

Effects of different building blocks designs on the statistical characteristics of Automated Zone-design Tool output areas

Mokhele T^{a1}, Mutanga O^b and Ahmed F^c

^aResearch Methodology and Data Centre, Human Sciences Research Council, South Africa;

^bSchool of Agriculture, Earth and Environmental Sciences, University of KwaZulu-Natal, South Africa;

^cSchool of Geography, Archaeology & Environmental Studies, University of The Witwatersrand, South Africa

DOI: <http://dx.doi.org/10.4314/sajg.v6i2.2>

Abstract

Prior to any census, the country usually gets demarcated into small geographic units called census enumeration areas, districts or blocks. In most countries, these small geographic units are also used for census dissemination. In cases where they are not used for census release, they are normally used as building blocks for developing output areas or they are aggregated to higher spatial levels in an effort to preserve privacy or confidentiality. Buildings blocks are therefore, of significant importance towards results that could be drawn from either aggregated higher level or from output areas developed using these small geographic areas. This paper aimed at evaluating the effects of different building blocks on the statistical characteristics of output areas generated using the Automated Zone-design Tool (AZTool) computer program. Different spatial layers (such as Enumeration Areas (EAs), Small Area Layers (SALs) and SubPlaces) from the 2001 census data were used as building blocks for the generation of census output areas with AZTool program in both rural and urban areas of South Africa. One way-Analysis of Variance (ANOVA) was also performed to determine statistical significance of the AZTool results. Results showed that the AZTool output areas generated from smaller areas (EAs and SALs) tend to be more homogeneous than the ones generated from larger areas (SubPlaces) when using dwelling type and geotype as homogeneity variables. The output areas from smaller areas also had narrower population distribution and more compact shapes than their counter-parts. In addition, the AZTool optimised output areas from the smaller areas allowed a clear distinction of the scale effects than output areas

¹ Corresponding Author: Email: TAMokhele@hsrc.ac.za

from larger areas. It was concluded that indeed different building blocks did have an impact on the statistical qualities of the AZTool optimised output areas in both rural and urban settings in South Africa.

Keywords: AZTool; Building blocks; Enumeration areas; Output areas; Scale effects; Small Area Layers; SubPlaces.

1 Introduction

It is common practice around the world that before any census is conducted the country usually gets demarcated into census Enumeration Areas (EAs). These small areas are normally designed in such a way that they are of a size enough to be covered by one census enumerator within the census period. In South Africa, these small areas (such as EAs) normally contain from 100 to 250 households (Stats SA, 2003; Stats SA and HSRC, 2007; Verhoef and Grobbelaar, 2005). The criteria for the design of EAs are that: firstly, they should not overlap; secondly, they should be compact without pockets or disjointed sections and should cover the entire country; thirdly, they should have boundaries that could be identified on the ground; and last but not the least, they should be of approximately equal population size to enable an enumerator to cover each one in the allocated census period (Stats SA and HSRC, 2007; Mokhele *et al*, 2016). Before 1996, the boundaries for EAs were hand-drawn, which is traditional demarcation (Laldaparsad, 2007). The 1996 census represented a transition from traditional demarcation and mapping methods towards an electronic geographic database. For the 2001 census, Geographic Information System (GIS) technology was used to draw EAs (80787 EAs) and for map production. For instance, about 80% of the 2001 EA demarcation was done in the office on a GIS using photography and digital topographical maps (Stats SA, 2003; Laldaparsad, 2007). For the other 20%, field inspection and other alternatives were considered (Stats SA, 2003; Laldaparsad, 2007). For the 2011 census, there were 103576 newly demarcated EAs across the country.

The 2001 census was not released at EA level in an effort to preserve confidentiality of individuals, but instead data was released at SubPlace level. These SubPlaces were too large for most census data users and did not have tighter and narrower population distribution hence comparability of areas with respect to population size was a challenge. A new spatial layer, the SAL, was therefore created using the non-zone design approach in 2005 for release of the 2001 census data at lower

level. A similar non-zone design approach was also employed in the creation of SAL for the 2011 census data (Mokhele *et al.*, 2016).

Some countries such as Australia have moved towards having nationally consistent small areas, mesh blocks, which would be used as a stable basis for their output zones and systems for many years to come (Cockings *et al.*, 2013). This move perhaps is worth an investigation in South Africa as this would allow trend analysis and comparisons between different censuses at smaller areas. Generally, geographic shape compactness is of concern with regard to urban morphology, political districting, and accuracy of enumeration unit values (MacEachren, 1985). Zones or areas with compact shapes whose boundaries follow recognisable features on the ground are often desirable for mapping purposes whereas homogeneity of population size is often preferable for statistical analysis (Cockings *et al.*, 2013). Social homogeneity of zones or areas on the other hand could also be of high importance as this could be used as an indication of where resource allocation or service deliveries should be prioritized by governments and Non-Government Organisations (NGOs). However, practical considerations often out-compete more conceptual aspects when designing these small areas or building blocks design (Cockings *et al.*, 2013).

In most countries, these small geographic units are also used for census dissemination. In cases where they are not used for census release, they are normally used as building blocks for developing output areas or zones or they are aggregated to higher spatial levels (Cockings *et al.*, 2013). This aggregation is often done on the basis of geographical location and usually data are made available at two or more spatial levels (Flowerdew, 2011). It is noteworthy that small areas or building blocks would always be of high importance for the dissemination of national population statistics due to confidentiality issues even if census is replaced by other systems such as registers or any administrative datasets, like in Denmark and Finland (Valente, 2010). Buildings blocks are therefore of significant importance towards results that could be drawn from either aggregated level or from output areas developed using these small geographic areas.

The fact that census data is collected for individual households but is usually released at higher levels to preserve confidentiality raises some concerns. This results in a problem called the Modifiable Areal Unit Problem (MAUP) originally discovered by Gehlke and Biehl (1934). The

MAUP has two components, the scale effect and the zonation (Openshaw, 1977; 1984; Reynolds, 1998; Ratcliffe and McCullagh, 1999; Kitchin and Tate, 2000; Heywood *et al.*, 2002; Duque *et al.*, 2007; Dumedah *et al.*, 2008; Ralphs and Ang, 2009; Flowerdew, 2011; Mokhele *et al.*, 2016). These two effects occur due to the fact that spatial processes generating the observed data may exist at scales and for particular areal units that may be reflected more or less accurately by the boundaries that are used (Manley *et al.*, 2006).

Cockings *et al.* (2013) evaluated the influence of two sets of building blocks (street blocks and postcodes) on output zone characteristics using six local authorities in England and Wales. Their findings indicated that postcodes were more effective building blocks than street blocks as they provided more uniform population and household sizes. On the other hand street blocks were found to produce more compact output zones with greater internal homogeneity of tenure and accommodation type. They also found that the scale effect of the modifiable areal unit problem and the specific geographical patterning of variables were important factors when designing building blocks. Therefore, this paper was aimed at evaluating the effects of different building blocks on the statistical characteristics of the AZTool optimised census output areas in South Africa.

2 Methods

Two provinces were selected for this study; Free State and Gauteng, which were representatives of rural and urban areas, respectively (Mokhele *et al.*, 2016). In each province, different spatial or geography levels such as district, municipality and mainplace were selected for subsequent analysis. There were no provincial boundary changes for Free State province in 2011 and its total population did not change substantially between 2001 and 2011 hence comparisons of the two censuses data could be undertaken where necessary for this study area. As both rural and urban settings were represented, findings from these study areas are likely to apply in many other parts of South Africa.

The original 2001 census data from Statistics South Africa (Stats SA) at SAL and SubPlace levels for the two provinces were extracted. It is noteworthy to mention that the 2001 census estimates (HSRC, 2005) were used for EA-level data as this data was not accessible from Stats SA. The

2011 census data at SAL level was also extracted for Free State province to allow comparison with the 2001 census data as this province did not exhibit a significant population change between 2001 and 2011 as well as its boundaries which did not change. The extracts from the data included total population, homogeneity variables (dwelling type and geotype) as well as spatial levels related information. Therefore, different spatial layers (2001 EAs, 2001 Subplaces, 2001 and 2011 SALs) were used as building blocks for the generation of census output areas in order to determine the impact of building blocks of output areas.

These output areas were generated using the Automated Zone-design Tool (AZTool) version 1.0.3 (Cockings *et al.*, 2011) with pre-defined design criteria such as minimum population threshold, population target, shape and homogeneity. The Automated Zone-design Tool program algorithms usually take input building blocks and iteratively aggregate them into larger output areas from an initial random aggregation, by checking the effect of swapping individual building blocks between output areas based on criteria set by the user (Openshaw, 1977; Mokhele *et al.*, 2016). All the AZTool output areas were generated using different building blocks with a population threshold of 500 (as practised by Stats SA) and a population target of 1000 to ensure confidentiality limit (Verhoef and Grobbelaar, 2005; Mokhele *et al.*, 2016). Confidentiality limit is minimum population that is used for the dissemination of census data in order to avoid personal information disclosure. The Intra-Area Correlation (IAC), described as direct measure of within-area homogeneity and between-area heterogeneity, was used to measure the degree of homogeneity within the AZTool output areas (Tranmer and Steel, 1998; 2001; Martin *et al.*, 2001; Flowerdew, 2011). The IAC values range from 0 to 1. For instance, a higher IAC value indicates a higher degree of homogeneity within-area and a higher degree of heterogeneity between areas (Tranmer and Steel, 1998; Martin *et al.*, 2001; Cockings *et al.*, 2013). In terms of the shape of output areas, shape compactness was explored. Shape compactness of the area means the degree to which the area has a compact (rather than linear) shape (MacEachren, 1985; Mokhele *et al.*, 2016). This is mainly prioritised if the purpose of the output areas is just for mapping. The overall Perimeter Squared per Area (P2A) was employed as a measure of shape compactness (MacEachren, 1985; Cockings and Martin, 2005; Haynes *et al.*, 2007; Mokhele *et al.*, 2016). Briefly, low P2A mean values indicate more compact shapes. The Statistical Package for the Social Sciences (SPSS) was also employed for further statistical analysis such as Analysis of Variance (ANOVA).

3 Results

3.1 Effect of building blocks on statistical qualities of output areas in rural settings

Table 1 summarises characteristics of output areas developed using three different building blocks (EAs, SALs and SubPlaces) at the rural settings. The confidentiality limit of 500 persons was adhered to for all output areas from the three different building blocks. The AZTool output areas from the EAs had slightly higher population means and lower standard deviations than the ones developed with SALs as building blocks. This means that the SALs built output areas were slightly tighter than the ones created from the EAs with regard to population distribution. The output areas from the SubPlaces on the other hand had higher population means and higher standard deviations. With regard to shape compactness, the lower P2A mean values indicated that output shapes were more compact whereas higher P2A mean values indicated that output areas were less compact. The P2A mean values for output areas from the EAs and the SALs were almost similar but the latter had slightly higher standard deviations at all levels. The output areas from the SubPlaces had higher P2A means and higher standard deviations than the ones generated from the EAs and the SALs at all spatial levels. Clearly, this shows that the output areas created using the EAs and the SALs were significantly ($p < 0.05$) more compact than those developed using the SubPlaces as building blocks. The post-hoc test results showed that P2A means for output areas from both the EAs and the SALs were not significantly different ($p > 0.05$). The results further indicated that the difference between P2A means of those generated from the SubPlaces and the EAs and the difference between P2A means of those created from the SubPlaces and the SALs was statistically significant ($p < 0.05$).

For homogeneity, only the AZTool optimised output areas from the EAs and SubPlaces yielded reasonable results. The SALs ones did not have enough homogeneity variables hence the IAC score produced not a number i.e. the SALs data did not have the dwelling type variable. At lower levels, the IAC scores for the output areas from the EAs were lower than those from the SubPlaces while at higher spatial levels the opposite was the case. The IAC score for the output areas developed using the EAs as building blocks was 0.59 while that of using the SubPlaces was 0.51

at provincial level. These statistics indicate that the output areas from the EAs were less homogeneous than those from the SubPlaces at lower levels (mainplace and municipality) while EAs output areas were more homogeneous than the SubPlaces ones at the provincial level. The two sets of output areas were homogeneously the same at the district level as they both had IAC score of 0.56.

Table 1: Statistical characteristics of output areas from the EAs, the SALs and the SubPlaces for Free State

	Number of Zones		Population			Shape		Homogeneity
	Min	Max	Mean	SD	Mean	SD	IAC	
EA Output Areas								
Phuthaditjhaba	48	649	2704	1113	346	28	9	0.21
Maluti-a-Phofung	349	610	2704	1027	232	32	13	0.50
Thabo Mofutsanyane	667	581	5292	1087	403	33	13	0.56
Free State	2440	547	9269	1101	489	31	12	0.59
SAL Output Areas								
Phuthaditjhaba	53	701	1602	1003	229	26	14	N/A
Maluti-a-Phofung	361	646	2071	999	230	32	14	N/A
Thabo Mofutsanyane	726	615	6701	1000	353	32	15	N/A
Free State	2707	619	7747	1000	342	31	13	N/A
SubPlace Output Areas								
Phuthaditjhaba	10	722	10507	5318	3902	47	32	0.32
Maluti-a-Phofung	53	516	22496	6807	4657	41	19	0.55
Thabo Mofutsanyane	112	631	26411	6482	4475	39	18	0.56
Free State	352	530	93782	7690	7890	37	21	0.51

3.2 Effect of building blocks on statistical qualities of output areas in urban settings

Table 2 presents the statistical qualities from similar analysis but this time for urban areas. For the mean population target, similar trends were noticed as the output areas from the SALs were having lower means (almost similar to the target mean) and standard deviations than those developed using the EAs. In addition, the AZTool output areas from the SubPlaces also had higher population means and higher standard deviations than those developed from the EAs and the SALs as in rural areas. Similar trends as to those in rural areas were also seen for the optimised output areas from all the three different building blocks from the shape compactness of the shape point of view. However, the P2A mean values of the output areas from the SubPlaces were not as higher as they

were in rural areas but they were statistically different from their counter-parts with one-way ANOVA revealing p-value less than 0.05 ($p = 0.006$).

In contrast to rural areas, the IAC scores for the output areas from the EAs were higher than those of the output areas from the SubPlaces at lower levels. At higher level, provincial level, the IAC score for the output areas developed using the EAs as building blocks was still higher than those from the SubPlaces. This highlights that for the urban areas, the automated zone design output areas generated using the EAs as building blocks were more homogeneous than their counter-parts at all spatial levels.

Table 2: Statistical characteristics of output areas from the EAs, SALs and SubPlaces for Gauteng

	Number of Zones	Population				Shape		Homogeneity
		Min	Max	Mean	SD	Mean	SD	IAC
EA Output Areas								
Pretoria	500	621	5026	1056	320	28	11	0.09
City of Tshwane	1276	502	8802	1203	514	27	10	0.46
Gauteng	7253	501	9627	1214	520	27	9	0.45
SAL Output Areas								
Pretoria	525	640	4227	1001	281	27	12	N/A
City of Tshwane	1527	587	8092	1000	429	26	11	N/A
Gauteng	8837	586	8400	1000	365	26	11	N/A
SubPlace Output Areas								
Pretoria	74	623	26773	7100	5425	29	12	0.06
City of Tshwane	135	577	82002	11311	11930	29	14	0.42
Gauteng	967	523	131662	9139	10052	31	17	0.40

3.3 Effect of building blocks from different censuses in rural settings

As the 2001 SALs did not have enough homogeneity variables, the 2011 SALs (from the 2011 census data) were used as building blocks to determine effects of the SALs on the statistical qualities of AZTool output areas in terms of degree of homogeneity. Only SAL data for Free State province was extracted from the 2011 census data as this province did not change boundaries from 2001 and as its total population had only slight increase while Gauteng province had changes on its provincial boundaries and its total population increased substantially. It is noteworthy to

mention that the boundaries of all other lower spatial levels changed from the 2001 census, hence only the provincial level of the 2011 census results could be compared with the 2001 census ones. Although both the 2001 EAs and 2011 SALs data had the dwelling type and the geotype homogeneity variables, there were slight differences in terms of their categories, that is, the 2001 EAs had 9 fields for dwelling type and 4 for geotype while the 2011 SALs data had 12 fields (cluster house, townhouse and caravan as extra fields) for dwelling type and 3 for geotype (Only Urban, no more Formal and Informal Urban).

The 2001 EAs optimised output areas had slightly higher population mean of 1101 and higher standard deviation of 489 compared with 1056 and 264 of the output areas generated using the 2011 SALs as building blocks (Table 3). This highlights that the output areas from the 2001 EAs were slightly less tight than their counter-parts with regard to population distribution. However, it should be mentioned that the population means for the optimised output areas from both sets of building blocks were close to the target mean of 1000 people which was set on the design criteria. The output areas from the two sets of building blocks were similar with regard to shape compactness. The AZTool output areas developed from the 2001 EAs were more homogeneous than the ones created using the 2011 SALs as building blocks with IAC score of 0.59 and 0.55 respectively. Table 3 further indicates that there is continuous decreasing trend with regard to IAC scores for output areas from smaller areas to larger areas as output areas from the SubPlaces recorded the lowest IAC value of 0.51.

Table 3: Statistical characteristics of output areas using 2001 EAs, 2011 SALs and 2001 SubPlaces as building blocks for the Free State province

Free State Province	Number of Zones	Population			Shape		Homogeneity	
		Min	Max	Mean	SD	Mean	SD	IAC
2001 EA Output Areas	2440	547	9269	1101	489	31	12	0.59
2011 SAL Output Areas	2596	594	5586	1056	264	31	11	0.55
2001 SubPlaces Output Areas	352	530	93782	7690	7890	37	21	0.51

The fact that almost half (42.7%) of the SALs for the 2011 census breached the confidentiality limit of 500 people in Free State province, prompts further generation of census output areas in South Africa, if confidentiality is taken seriously. The argument is that there is a pressing need for the creation of the 2011 census output areas which truly respect confidentiality limit as much as

possible. The AZTool program was then used to explore effects of different homogeneity variable pairs on statistical qualities of census output areas using the 2011 SALs as building blocks in Free State province at all spatial levels. The homogeneity variable pairs were: dwelling type and geotype; tenure type and geotype; dwelling type and tenure type; and all three homogeneity variables together.

Results highlighted that statistical qualities of AZTool output areas developed using different combinations of homogeneity variable pairs were slightly similar in terms of population means and shape compactness. The statistical characteristics differed when it comes to degree of homogeneity. Figure 1 shows that tenure type and geotype homogeneity variable pair had higher IAC scores than all the other variable pairs at all spatial levels. The dwelling type and geotype homogeneity variable pair became second, and then all three homogeneity variable pair and lastly dwelling type and tenure type. The dwelling type and tenure type homogeneity variable pair had very low IAC scores; hence if the social homogeneity is one of the design criteria, this pair could not be used. For example, at provincial level, the pair resulted in output areas that were almost three times less homogeneous and two times less homogeneous than the ones from tenure type and geotype and all three homogeneity variable pairs, respectively.

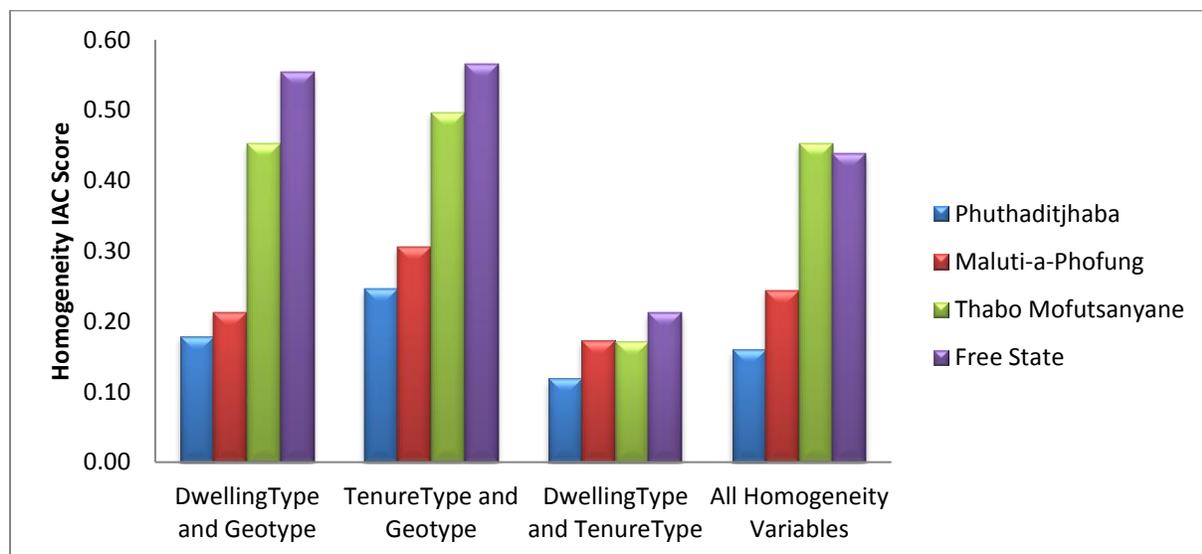


Figure 1: Different homogeneity variable pairs' IAC scores for AZTool output areas in Free State

4 Discussion

Findings of this study highlight that different building blocks do have an impact on the statistical qualities of the AZTool optimised output areas. Generally, all output areas from the three different building blocks adhered to the confidential limit of 500 persons; this is a huge success from personal privacy perspective. When the EAs and the SALs were used as building blocks in all study areas, statistics showed that output areas from the EAs had slightly higher population means and lower standard deviations than the ones from the SALs. However, the means from the two sets of optimised output areas were close to user-defined population target mean. Clearly, this highlights that the EAs output areas had slightly broader population distributions than their counter-parts. This might be due to the fact that EAs had maximum population of 9269 and low population average of 519 compared to 6701 and 782, respectively, of the SALs. This shows that the AZTool program had to do more effort to bring the mean value of 519 to the target mean of 1000 than it was for 782 to 1000. Unsurprisingly, the output areas from the SubPlaces had higher population means and standard deviations than the ones from the two sets of buildings blocks at all levels in both rural and urban settings. This was expected as the SubPlaces are much bigger in size than the two sets of building blocks and the two sets nest within the these SubPlaces in the South African geography hierarchy.

With regard to shape compactness, the optimised output areas generated from the EAs and the SALs were almost similar at all study areas. The output areas from the SubPlaces were the less compact compared to the ones from EAs and the SALs at all spatial levels in both rural and urban settings. However, the P2A mean values and standard deviations of the output areas from the SubPlaces in urban areas were not as high as they were in rural areas; in fact they were close to the mean values and standard deviations of output areas from the EAs and the SALs. Therefore the effects of different building blocks on the shape characteristics of the AZTool output areas tend to be noticed more in the rural areas, especially between lower level building blocks (EAs and SALs) and higher level building blocks (SubPlaces). Findings from all the three different building blocks further showed that the AZTool optimised output areas from urban areas were more compact than their counter-parts at all levels of geography. In support of these findings, Cockings

et al. (2013) discovered that output areas in rural areas (Isle of Anglesey) were less compact than those in the urban areas (Camden, Manchester).

For degree of homogeneity, only the AZTool optimised output areas from the EAs and SubPlaces yielded reasonable results. The SALs ones did not have enough homogeneity variables hence did not produce reasonable output results. Therefore 2011 SALs were explored, but only for the Free State province as it did not change its provincial boundaries from 2001. In general, the output areas created using the EAs as building blocks were more homogeneous than those created from the SubPlaces in both rural and urban settings. Few exceptions were found in rural areas where output areas from the EAs at lower geographic levels (Mainplace and municipality) were less homogenous than the ones from the SubPlaces.

In terms of homogeneity at the SAL level, findings from the Free State province (only at provincial level) showed that the AZTool output areas created using 2011 SALs as building blocks were less homogeneous than the ones from EAs but more homogenous than those from the SubPlaces. This is indicative that the AZTool output areas generated from smaller areas tend to be more homogeneous than the ones generated from larger areas when using dwelling type and geotype as homogeneity variables. Similarly, it was found that there was a tendency for smaller areas to capture more between-neighbourhood variations than larger areas, hence clustering appeared to be most marked at the very local scale (Haynes *et al.*, 2008).

In addition to measuring the degree of homogeneity, IAC scores could also be used as an assessment of magnitude of the scale effect because the IAC scores are adjusted for population size (Manley *et al.*, 2006; Flowerdew, 2011). Generally, the higher IAC scores indicate the higher scale effects. The higher IAC scores produced by output areas from smaller areas indicate that scale effects are clearly identified when smaller areas are used as building blocks than when larger areas are considered. This also supports arguments by previous studies such as (Openshaw, 1984; Cockings *et al.*, 2013) that the scale effect of the MAUP is generally greater than the zonation effect.

In general, the AZTool output areas from the SubPlaces had higher population means and higher standard deviations than those developed from the EAs and the SALs at all levels in both rural and urban areas. This shows that the SubPlaces are not ideal building blocks from user's perspective as comparisons of individual areas in terms of population size is not possible. In addition, the output areas from the SubPlaces were less compact in shape and less homogenous than the output areas from their counter-parts.

When looking at different combinations of homogeneity variable pairs, it was found that tenure type and geotype homogeneity variable pair and the dwelling type and geotype homogeneity variable pair made it more possible to identify scale effects than all three homogeneity variable pair and the dwelling type and tenure type. The dwelling type and tenure type homogeneity variable pair had very low IAC scores which indicated that output areas were less heterogeneous between each other hence low scale effect.

Among limitations to this study was the accessibility of data at lower levels from Stats SA. The accessibility of census data at household level could have allowed the exploration of other building blocks design such as grid squares which were found to minimize the effect of MAUP in France by Sabel *et al.* (2013). In addition, the 2011 census data at the SAL level excluded zero-populated areas; therefore this resulted in 15 isolated building blocks being picked by the AZTool program in Free State province. They were excluded for further analysis as the AZTool works with contiguous building blocks. Even though these isolated building blocks constituted only 0.15% of the total population of Free State province, they might have some slight contribution on the statistical characteristics of the AZTool output areas generated using the 2011 SALs as building blocks.

5 Conclusions

It was concluded that based on results from this study, different building blocks did have an impact on the statistical qualities of the AZTool optimised output areas in both rural and urban settings in South Africa. Although the output areas from the smaller areas (EAs and SALs) were almost similar, they differed slightly. The output areas generated from the EAs were slightly more

homogeneous while the ones from the SALs had slightly narrower population distributions. Therefore, between these two building blocks, a choice would depend on the user needs. For instance, for better allocation of resources and prioritising targeting of the population that is the most in need, the EAs would be better suited as building blocks than the SALS. However, the latter would be ideal for statistical analysis as individual areas are comparable in terms of population size. The SubPlaces on the other hand are not ideal to be used as building blocks as they produced output areas with higher population means and higher standard deviations than those developed from the EAs and the SALs at all levels in both rural and urban areas. In addition, the output areas generated from the SubPlaces were less compact in shape and less homogenous than the output areas from their counter-parts. From these findings, it was also concluded that that the AZTool output areas generated from smaller areas tend to be more homogeneous than the ones generated from larger areas when using dwelling type and geotype as homogeneity variables. In addition, the AZTool optimised output areas from the smaller areas allowed a clear distinction of the scale effects than output areas from larger areas. The accessibility of census data at lower levels such as household would have allowed exploration of other building blocks such as grid squares. Therefore, it is recommended that such data should be made accessible even if it is under secure condition for future research.

6 Acknowledgements

The authors would like to greatly thank Prof David Martin and Dr Samantha Cockings for their permission to use the AZTool software, which is copyright David Martin, Samantha Cockings and University of Southampton. This paper is part of the doctoral thesis work that was submitted to the University of KwaZulu-Natal.

7 References

- Cockings, S. and Martin, D. (2005). Zone design for environment and health studies using pre-aggregated data. *Social Science and Medicine*, 60, 2729 – 2742.
- Cockings, S., Harfoot, A., Martin, D. and Hornby, D. (2011). Maintaining existing zoning systems using automated zone-design techniques: methods for creating the 2011 census output geographies for England and Wales. *Environment and Planning A*, 43, 2399 – 2418.

- Cockings, S., Harfoot, A., Martin, D. and Hornby, D. (2013). Getting the foundations right: spatial building blocks for official population statistics. *Environment and Planning A*, 45, 1403 – 1420.
- Dumedah, G., Schuurman, N. and Yang, W. (2008). Minimizing effects of scale distortion for spatially grouped census data using rough sets. *Journal of Geographical Systems*, 10, 47 – 69.
- Duque, J.C., Ramos, R. and Surinach, J. (2007). Supervised regionalization methods: A survey. *International Regional Science Review*, 30 (3), 195 – 220.
- Flowerdew, R. (2011). How serious is the Modifiable Areal Unit Problem for analysis of English census data? *Population Trends*, 145, 106 – 118.
- Gehlke, C.E. and Biehl, K. (1934). Certain effects of grouping upon the size of the correlation coefficient in census tract material. *Journal of the American Statistical Association*, 29, 169 – 170.
- Haynes, R., Daras, K., Reading, R., and Jones, A. (2007). Modifiable neighbourhood units, zone design and residents' perceptions. *Health and Place*, 13, 812 – 825.
- Haynes, R., Jones, A., Reading, R., Daras, K. and Emond, A. (2008). Neighbourhood variations in child accidents and related child and maternal characteristics: does area definition make a difference? *Health and Place*, 14, 693 – 701.
- Heywood, I., Cornelius, S. and Carver, S. (2002). *An introduction to Geographical Information Systems*. 2nd ed. Pearson Prentice Hall, London.
- HSRC, (2005). 2001 census EA estimates. Human Sciences Research Council in collaboration with Prof DJ Stoker. Pretoria, South Africa.
- Kitchin, R. and Tate, N.J. (2000). *Conducting research in human geography: Theory, methodology and practice*. Pearson Prentice Hall, London.
- Laldaparsad, S. (2007). Census mapping and the use of geo-spatial technologies (A case of South Africa). *United Nations Expert Group Meeting on Contemporary Practices in Census Mapping and Use of Geographical Information Systems*, 29 May – 1 June 2007, New York.
- MacEachren, A. (1985). Compactness of geographic shape: comparison and evaluation of measures. *Geografiska Annaler B*, 67, 53 – 67.

- Manley, D., Flowerdew, R. and Steel, D. (2006). Scales, levels and processes: studying spatial patterns of British census variables. *Computers, Environment and Urban Systems*, 30, 143 – 160.
- Martin, D., Nolan, A. and Tranmer, M. (2001). The application of zone-design methodology in the 2001 UK Census. *Environment and Planning A*, 33, 1949 – 1962.
- Mokhele, T., Mutanga, O. and Ahmed F. (2016). Development of census output areas with AZTool in South Africa. *South African Journal of Science*, 112, 7/8, 1 – 7.
- Openshaw, S. (1977). A geographical solution to scale and aggregation problems in region-building, partitioning and spatial modelling. *Transactions of the Institute of British Geographers, NS*, 2 (4), 459 – 472.
- Openshaw, S. (1984). Problem concepts and techniques in modern geography: The Modifiable Areal Unit. *CATMOG* 38, 1 – 39.
- Ralphs, M. and Ang, L. (2009). Optimized geographies for data reporting: Zone design tools for census output geographies. *Statistics New Zealand Working Paper No 09–01*. Statistics New Zealand, Wellington.
- Ratcliffe, J.H. and McCullagh, M.J. (1999). Hotbeds of crime and the search for spatial accuracy. *Journal of Geographical Systems*, 1, 385 – 398.
- Reynolds, H.D. (1998). *The Modifiable Unit Problem: Empirical analysis by statistical simulation*, Unpublished PhD Thesis. University of Toronto, Canada.
- Sabel, C.E., Kihal, W., Bard, D. and Weber, C. (2013). Creation of synthetic homogeneous neighbourhoods using zone design algorithms to explore relationships between asthma and deprivation in Strasbourg, France. *Social Science and Medicine*, 91, 110 – 121.
- Stats SA, (2003). Census 2001. How the count was done. Statistics South Africa, South Africa.
- Stats SA and HSRC, (2007). *Using the 2001 census: Approaches to analysing data*. A collaboration between Statistics South Africa (Stats SA) and the Human Sciences Research Council (HSRC), 244 p, South Africa.
- Tranmer, M. and Steel, D. (1998). Using census data to investigate the causes of the ecological fallacy. *Environment and Planning A*, 30, 817 – 831.

- Tranmer, M. and Steel, D. (2001). Using local census data to investigate scale effects. In Tate N.J. and Atkinson P.M. (eds). *Modelling scale in geographical information science*. John Wiley and Sons, Chichester. pp. 105 – 122.
- Valente, P. (2010). Census taking in Europe: How are populations counted in 2010? *Population and Societies*, 467. Available from:
http://www.ined.fr/fichier/s_rubrique/19135/pesa467.en.pdf [Accessed 20 February 2015].
- Verhoef, H. and Grobbelaar, N. (2005). The development of a Small Area Layer for South Africa for census data dissemination. Statistics South Africa. Available from:
<http://www.cartesia.org/geodoc/icc2005/pdf/poster/TEMA26/HELENE%20VERHOEF.pdf> [Accessed 19 November 2009].