South African Journal of Geomatics, Vol. 11. No. 1, February 2022

The statistical qualities of the zone design census output areas

T Mokhele[1], O Mutanga[2] and F Ahmed[3]

[1] Geospatial Analytics, eResearch Knowledge Centre, Human Sciences Research Council, South Africa, TAMokhele@hsrc.ac.za
[2] School of Agriculture, Earth and Environmental Sciences, University of KwaZulu-Natal, South Africa
[3] School of Geography, Archaeology and Environmental Studies, University of Witwatersrand, South Africa

DOI: http://dx.doi.org/10.4314/sajg.v11i1.1

Abstract

The statistical qualities of census output areas are of great importance especially when the purpose of output areas is to understand the statistical properties of the population rather than mapping. If the purpose of creating census output areas is solely for displaying results in a map format, shape compactness of output areas is prioritised. In that case, other statistical characteristics such as population, population mean and social homogeneity are often ignored. This paper explored the statistical qualities of the Automated Zone-design Tool (AZTool) generated census output areas using the 2001 census Enumeration Areas (EAs) as building blocks in South Africa. The statistical qualities were mainly based on population target mean, minimum population threshold, social homogeneity as well as shape compactness. The homogeneity variables that were selected from the 2001 census data were dwelling type and geotype. The results showed that the AZTool generated output areas substantially outperformed the original EAs and Small Area Layers (SALs) in terms of the minimum population threshold and population distribution statistical qualities. It is worth noting though that the AZTool output areas were less compact and homogeneous than the original EAs in both urban and rural settings. The fact that a minimum population threshold of 500 was respected by the AZTool output areas in both rural and urban settings was a huge success from confidentiality point of view. It was concluded that the AZTool could be utilized to produce robust and high-quality optimised output areas for population census dissemination in South Africa.

Keywords: *AZTool; Census; Enumeration areas; Output areas; South Africa.*

1. Introduction

The statistical qualities of census output areas are of great importance especially when the purpose of output areas is to understand the statistical properties of the population rather than mapping only. In this study, statistical qualities are based on the characteristics of output areas regarding their shape, social homogeneity and population targets. For instance, if the purpose of creating census output areas is solely for displaying results in a map format, shape compactness of output areas is prioritised. In

that case, other statistical characteristics such as population, population mean and social homogeneity are often ignored.

Automated Zone-design Tool (AZTool) software has been utilized to produce robust and high-quality optimised output areas where population targets, social homogeneity and shape compactness can be pre-defined. The AZTool program works by iteratively combining and recombining sets of building blocks to create output areas which optimise a set of pre-specified design criteria (Cockings *et al.,* 2011; Sabel *et al.,* 2013; Mokhele *et al.* 2016). It was developed by Cockings, Martin and Harfoot at the University of Southampton in 2006. Further details on the history of the AZTool can be found in Mokhele *et al.* (2016).

Applications of the AZTool software are well described in the following references (Flowerdew *et al.,* 2008; Ralphs and Ang, 2009; Cockings *et al.*, 2011; 2013; Martin *et al.*, 2013; Sabel *et al.*, 2013; Mokhele *et al.*, 2016; 2017). For instance, Cockings *et al.* (2011) employed the AZTool to modify the 2001 Census output geographies within six local authority districts in England and Wales in order to make them suitable for the release of contemporary population-related data. This was done such that zones that still meet the design criteria were retained while those that were no longer fit for purpose were split or merged. The use of the AZTool for maintenance of an existing system was found to be a more iterative and constrained problem than designing a completely new system; design constraints frequently had to be relaxed and manual intervention was occasionally required (Cockings *et al.,* 2011). In addition, their findings suggested that it would be easier to resolve under-threshold zones than over-threshold zones.

Martin *et al.* (2013) further explored the application of the AZTool for creating workplace zones (WZ) with England and Wales 2001 census microdata. They found that the prototype areas displayed much improved statistical properties, with more uniform sizes of workforce, less extreme values and compliance by design with the specified threshold values. Their results further showed that there was a small number of WZs which could not be automatically resolved by using the parameters evaluated in their study. The reason being either no suitable neighbouring zones were available for merging or their constituent postcodes were inappropriately configured. Their approach was further adopted or incorporated in England and Wales 2011 census output plans.

None of these studies strictly focused on the statistical quality of the created optimised output areas or zones except the one by Ralphs and Ang (2009). They attempted to determine statistical quality of automatically developed geographies by comparing them with existing official geographies in New Zealand. They found that the automatically generated geographies substantially outperformed the existing geographies across almost all of their optimisation criteria. For instance, the automatically created geographies effectively satisfied minimum and target population thresholds, while the population distributions were much narrower in range than the existing reporting geographies. Therefore, this paper aimed to determine the statistical qualities of the AZTool generated census output areas using South African Enumeration Areas (EAs) as building blocks. Enumeration Areas (EAs) are smallest geography units used for census data collection in South Africa. The EAs typically contain between 100 and 250 households, do not overlap, have boundaries that can be identified on

the ground, and are of approximately equal population size to enable an enumerator to cover each unit within the census period.

2. Methods

Two out of the nine provinces in South Africa were selected for this study (Mokhele *et al.*, 2016; 2017). These were Free State and Gauteng provinces which were representative of rural and urban areas respectively. To get a better picture of the statistical qualities of the AZTool output areas at different geographic levels (the district, municipality and mainplace levels) were also analysed.

The 2001 census estimates data developed by HSRC (2005) were used to get data at the EA level as the original data was not accessible at this level from Statistics South Africa (Stats SA). The data for the two provinces that were extracted from these census data include total population, homogeneity variables as well as different spatial level boundaries. The homogeneity variables that were selected from the 2001 and 2011 census data are dwelling type and geotype. The dwelling type, also known as housing type, is the commonly used variable as proxy for social built environment homogeneity measure (Martin *et al.*, 2001; Ralphs and Ang, 2009) while the geotype (geographic type) has been used as a homogeneity rule for development of SAL which was used to disseminate the 2001 census data in South Africa (Verhoef and Grobbelaar, 2005; Mokhele *et al.*, 2016).

The EAs from the 2001 census data were used as building blocks for the development of optimised census output areas using the AZTool version 1.0.3 (Cockings *et al.*, 2011). The minimum population threshold, population target, shape and homogeneity criteria were pre-defined in the creation of these optimised output areas. A minimum population of 500 and a population target of 1000 were set (Verhoef and Grobbelaar, 2005; Mokhele *et al.*, 2016; 2017). For homogeneity, this study employed the Intra-Area Correlation (IAC) while Perimeter Squared per Area (P2A) was used as a measure of shape compactness (Mokhele *et al.*, 2016; 2017). Further statistical analyses such as Analysis of Variance (ANOVA) and Shapiro-wilk test were performed in Statistical Package for Social Sciences (SPSS).

3. Results

Figure 1 highlights the comparison of the original EAs used as building blocks with the AZTool census output areas in Phuthaditjhaba. Figure 1a shows that there was a significant number of areas that had less than 500 people. The original EAs population distribution also had large population range which means it could not be easy to compare individual areas based on population size. The higher variance further indicates that the original EAs had broader population distribution compared to the optimised AZTool output areas. In addition, the population means of the AZTool output areas were closer to the target mean of 1000 with lower standard deviations compared to the original EAs (Figure 1b). This indicates that the output areas had much narrower and tighter population distributions than their counterparts. The confidentiality limit of 500 people was also not breached

for output areas, which is a success from confidentiality point of view. This was further proven statistically by running Shapiro-wilk test which showed that the population distribution for the AZTool output areas was normal ($p > 0.05$) while for the counterpart it was not normal ($p < 0.05$).

To depict the general picture at the urban settings, a similar population distribution figure was displayed for Pretoria (Figure 2). This figure shows that similar trends to those of the rural areas were experienced. The AZTool output areas respected the confidentiality limit and had much tighter population distributions (Figure 2b). It is important to highlight that none of these population distributions was normal as the Shapiro wilk test revealed significant ($p < 0.05$) results in both cases.

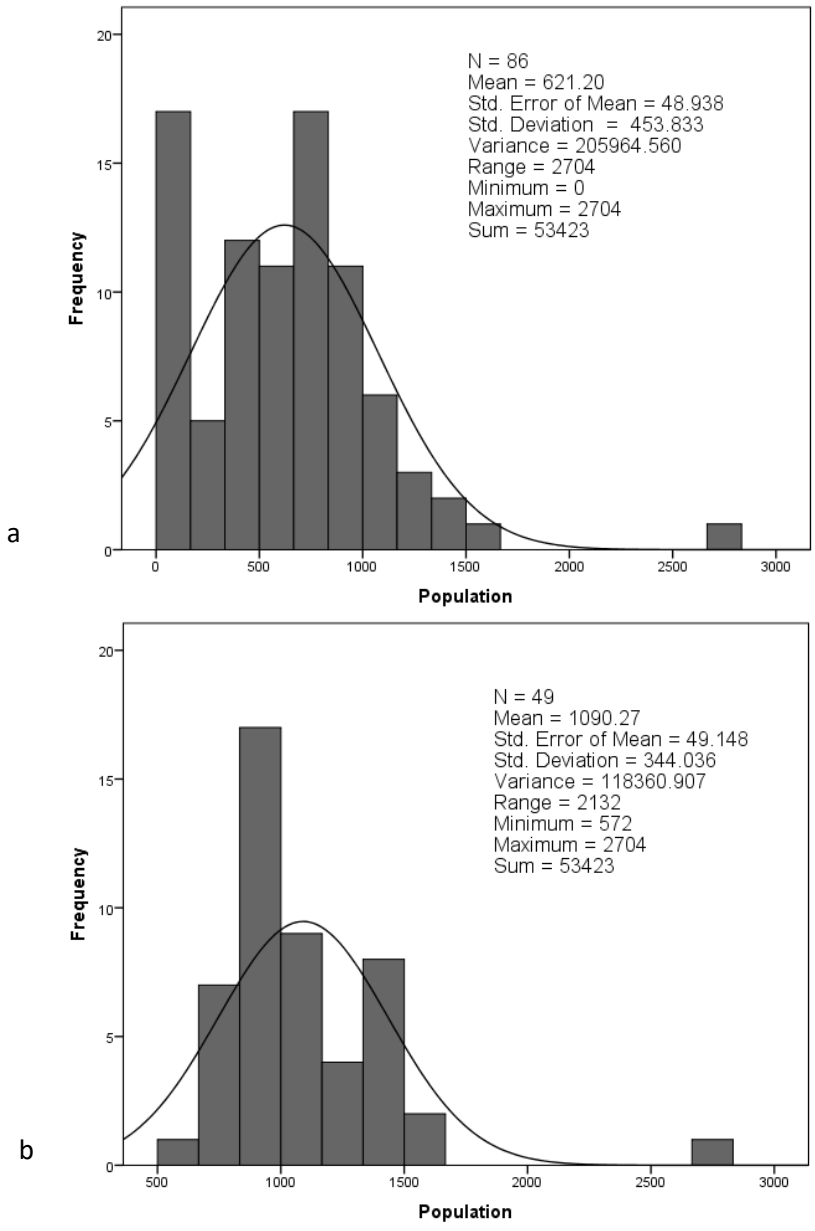

Figure 1. Population distribution for a) the original EAs and b) the AZTool census output areas for Phuthaditjhaba mainplace

The results showed that confidentiality was adhered to at all geographical levels in the AZTool output areas in both rural and urban areas compared to the original EAs where it was breached at all spatial levels. However, these newly created AZTool output areas had higher shape mean at all geographical levels indicating that they were slightly less compact compared to the original EAs in both rural and urban settings.

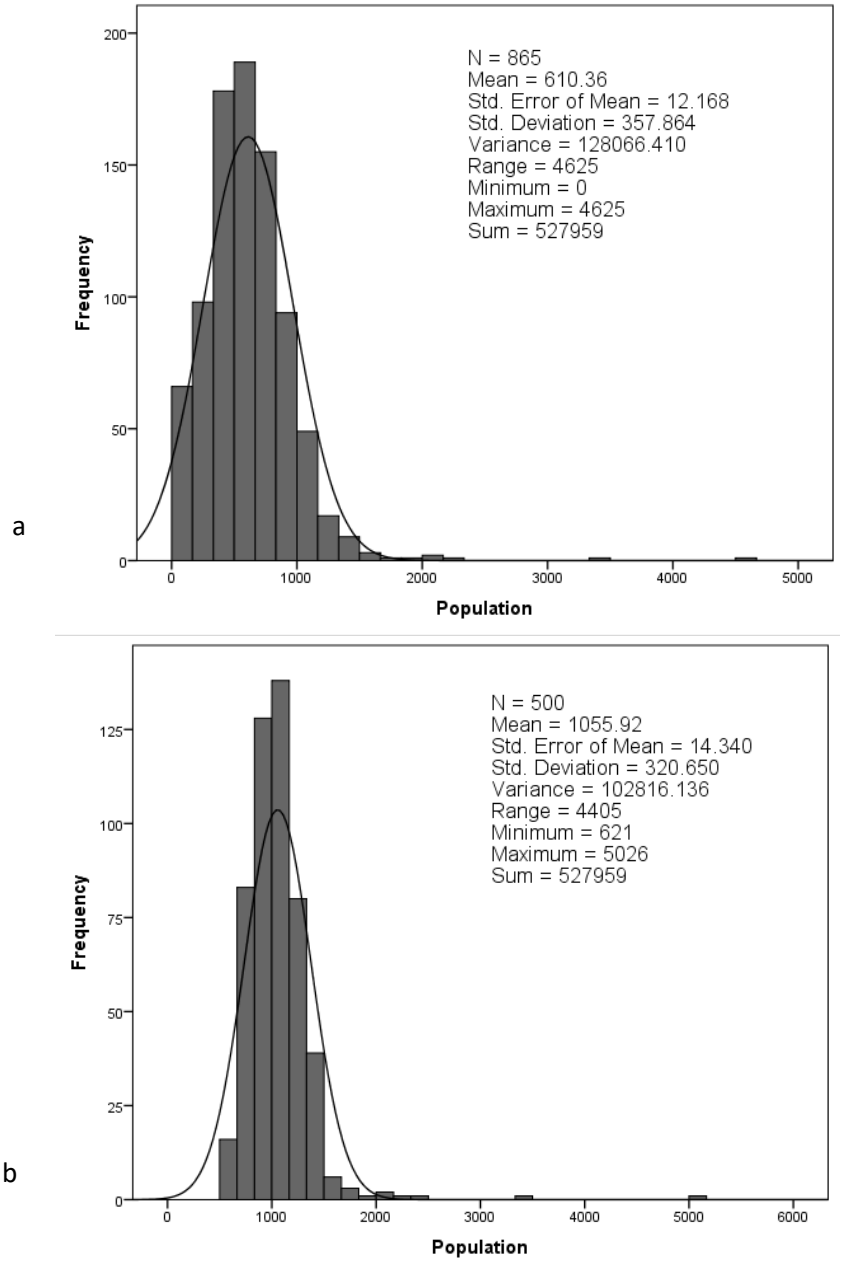

Figure 2. Population distribution for a) the original EAs and b) the AZTool census output areas for Pretoria mainplace

A further test was performed to see if increased number of the AZTool runs would improve statistical characteristics of output areas at the district level in both rural and urban areas. The results showed that increasing number of runs did not improve statistical qualities of optimised output areas in all areas (see Tables 1 and 2).

Table 1. Statistical outputs of Thabo Mofutsanyane district with different runs

Number	Output	Population					Shape			Homogeneity
of Runs	Areas	Min	Max	Mean	SD	Score	Mean	SD	Score	IAC
10	667	581	5292	1087	403	113616375	33	13	21695	0.56
20	667	516	5292	1087	404	113921055	32	13	21430	0.56
30	678	587	5364	1070	404	113876601	32	13	21670	0.56
40	676	527	5292	1073	403	113479469	32	12	21389	0.56
50	672	610	5364	1079	401	112337947	32	12	21633	0.56
100	669	581	5292	1084	401	112260839	32	12	21263	0.56
500	663	597	5292	1094	403	113704181	32	13	21364	0.56
1000	676	578	5364	1073	399	111041831	32	12	21593	0.56

Table 2. Statistical outputs of Tshwane district with different runs

Number	Output	Population					Shape			Homogeneity
of Runs	Areas	Min	Max	Mean	SD	Score	Mean	SD	Score	IAC
10	1276	502	8802	1203	514	389103794	27	10	33940	0.46
20	1273	517	8802	1205	514	390442028	26	10	33623	0.46
30	1262	517	8802	1216	507	383577766	27	10	33682	0.46
40	1267	507	8802	1211	509	385210614	27	10	33750	0.46
50	1265	517	8802	1213	512	388941006	27	11	33581	0.46
100	1271	502	8802	1207	512	387293712	27	10	33732	0.46
500	1273	517	8802	1205	506	379605664	27	10	33799	0.46
1000	1281	502	8802	1198	506	377658462	27	11	33970	0.46

Different weights for homogeneity, population target and shape were also explored to see their statistical effects on the output areas. For instance, when homogeneity weight was set to the weight of 200, 300, 400, 500, and 1000 respectively, the other two (population and shape weights) were left at default weight of 100 and vice versa. Figure 3 shows that different shape weights make a substantial improvement on the shape measure of the output areas. There is clear evidence that when the shape (P2A) weight increases, the shape measure decreases, resulting in more compact output areas. For instance, when the shape weight increased from 100 – 1000, the P2A measure decreased from 1340 – 664.

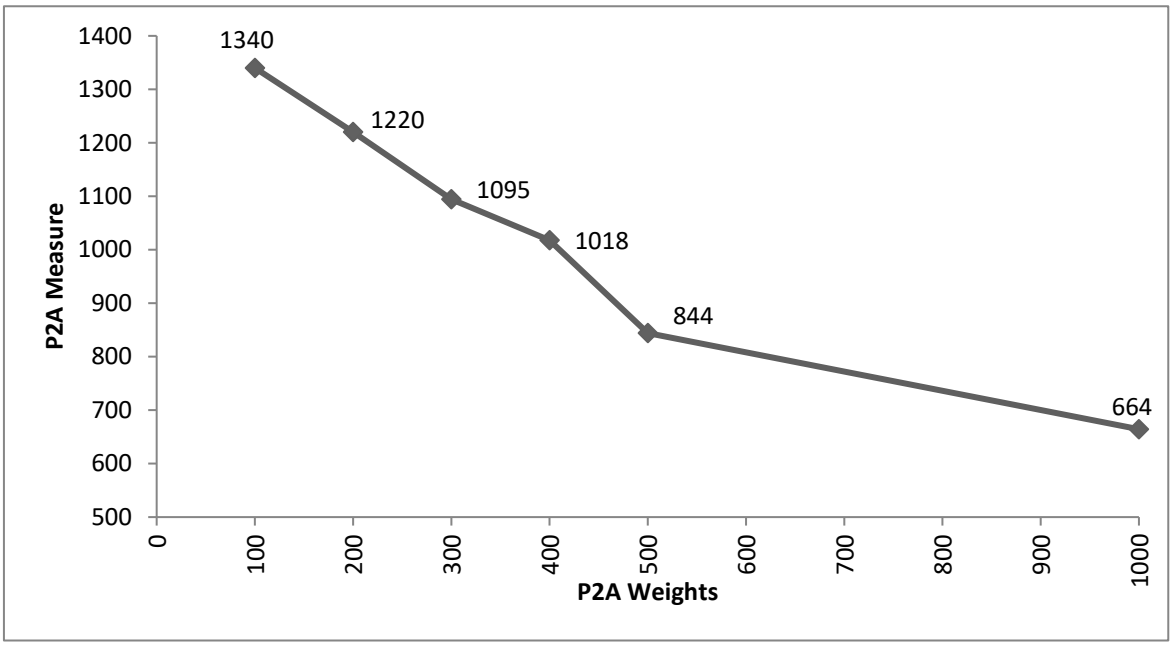

Figure 3. Effects of different shape weights on the P2A measure of output areas for Phuthaditjhaba mainplace

Effects of different population weights on the population characteristics of the AZTool output areas were also explored for Phuthaditjhaba. Figure 4 highlights that both minimum and maximum population did not change when different population weights were applied. The population target means changed a bit but were also constant after population weights of 500 and1000 were considered.

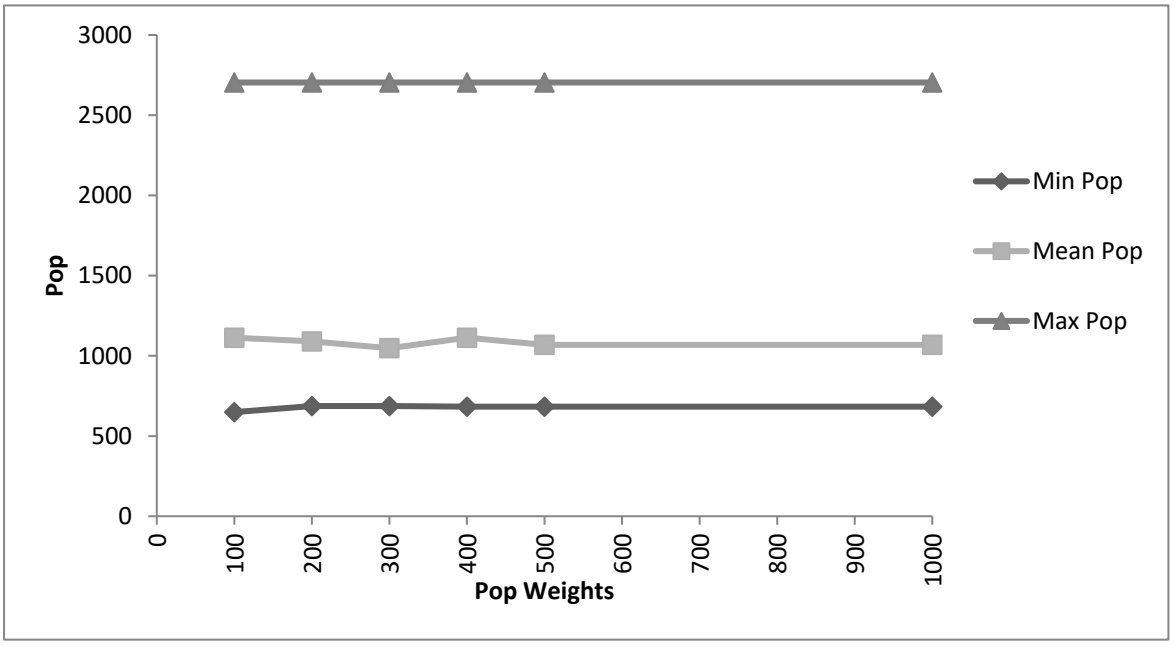

Figure 4. Effects of different population weights on the population characteristics of the AZTool output areas for Phuthaditjhaba mainplace

Figure 5 shows the impact of different shape weights on the AZTool optimised output areas for Phuthaditjhaba. Clearly, the visual displays highlight that there is improvement from Figure 5a (original EAs) to Figure 5b (output areas with shape weight of 100) in terms of shape compactness. The shape weights of 500 and 1000 show even more compact shapes (Figures 5c and d). This indicates that, if the priority to have more compact output areas, especially for mapping, different weights could be applied for Phuthaditjhaba, especially higher weights. It is noteworthy that this application of higher shape weights would come at a compromise of other design criteria such as population target and social homogeneity.

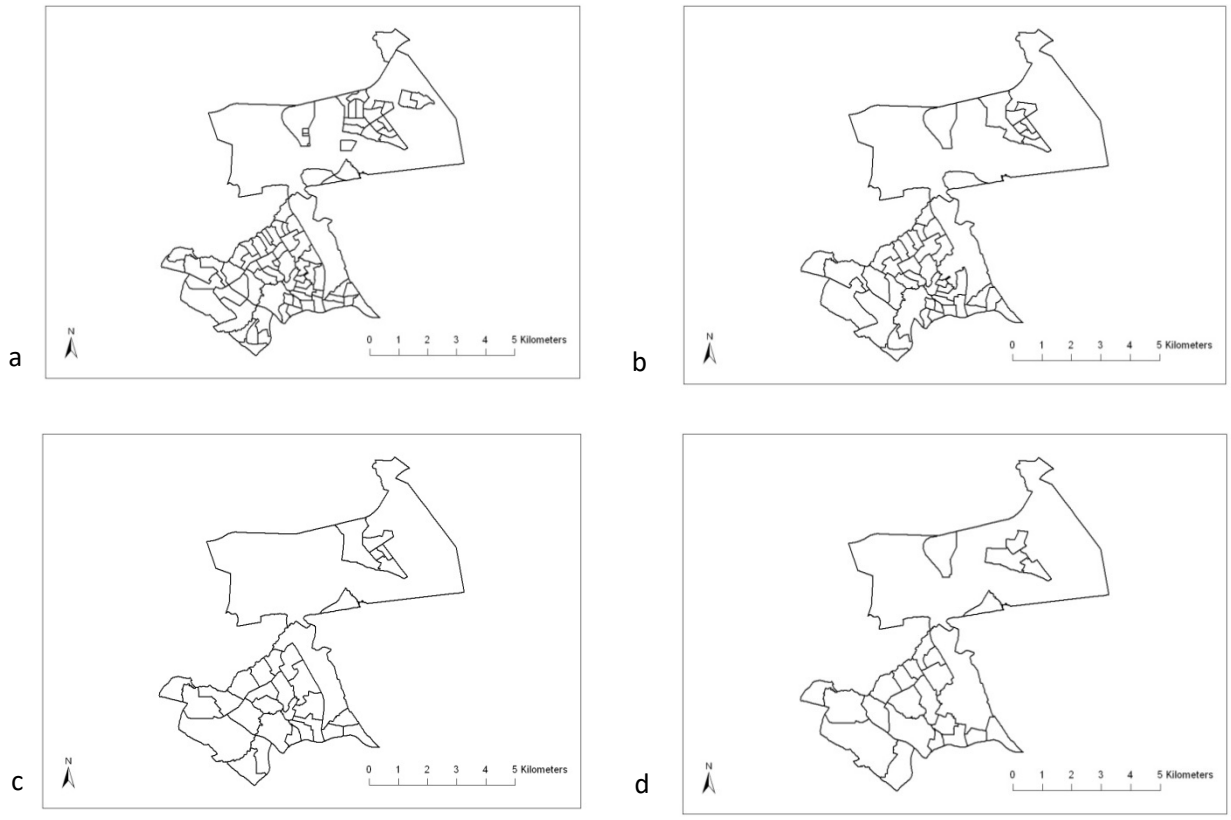

Figure 5. Phuthaditjhaba mainplace a) original EAs, b) P2A100weight, c) P2A500weight, and d) P2A1000weight output areas

The 2011 census data was released at the SAL level, however there was a significant number of areas that were below the official minimum threshold of 500 people, especially in Free State whereby almost half (42.2%) of the areas had below 500 people compared to around 27% in Gauteng. Therefore, the SALs from the 2011 census data were also used as building blocks in an effort to further determine statistical qualities of the AZTool generated output areas. The same criteria set for the generation of output areas using the EAs were employed. The results highlight that the AZTool output areas substantially outperformed the original SALs with regard to confidentiality as none of the output areas were below the 500 minimum population thresholds (Table 3). In addition, the population means of the output areas were closer to the set population target of 1000 than the ones of the original SALs at all spatial levels. Hence the output areas had tighter population distribution than the original SALs. The output areas were less compact compared to the SALs at all spatial levels as

they had significantly ($p < 0.05$) higher P2A means than their counterparts. Regarding homogeneity, the SALs produced results at higher level (provincial level) only. Hence only this level could be compared with IAC score for the optimised output areas. Results also highlight that the optimised output areas were less homogeneous than the original SALs.

Table 3. Statistical characteristics of the original SALs and the AZTool generated output areas at all levels in the Free State province

	Number	Population				Shape		Homogeneity
	of Zones	Min	Max	Mean	SD	Mean	SD	IAC
SALs								
Phuthaditjhaba	105	42	1065	521	128	25	8	N/A
Maluti-a-Phofung	729	15	1080	460	122	27	9	N/A
Thabo Mofutsanyane	1513	9	1326	486	167	26	8	N/A
Free State	5114	9	5586	536	228	25	9	0.62
Output Areas								
Phuthaditjhaba	51	639	1677	1072	210	33	14	0.18
Maluti-a-Phofung	334	642	1563	1005	166	35	13	0.21
Thabo Mofutsanyane	721	612	1674	1021	188	33	12	0.45
Free State	2596	594	5586	1056	264	31	11	0.55

4. Discussion

The results showed that confidentiality was largely adhered to at all geographical levels in the AZTool output areas in both rural and urban areas compared to the original EAs where the minimum population was zero at all geographic levels. Census data or national statistics must be released at level where disclosure of personal information of individuals, households, or organisations is avoided by all means, even if other systems such as registers or any administrative datasets are used to collect these data (Valente, 2010; Cockings *et al.*, 2011; Flowerdew, 2011). Furthermore, the AZTool optimised output areas had much narrower and tighter population distributions than the original EAs. This was further proven statistically by Shapiro-wilk test results which showed that the population distribution for the AZTool output areas was normal ($p > 0.05$) whereas for the one of the EAs was not normal ($p < 0.05$). However, these newly created AZTool output areas had higher shape mean at all geographical levels indicating that they were statistically ($p < 0.05$) slightly less compact compared to the original EAs in both rural and urban settings. This shows that a compromise had to be considered at some point (Ralphs and Ang, 2009; Cockings and Martin, 2005; Drackley *et al.*, 2011).

Findings from this study also showed that different shape weights had a great improvement on the visual display of the output areas. This was proven by the fact that when the criterion for the shape was set to carry ten times more weight than population and homogeneity, the shapes of output areas were more circular and less elongated. It is noteworthy that this application of higher shape weights would of course come at a compromise of other design criteria such as population target and social

homogeneity. No previous studies which reported on direct impact of different AZTool weights on the statistical qualities of the optimised output areas were found for comparative purposes.

In addition, when the 2011 census data was explored, the results highlighted that the AZTool output areas substantially outperformed the original SALs with regard to confidentiality as none of the output areas were below the 500 minimum population thresholds. The population means of the output areas were closer to the set population target of 1000 than the ones of the original SALs at all spatial levels. Hence, the AZTool optimised output areas had tighter population distribution than the original SALs (Ralphs and Ang, 2009; Martin *et al.,* 2013). The output areas were less compact compared to the SALs at all spatial levels. Regarding homogeneity, the SALs produced results at higher level (provincial level) only. Hence only this level could be compared with IAC score for the optimised output areas. Results also showed that the output areas were less homogeneous than the SALs.

The fact that homogeneity of output areas can be specified for this tool means that areas with similar socio-economic and socio-demographic characteristics can be grouped to together for to form an output area. This means that there can be better allocations of resources by government as the output areas will not be mixture of rich and poor residents. Hence this tool may be used for spatial planning, transformation and equity in the context of South Africa.

The findings from this study have a potential to influence policy and practice of government stakeholders, such as Stats SA, for future census disseminations. Stats SA is the official National Statistics Central Office in South Africa. The lead author in this paper had already trained some Stats SA officials on the AZTool applications in the creation of census output areas in South Africa. Stats SA had started exploring the possibilities of using this AZTool for 2021 census disseminations. There was a seminar where Stats SA official presented their intentions of exploring this AZTool for 2021 census.

5. Conclusions

It was further proven that the AZTool generated output areas substantially outperformed the original EAs and the SALs in terms of minimum population threshold and population distribution statistical qualities. To substantiate this, Shapiro-wilk test results showed that the population distribution for the AZTool output areas was normal ($p > 0.05$) whereas for the one of the EAs was not normal ($p < 0.05$). However, the AZTool output areas were less compact and homogeneous than the original EAs in both urban and rural settings. The fact that confidentiality limit of 500 persons was respected by the AZTool output areas in both rural and urban settings was a huge success from a confidentiality point of view. Results further showed that different shape weights had a great improvement on the visual display of the AZTool output areas. For instance, when the criterion for the shape was set to carry ten times more weight than population and homogeneity, the shapes of output areas were more circular and less elongated. It was concluded that the AZTool could be utilized to produce robust and high-quality optimised output areas for population census disseminations in

South Africa. However, a compromise had to be taken when setting the criterion based on the purpose the output areas would be utilised for.

6. Acknowledgements

Great thanks to Profs David Martin and Samantha Cockings for their permission to use the AZTool software, which is copyright David Martin, Samantha Cockings and University of Southampton. This article was extracted from 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*, vol. 60, pp. 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*, vol. 43, pp. 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*, vol. 45, pp. 1403 – 1420.

Drackley, A, Newbold, KB, and Taylor, C 2011 'Defining socially-based spatial boundaries in the Region of Peel, Ontario, Canada' *International Journal of Health Geographics*, vol. 10, no. 38, pp. 1 – 12.

Flowerdew, R, Manley, DJ and Sabel, CE 2008, 'Neighbourhood effects on health: Does it matter where you draw the boundaries?', *Social Science and Medicine*, vol. 66, pp. 1241 – 1255.

Flowerdew, R 2011, 'How serious is the Modifiable Areal Unit Problem for analysis of English census data?', *Population Trends*, vol.145, pp. 106 – 118.

HSRC 2005, *2001 census EA estimates*, Human Sciences Research Council in collaboration with Prof DJ Stoker, Pretoria, South Africa.

Martin, D, Nolan, A and Tranmer, M 2001, 'The application of zone-design methodology in the 2001 UK Census', *Environment and Planning A,* vol. 33, pp. 1949 – 1962.

Martin, D, Cockings, S and Harfoot, A 2013, 'Development of a geographical framework for census workplace data', *Journal of Royal Statistical Society*, vol. 176, no. 2, pp. 1 – 18.

Mbogoni, M 2012 'Report of the United States of America on the 2010 World Programme on population and housing censuses', *United Nations International Seminar on population and housing censuses: Beyond the 2010 Round,* Republic of Korea, 27 - 29 November 2012.

Mokhele, T, Mutanga, O and Ahmed, F 2017, 'Effects of different building blocks designs on the statistical characteristics of Automated Zone-design Tool output areas', *South African Journal of Geomatics*, vol. 6, no. 2, pp. 155 - 171.

Mokhele, T, Mutanga, O and Ahmed, F 2016, 'Development of census output areas with AZTool in South Africa', *South African Journal of Science*, vol. 112, no. 7/8, pp. 1 – 7.

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.

Sabel, CE, 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*, vol. 91, pp. 110 – 121.

Valente, P 2010, 'Census taking in Europe: How are populations counted in 2010?', *Population and Societies,* vol. 467, Viewed 20 February 2015,
<http://www.ined.fr/fichier/s_rubrique/19135/pesa467.en.pdf>.
Verhoef, H and Grobbelaar, N 2005, *The development of a Small Area Layer for South Africa for census data dissemination*, Statistics South Africa', Viewed 19 November 2009, <http://www.cartesia.org/geodoc/icc2005/pdf/poster/TEMA26/HELENE%20VERHOEF.pdf>.