the census data. The lowest agreement was for high-rise apartments (flats). This could be attributable to differences in the unit of measure used in the different sources, e.g. GTI counts dwellings and therefore a block of flats would count as one unit, whilst in the census, each household within a block of flats is counted. It was concluded that the census EA type data was satisfactory for the purpose of classifying households required for urban modelling applications.

Table 3: Interrater statistics obtained in an assessment of validity of the census land-use (dwelling types) counts per small area against the GTI building-based land-use data

Dwelling type	**Observed agreement (as %)**
Formal semi/fully detached house	85
Informal house NOT in backyard	75
Informal house IN backyard	69
Flats or apartment	68
Cluster house	80

4.1.2 Comparing simulated and actual growth in household numbers

Urban growth simulations for the period between the 2001 and 2011 censuses were done in UrbanSim using household classification information. To validate the results, the simulated growth in number of households was compared with the actual growth during the same period, according to censuses 2001 and 2011 and GTI's building-based land-use dataset. A larger spatial unit of analysis (modified sub-places) was considered. The method used for adjusting census counts for changes in sub-place boundaries is described in Section 5.4.

Most of the outliers in the comparisons of simulated and actual household growth in Figure 7(a) could be explained. For instance, some outliers were attributable to delays in legal processes experienced by real-life developers but not known to developers in the model. If these outliers are excluded, the predictive accuracy of the model is estimated at about two housing units per hectare. The comparison also suggests that errors inherent in the model are similar in magnitude to the

differences between actual observations of two reputable sources (Figure 7(b)) over approximately the same period.

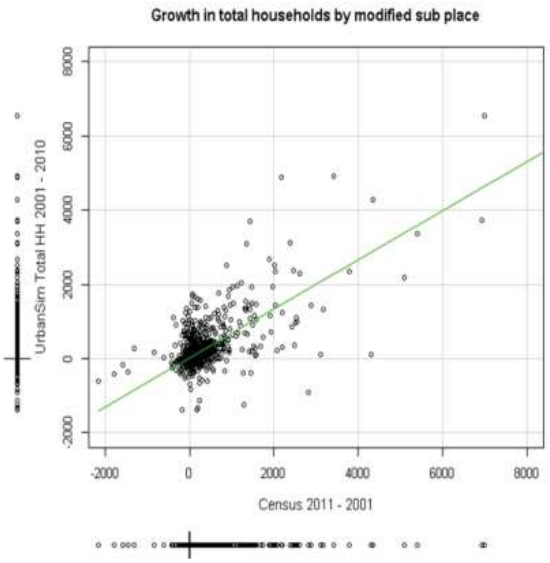

Figure 7 (a): Household growth between 2001 and 2011: Comparing UrbanSim predictions with Census 2011 data

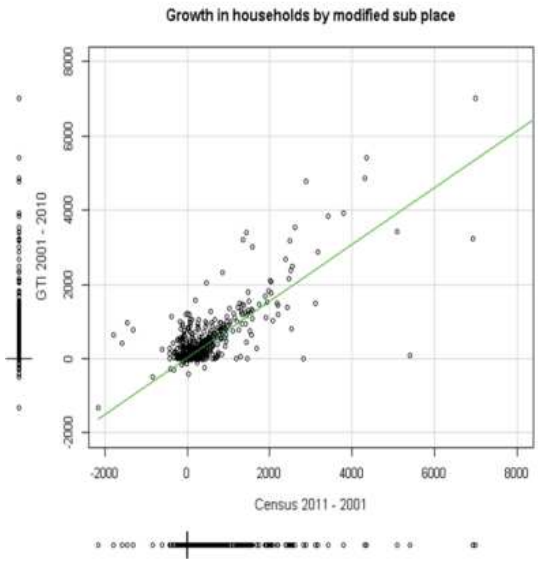

Figure 7 (b): Household growth between 2001 and 2011: Comparing GTI dwelling counts with Census 2011 data

4.2 Catchment analysis results

For the schools catchment analysis in Tshwane, distances between school locations from the unverified and the verified inventories were determined. When comparing

school locations between these two inventories, 65% of schools were within 100 m of the correct locations, while 2% (nine schools) were more than 5 km from the correct location. While the latter discrepancy seems insignificant, the financial implications are significant. For instance, the locations of ten new primary and ten new secondary schools were determined in Flowmap using the two inventories. With the verified inventory, only six new primary schools were needed, and the locations of two new secondary schools differed from those determined through the unverified inventory. Therefore, using an inaccurate inventory of schools would have led to an infrastructure misspend of over R 120 million (Schmitz and Eksteen, 2014).

4.3 Urban air quality mapping

The purpose of mapping the PM_{10} exceedances was to assess which areas have poor air quality and to determine what drives the observed patterns. An understanding of such drivers can inform strategies to control pollution at the source. Statistical mapping of the observed PM_{10} annual exceedance rate proceeded by means of Kriging, with the data

from 36 air quality monitors considered. The results from implementing ordinary Kriging show areas west of Gauteng, highlighted by the box in Figure 8 (top-left), as having high exceedance rates. Upon inspection, the high PM_{10} concentrations occur in high-density residential areas, hence the decision was to incorporate dwelling data in the Kriging model.

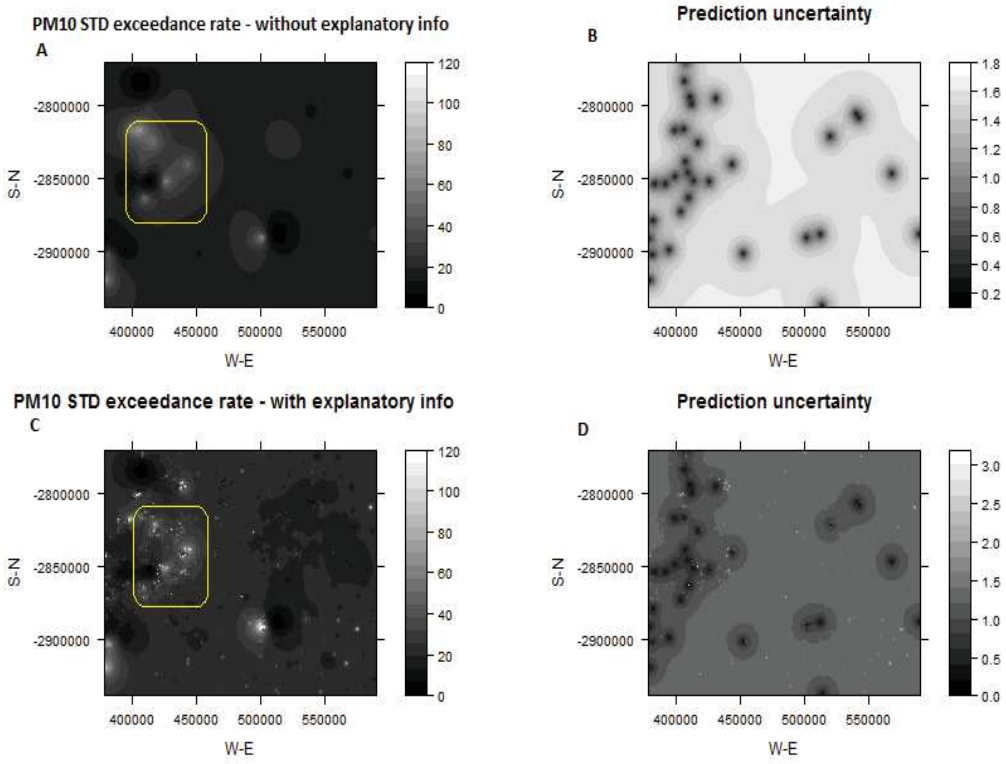

Figure 8: Resulting PM_{10} exceedance rate map, with and without use of explanatory information (A and C, respectively). The corresponding prediction uncertainty surfaces are shown as maps B and D

A description of the four explanatory variables considered in Kriging with external drift is given in Section 2.3. The number of dwellings per SAL was not a statistically significant explanatory variable, whereas the percentage of informal dwellings and biomass energy used for cooking and heating were significant. The advantage of using additional explanatory spatial variables that are significantly correlated with the response variable is an improved map. Kriging with external drift, using the percentage of informal dwellings per small area as the explanatory variable, resulted in a more nuanced pattern of spatial variation, as observed on the bottom-left map in Figure 8 (in comparison to the top-left map). The maps on the right of Figure 8 show the prediction error variance, a measure of precision. Precision was lowest (high prediction error variance) in areas without air quality stations (as expected). However, in adding an explanatory variable prediction, uncertainty is reduced – especially in areas without air quality monitors.

Biomass energy use variables were then included in the model, in addition to the informal dwelling percentage. Similar patterns were observed for the Highveld region. Results for

the Ekurhuleni municipal area were extracted and are shown in Figure 9(a). Considering Figure 9(b), it can be deduced that locations of poor air quality in Ekurhuleni (high PM_{10} exceedance rates), are characterised by typical township housings, including four-room type houses and RDP houses, inter-twined with informal dwellings as backyard shacks within townships, and stand-alone shacks in informal settlements. Therefore, this study found that domestic fuel combustion was a significant contributor to poor air quality in urban areas in Gauteng, confirming the results of previous environmental health studies in this region (Norman et al., 2007; Wright et al., 2011). Financial constraints could be the reason for such biomass energy use in these areas; hence strategies to curb pollution would need to consider the socio-economic conditions in these areas.

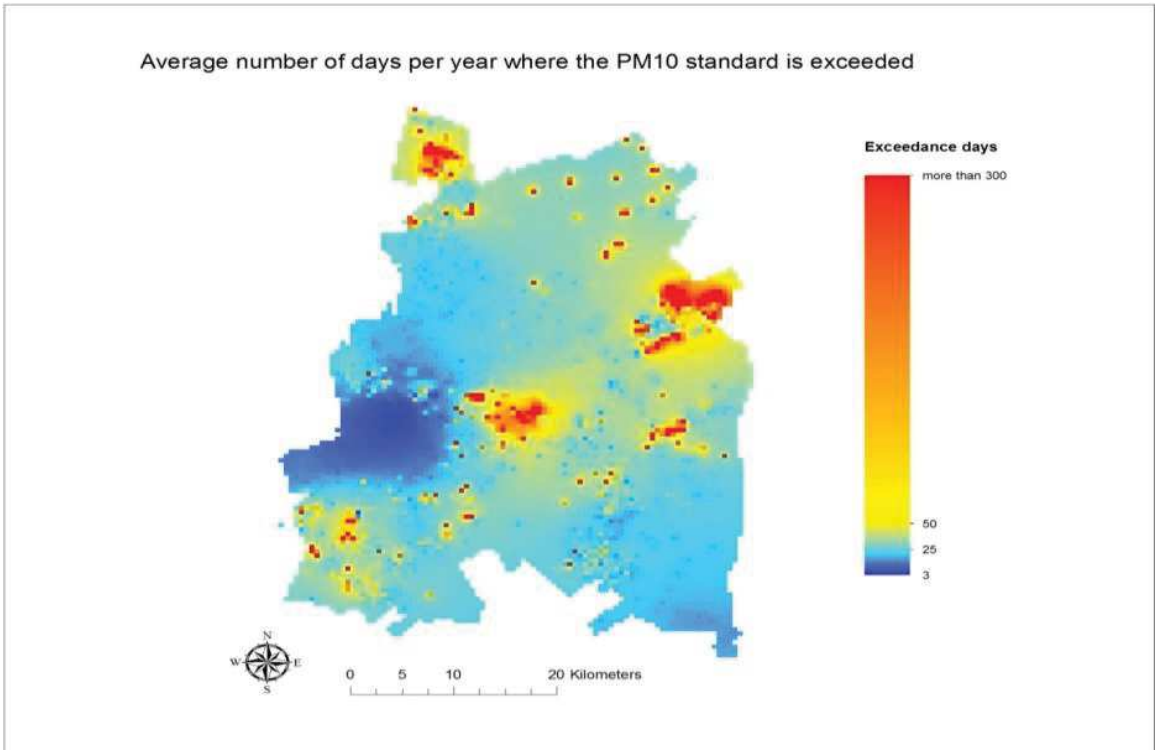

Figure 9(a): Predicted number of daily exceedance per year of the PM_{10} standard in Ekurhuleni with informal dwelling proportion and domestic use of biomass energy used as explanatory information

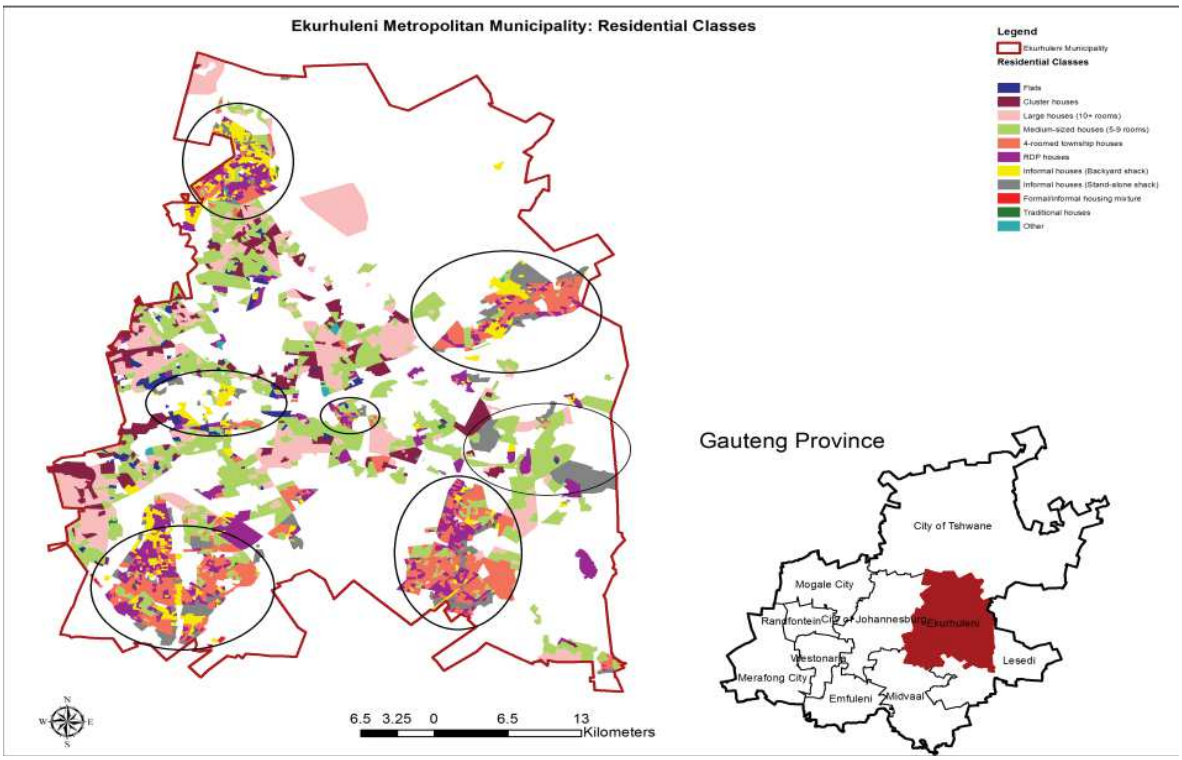

Figure 9(b): Map of Ekurhuleni Metropolitan Municipality with residential classes derived from the first two stages (consolidating 'dwelling location' and 'dwelling condition' clusters) of the multi-stage k-means procedure

5. Discussion

Governments, specifically municipalities, have a responsibility in ensuring there is adequate infrastructure for stimulating and sustaining local economic growth and for securing a better quality of life for their citizens. Planning in metropolitan areas in South Africa, particularly in Gauteng,

has moved away from master spatial planning into alternatives that advocate for linkages between human settlement planning, infrastructure development and ecosystem preservation (Gotz et al., 2004; Todes et al., 2010; Todes, 2012). A key resource for this approach to planning is the availability of good quality socio-demographic and housing data, such as national census data, which can be disaggregated to small areas. This was illustrated in this paper through three studies, showing how the South African Census 2011 data were processed to successfully assess demand growth for municipal services and infrastructure, and to identify areas of concern for pollution control. Another desirable property of census data for urban planning is the possibility of linking these data to household and travel surveys.

An important finding to emerge from the case studies was the importance of the geo-demographic classification from census data and the potential to use these in other urban modelling initiatives. In geo-demographic classification, comprehensive small area information from a census is condensed into classes or segments that describe the people

(households) and the characteristics of the areas in which they live (Ojo et al., 2013). The household classification tool was developed for modelling within UrbanSim. However, we demonstrated that these household classes are also applicable in interpreting the spatial distribution of exceedances of PM_{10} air quality standards, in our case leading to the conclusion that informal settlements are areas of poor air quality due to alternative energy similarly to findings of previous environmental health studies in Gauteng. Geo-demographic classification of census data to support planning decisions has a long history in the United Kingdom and the United States of America (Singleton and Spielman, 2014), with uptake in developing nations also gaining momentum. For instance, for the Philippines, a three-tier hierarchical geo-demographic classification was developed to support policy related questions (Ojo et al., 2013) and for Nigeria, 774 local government areas based on geo-demographic classification were created to enable understanding of the local population needs and provide evidence for planning and related policymaking (Ojo et al., 2012).

5.1 Challenges in calculating dwelling density per small area

Dwelling density was calculated using the number of dwellings in the census data and the area of the SAL, but Figure 10 gives a mapped example highlighting the problem of using density in this manner: the green polygons represent the small areas clustered as high-density formal housing, while the brown polygons are the low-density formal housing. On closer inspection, it was clear that the density of the dwellings is similar in these clusters, but the presence of open spaces or schools, i.e. multiple land-uses, in the brown polygons created a lower density value, implying these areas were less built-up than those in the green polygons. As a result, dwelling density had to be removed from the location characteristic factor in household segmentation, and only the census classification of EA type was used.

Figure 10: First layer of clusters based on density and location characteristics. Although small areas in green should be classified as those coloured in brown, they were misclassified because the presence of open spaces

5.2 Distortions caused by class definitions

It is known that in urban areas, such as in Gauteng, enumeration classification can be fraught with difficulties. In particular, some residential areas classified as informal in the census data give a clear impression of formalised housing on a map. Similarly, examples were found of small areas incorrectly classified as having collective living quarters. Although such

examples were found, it is difficult to check and rectify individual enumeration area types in a pure data clustering activity, involving 4,160 small areas say, but the consequence of not correcting such errors may be that certain areas are profiled incorrectly for demand for services. Therefore, a method for isolating such cases by using auxiliary information is of interest, particularly in areas that have developed into formal residential areas but are still being recorded as informal.

Another issue with dwelling type classification was the creation of outlier clusters due to lack of clarity in names. In particular, the definition of “traditional” is not clear, because one would assume that it refers to structures that involve traditional building materials, such as thatched roof huts. On close inspection of these small areas using Google Earth however, only one such small area could be seen that actually contained “traditional” dwellings according to this definition. Similar problems also arose with the “caravan/tent” classification. Therefore, going forward, it will be necessary to treat such atypical dwelling types in a different manner to that of other dwelling types in the clustering process, to prevent

creating outlier clusters from small areas which are in fact not outliers.

5.3 Problems identifying life-stage characteristics

During the cluster consolidation stages in our algorithm, clusters corresponding to factors 4 and 5 in Table 1 were initially derived, but were not considered for consolidation. Rather, the variables forming these factors were used descriptively on clusters formed by consolidating clusters corresponding to the first three factors. In particular, the 'life cycle stage and household (family) structure' factor was of interest in attempting to derive from the census data, the mobility of households within a small area. Life-stages of people in a household can be related to their probability of moving to other areas and their demand for services. With only the totals available for all categories in the life-stage key variables for each small area, it was impossible to group families, determine family types and subsequently derive information on their mobility. The clusters with people who were "never married" were the ones with all the children (above average values for the lower age groups), and small

areas dominated by married and divorced people with below average numbers of children seemed illogical. It was later established that all children are classified as “never married” in the data, even though they are too young to marry. Such clusters were difficult to interpret. If further information on family structures was available from the census data, it would enhance our understanding of the needs of households within a small area.

5.4 Consistency of spatial scales

The spatial scale of the data used in all three case studies was the small area. Consistency and appropriateness in spatial scales in census data is important in spatial urban planning, as the data are used to derive information for many factors and processes that interplay in the urban environment. The data are also used to track changes, through comparison with past censuses and other survey datasets, re-iterating the importance of consistency in spatial scales. An issue is that some small areas cover a much larger area than the average small area size. This is done to protect the anonymity of those living in less populated areas. However, for the purpose of

modelling, similarly sized spatial entities are often preferred and in the case of urban areas, the smaller the better, due to the high spatial density of buildings. This was a reason for the development of the Geographic Analysis Platform (GAP) geoframe with similar-sized mesozones to which statistics are disaggregated (Naudé et al., 2007).

Another spatial scale issue concerns sub-places. The country is divided into 22,108 areas delineating the boundaries of suburbs, sections and sub-villages. Suburbs are mostly found in built-up areas. Areas at the fringes of built-up suburbs are normally small or agricultural holdings. All other areas on the sub-place scale are then classed as non-urban (NU), e.g. Tshwane NU. Some NUs are large and consist of multiple, disjoint polygons, but are represented spatially as a single feature or area, as shown in Figure 11(a). The census data for these areas are also aggregated, as if the NU covers a single geographic location and not one that is dispersed over a large area. Therefore, it becomes a challenge to do a meaningful analysis where such NUs form a significant part of one's area of interest. A solution that was developed to enable urban growth modelling was to create a modified

spatial representation of the sub-places, referred to as "modified sub-places", as shown in Figure 11(b). These modified sub-places were derived by subdividing all sub-places larger than 3 km^2 into smaller hexagons, as shown in Figure 11(b). The 3 km^2 was chosen because it corresponded to the 75^{th} percentile of the area in Gauteng. By making use of GTI's auxiliary datasets indicating the actual locations of dwellings, the data for each of the modified sub-places were calculated and apportioned from the data for the entire NU.

Figure 11(a): This map shows a large area, consisting of multiple suburbs, which is classified as a single non-urban (NU) area at sub-place level for Tshwane municipality

Figure 11(b): Illustrating how large sub-places were modified into smaller hexagonal polygons, which were then named 'modified sub-places'

5.5 Lack of workplace data

Information on place of work, i.e. the geographic location where workers carry out their occupations, is important in planning. Although a question about the place of work had been included in the household questionnaires of some previous South African censuses (but not in the Census 2011), the availability and reporting of the number of workers by place of work at a practical geographical level have been scarce. In the previous census, the resolution of place of work was main-place, unless the answer to another question about working and living in the same sub-place place was true, in which case the place of work could be resolved to sub-place. It is not certain whether this is due to the quality of the responses or whether it can be attributed to inappropriateness of automatic geocoding, which is only possible where street addresses or aliases are available. The location of informal activities and of activities of variable location, such as construction, adds complications.

The lack of suitable workplace data in South Africa requires additional expensive and *ad hoc* surveys for various

studies. The United Nations recommends that place of work be included in national censuses (United Nations, 2007). In countries such as the United States of America, United Kingdom, Australia, New Zealand and Ireland, workplace data have been successfully collected during censuses and reported over a number of years. Sometimes the results of additional surveys, such as the American Community Survey (ACS), have been used to enhance the census information (Statistics New Zealand, 2013; UK Office for National Statistics, 2014). Statistics New Zealand's standard for workplace address is a nine-page document dealing with operational issues, classification criteria, outputs, coding process and related standards. This information has been used in numerous studies of commuting patterns, workplace population analyses, calculating daytime population and visualising worker movements at various geographical levels (Statistics New Zealand, 2009; UK Office for National Statistics, 2013; UK Office for National Statistics, 2014; McKenzie et al., 2014; Rapino et al., 2014).

One possible solution for the lack of place of work data we are considering is combining the census 10% sample with

other surveys (undertaken or facilitated by Statistics South Africa), such as the labour force and national household travel surveys. Each has particular strengths. The 10% sample provides the best characterisation of person and household attributes, the labour force survey the best information on the type of work, while the household travel survey has the best information on origin, destination and mode and cost of travel. This will, however, have to be done in collaboration with Statistics South Africa because it requires sample-specific information not ordinarily released for reasons of confidentiality.

6. Conclusions

There were three main contributions made by this paper. Firstly, successful experiences in using the South African Census 2011 small area data in modelling that resulted in valuable input for urban planning were shared. The second contribution is the use of statistical interrater measures to show the similarity of a census housing-type classification to an independent building-based land-use classification dataset. Thirdly, according to the authors' knowledge, this is the first

paper to highlight the challenge of lack of workplace information from the South African census and to present a possible way to overcome this problem.

7. Acknowledgements

We acknowledge Stats SA, the South African Weather Service, GeoTerra Image, and the South African Department of Basic Education for the data. The CSIR and Nuffic are acknowledged for the funding, and we appreciate the reviewer's comments, which have improved our contributions.

References

Adnan, M., Longley, P.A., Singleton, A.D. & Brunsdon, C. 2010. Towards real-time geodemographics: Clustering algorithm performance for large multidimensional spatial databases. *Transactions in GIS*, 14(3):1467-9671.

Bivand, R.S., Pebesma, E.J. & Gomez-Rubio, V. 2008. *Applied spatial data analysis with R.* Springer, New York.

Borning, A., Waddell, P. & Förster, R. 2008. UrbanSim: Using simulation to inform public deliberation and decision-making. In *Digital government,* pp. 439-464. Springer, New York.

Cicchetti, D.V. & Feinstein, A.R. 1990. High agreement but low kappa: II. Resolving the paradoxes. *Journal of clinical epidemiology,* 43(6):551-558.

Cilliers, S., Du Toit, M., Cilliers, J., Drewes, E. & Retief, F. 2014. Sustainable urban landscapes: South African perspectives on transdisciplinary possibilities. *Landscape and Urban Planning,* 125:260-270.

CSIR. 2011. *CSIR Focus on national service delivery support.* CSIR, South Africa.

http://www.csir.co.za/publications/pdfs/CSIR%20Service%20 Delivery%20FINAL.PDF.

De Jong, T. & Van der Vaart, N. 2013. *Manual Flowmap 7.4.2.* Faculty of Geographical Sciences, Utrecht University, Netherlands. http://flowmap.geo.uu.nl/downloads/FM742_Manual.pdf.

Dudeni-Tlhone, N., Holloway, J.P., Khuluse, S. & Koen, R. 2013. Clustering of housing and household patterns using 2011 population census. *Paper in proceedings of the 55th Annual Conference of the South African Statistical Association*, Polokwane, South Africa, 4-8 November 2013. pp. 23-30. http://researchspace.csir.co.za/dspace/handle/10204/7174

GeoTerra Image. 2012. Building-based land use - Product summary sheet. GeoTerra Image, South Africa. http://www.geoterraimage.com/pdfs/101%20Building%20Based%20Land%20Use.pdf

Green, C. & Argue, T. 2012. *CSIR Guidelines for the provision of social facilities in South African settlements.* First edition. CSIR, South Africa. http://www.csir.co.za/Built_environment/pdfs/CSIR_Guidelines. pdf.

Gotz, G., Allan, K. & Harrison, K. 2004. Sustainable cities: Built- and natural-environment trends and the state of the sustainable city. *In State of the cities report*. First edition, Boraine, A. (ed). South African Cities Network.

Huang, Z. 1998. Extensions to the k-means algorithm for clustering large datasets with categorical values. *Data Mining and Knowledge Discovery*, 2(3):283–304.

Jain, A.K., Murty, M.N., & Flynn, P.J. 1999. Data clustering: a review. *ACM computing surveys (CSUR)*, 31(3):264–323.

Khuluse, S. & Stein, A. 2013. Mapping the annual exceedance frequencies of the PM_{10} air quality standard - Comparing kriging to a generalized linear spatial model. *Poster presented at the 55th Annual Conference of the South African Statistical Association*, Polokwane, South Africa, 4-8 November 2013.
http://researchspace.csir.co.za/dspace/handle/10204/7780

MacQueen, J. 1967. Some methods for classification and analysis of multivariate observations. *In proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 1:281-297. University of California Press, USA.

McKenzie, B., Koerber, W., Fields, A., Benetsky, M. & Rapino, M. 2014. *Commuter Adjusted Daytime Population: 2006-2010.* United States Census Bureau, USA. https://www.census.gov/hhes/commuting/files/ACS/Commuter% 20 Adjusted%20Population%20Paper.pdf

Naudé, A., Badenhorst, W., Zietsman, L., Van Huyssteen, E., and Maritz, J. 2007. *Geospatial Analysis Platform – Version 2: Technical overview of the mesoframe methodology and South African Geospatial Analysis Platform*. CSIR Report, CSIR/BE/PSS/IR/2007/0104/B. CSIR, South Africa.

Norman, R., Cairncross, E., Witi, J., Bradshaw, D. & the South African Comparative Risk Assessment Collaboration Group. 2007. Estimating the burden of disease attributable to urban outdoor air pollution in South Africa in 2000. *South African Medical Journal,* 97(7):782-790.

Ojo, A., Vickers, D. & Ballas, D. 2012. The segmentation of local government areas: Creating a new geography of Nigeria. *Applied spatial analysis and policy*, 5(1):25-49.

Ojo, A., Vickers, D. & Ballas, D. 2013. Creating a small scale area classification for understanding the economic, social and housing characteristics of small geographical areas in

the Philippines. *Regional Science Policy & Practice*, 5(1):1757-7802.

UK Office for National Statistics. 2014. *Workplace population analysis, 2011 census.* Office for National Statistics, London, United Kingdom. http://www.ons.gov.uk/ons/rel/census/2011-census/workplace-population-statistics-for-workplace-zones-and-middle-layer-super-output-areas--msoas--in-england-and-wales/workplace-population-analysis--2011-census.html.

UK Office for National Statistics. 2014. *Workplace zones (WZ).* Office for National Statistics, London, United Kingdom. http://www.ons.gov.uk/ons/guide-method/geography/beginner-s-guide/census/workplace-zones--wzs-index.html

UK Office for National Statistics. 2013. *The workday population of England and Wales: An alternative 2011 census output base.* Office for National Statistics, London, United Kingdom. http://www.ons.gov.uk/ons/rel/census/2011-census/workday-population-statistics-for-output-areas-in-england-and-wales--part-1-/rpt-workday-population-of-england-and-wales.html.

Piketh, S.J., Otter, L.B., Burger, R.P., Walton, N., Van Nierop, M.A., Bigala, T., Chiloane, K.E. & Gwaze, P. 2004. *Brown haze II report: Cape Town.* Climatology Research Group, University of Witwatersrand, South Africa. http://www.capetown.gov.za/en/CityHealth/AirQualityManagement/BrownHaze/Pages/BrownHazeIIStudy.aspx.

Rapino, M.A., Koerber, W.K. & McKenzie, B. 2014. *How can we best visualize worker movement throughout the day?* SEHSD Working Paper No. 2014-01, United States Census Bureau, USA. https://www.census.gov/hhes/commuting/files/ACS/worker-movement-paper.pdf.

RSA *Government Gazette*. 2009. *National ambient air quality standards. National environmental management: Air quality act 2004 (Act No. 39 of 2004), 534(32816).* https://www.environment.gov.za/sites/default/files/legislations/nemaqa_airquality_g32816gon1210.pdf.

Singleton, A.D. & Spielman, S.E. 2014. The past, present, and future of geodemographic research in the United States and United Kingdom. *The Professional Geographer*, 66(4):558-567.

Schmitz, P. & Eksteen, S. 2014. The Effect of GIS data quality on infrastructure planning: School accessibility in the City of Tshwane, South Africa. *In proceedings of the second AfricaGEO conference*, J. Whittal & S. Motala (eds), Cape Town, South Africa, 1-3 July 2014. http://researchspace.csir.co.za/dspace/handle/10204/7874

Statistics New Zealand. 2013. *2013 Census information by variable.* http://www.stats.govt.nz/Census/2013-census/info-about-2013-census-data/information-by-variable.aspx#

Statistics New Zealand. 2009. *Commuting patterns in New Zealand: 1996-2006.* http://www.stats.govt.nz/browse_for_stats/Maps_and_geography/Geographic-areas/commuting-patterns-in-nz-1996-2006.aspx

Statistics South Africa. 2012a. *Census 2011 statistical release – P0301.4.* Statistics South Africa, Pretoria. http://www.statssa.gov.za/publications/P03014/P030142011.pdf

Statistics South Africa. 2012b. *Census 2011 metadata.* Statistics South Africa, Pretoria. http://www.statssa.gov.za/census/census_2011/census_prod ucts/Census_2011_Metadata.pdf.

Todes, A., Karam, A., Klug, N. & Malaza, N. 2010. Beyond master planning? New approaches to spatial planning in Ekurhuleni, South Africa. *Habitat International*, 34(4):414-420.

Todes, A. 2012. Urban growth and strategic spatial planning in Johannesburg, South Africa. *Cities*, 29(3):158-165.

United Nations. 2007. *Principles and recommendations for population and housing censuses: Revision 2.* Series M No. 67/Rev.2 edn, Statistical Paper ST/ESA/STAT/SER.M/67/Rev.2. New York, USA. http://unstats.un.org/unsd/publication/seriesM/seriesm_67Rev2e.pdf

Vickers, D. & Rees, P. 2007. Creating the UK national statistics 2001 output area classification. *Journal of the Royal Statistical Society: Series A (Statistics in Society),* 170(2):379-403.

Waddell, P. 2005. Confronting the bane of endogeneity in modelling urban social dynamics. *In proceedings of the Workshop on Modelling Urban Social Dynamics,* University of Surrey, 7-8 April 2005. http://www.urbansim.org/pub/Research/ResearchPapers/surrey.pdf.

Waldeck, L. 2013. *Overview of urban growth simulation: With examples from different cities.* Presented at the CSIR Durban regional office. CSIR, South Africa. http://researchspace.csir.co.za/dspace/handle/10204/7264

Wegener, M. 1994. Operational urban models: state of the art. *Journal of the American Planning Association,* 60(1):17-29.

Wright, C., Garland, R., Thambiran, T. & Diab, R. 2011. Air quality: A South African perspective. *Journal of the Institution of Environmental Sciences,* 20(1):25-27.

Yoon, K.A., Kwon, O.S. & Bae, D.H. 2007. An approach to outlier detection of software measurement data using the k-means clustering method. *In First International Symposium on Empirical Software Engineering and Measurement. IEEE*, 443–445.