The Effect of GIS Data Quality on Infrastructure Planning: School Accessibility in the City of Tshwane, South Africa

Dr Peter Schmitz[1,2], Sanet Eksteen[3]

[1]CSIR Built Environment, pschmitz@csir.co.za
[2]Centre for Geoinformation Science, University of Pretoria
[3]Centre for Geoinformation Science, University of Pretoria, sanet.eksteen@up.ac.za

This paper has been through a process of double-blind peer review

Abstract

Worldwide geospatial datasets have become more readily available to non-expert users via the Internet and have changed the context in which spatial data is used. The quality of the available datasets can only be assessed if the corresponding metadata is available as well. The lack of metadata and non-expert knowledge of users often lead to the exploitation of data in different applications. In South Africa, the promulgation of the Spatial Data Infrastructure Act has created the environment for the capturing and sharing of spatial data and the publishing of metadata. The aim of this paper is to illustrate the financial cost involved when using poor quality data in planning for infrastructure such as schools. The positional accuracy of the quality of the dataset is evaluated by determining the locations of new schools using an original dataset and a verified dataset for primary and secondary public schools in the City of Tshwane. The result of the study is expressed in monetary terms and indicates that substantial amounts of money could be misspent if the necessary metadata to evaluate the quality of a dataset is not available. The study is not aimed at criticising the good work done by the Department of Basic Education but to illustrate the impact that data of insufficient quality may have on decision making.

Keywords: GIS data quality, school accessibility, Spatial Data Infrastructure Act, metadata

1. Introduction

Worldwide geospatial data has become more readily available via the Internet to the non-expert user and has changed the context in which spatial data is used. Metadata should assist users to evaluate the quality of data before applying it in different GIS projects (Devilliers et al., 2010). The lack of metadata and the misuse of data by non-expert users have often lead to the misuse of data in different applications with a potentially significant impact on political, societal, and economic decision-making (Agumya & Hunter, 2002). In South Africa the promulgation of the *Spatial Data Infrastructure Act* (SDI act) has created the environment and awareness for the capture and sharing of quality spatial data and the publishing of metadata (SDI act, 2003). However, recent studies have indicated that the availability of metadata is an issue of concern (Olfat et al., 2012) and needs to be addressed before datasets are used for economic or political decision making.

While South Africa, like the rest of Africa, faces many societal, economic and political challenges the lack of basic education remains high on the political agenda (Hill et al., 2012; Fiske & Ladd, 2005; OBG, 2013). While government has put in substantial efforts to eradicate inequalities in the South African schooling system, there remains a shortage of classrooms in primary and secondary schools (OBG, 2013). This has given rise to the question of what the financial effect will be of using poor quality data for the planning of infrastructure such as schools.

In the light of the above, the objective of this study is to demonstrate the impact of data quality and the lack of metadata may have on infrastructure planning. This study focuses specifically on primary and secondary government schools in the City of Tshwane in Gauteng, South Africa.

The remainder of this section discusses the current state of data availability, data quality, and metadata in general. Section 3 describes the current situation of schools in South Africa while section 4 describes the methodology followed to conduct this research. The results and discussion of this research are presented in section 5 followed by the conclusion in section 6.

2. Data Availability, Data Quality and Metadata

The impact of poor data quality on GIS projects have been investigated and emphasised in various studies (Batini et al., 2009, Devilliers et al., 2010). According to Devilliers et al (2010) the lack of a commonly accepted definition of data quality has contributed towards different interpretations of the terminology. For the purpose of this article the definition of Bernhardson (2002) will be used. In terms of this definition data quality is defined as the characteristics of spatial data that describe the extent to which it is fit for use. The quality of a dataset should be evaluated according to certain criteria. These criteria include positional accuracy, attribute accuracy, temporal accuracy, logical consistency, and data completeness (Bernhardson, 2002). Currently there are no specific standards for data quality in South Africa and the ISO 19157 standard is used by some organisations to determine standards for data capturing and evaluation (ISO, 2013).

Metadata on the other hand is described as ‘data about data’ (Bernhardson, 2002). Although metadata has various uses, the most important one in context of this paper is to determine the usability of a dataset for an intended purpose (Jacobs & Chase, 2011). Datasets are only of value to users if it has metadata that allows the user to evaluate the quality of the data, however, different users require different levels of quality from the same datasets (Cooper, 1993). Without metadata, the fitness of a dataset can therefore not be evaluated.

3. Infrastructure Planning – Schools In the City of Tshwane

According to statistics released by the Department of Education there are almost 12.5 million learners attending 26 000 schools in South Africa (School Realities, 2012) spanning from primary schools (grade 0, or grade R, to grade 7) to secondary schools (grade 8- 12) (SAinfo Reporter, 2013). Approximately six percent of the schools in South Africa are independent or private schools while 94% are classified as government funded public schools (School Realities, 2012). Under the School Act of 1996 education is compulsory for all South Africans from grade 1 (aged 7) to grade 9 (or until a learner reaches the age of 15 years) (SAinfo Reporter, 2013). According to the newest released statistics, the average learner to teacher ratio of schools in South Africa is 30.4:1 (School Realities, 2012).

This study focuses specifically on the City of Tshwane located in Gauteng Province of South Africa. The location of the study area is indicated on the map in Figure 1.

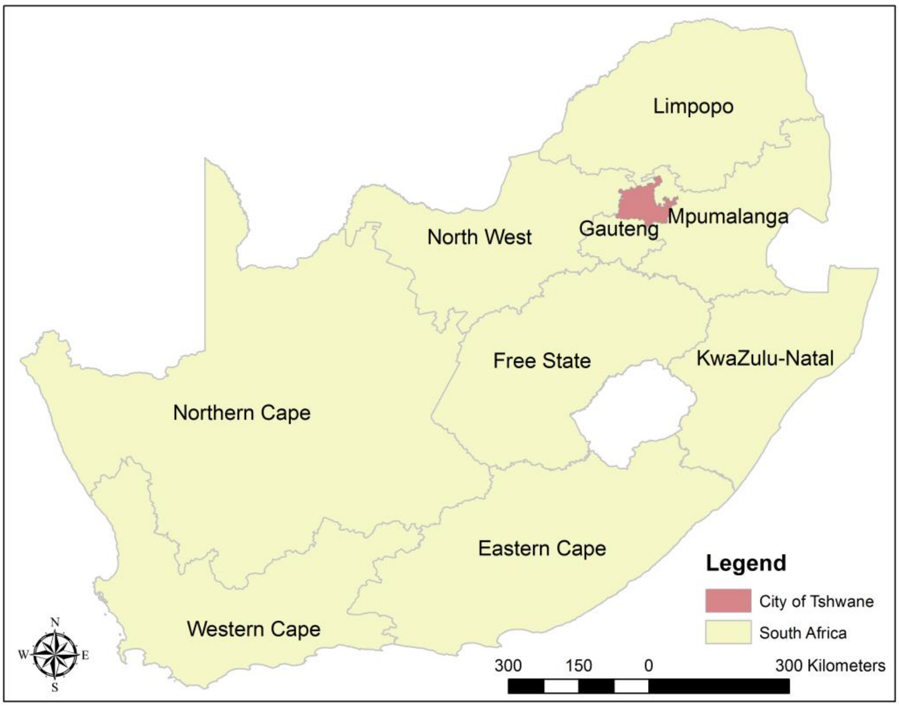

Figure 1: Location of the study area: City of Tshwane.

There are 501 public schools located inside the boundaries of the City of Tshwane with almost 53 000 learners of a typical school going age (6 – 18 years) (Stats SA, 2012) There are 18 combined schools that cater for both primary and secondary school learners, 319 primary schools and 164 secondary schools. According to the Government Gazette (2008) the ideal learner space is 29 children per classroom for a primary school and 30 children per classroom for a secondary school. The high learner-space radius means that there are children for whom the basic right of education is not accessible in the City of Tshwane. The cost of building a new primary school in Gauteng is estimated at R40 million while the cost for a secondary school is estimated at R45 million (CSIR BE, 2012). Due to the high costs of building primary and secondary schools, lack of proper planning or planning based on erroneous data could escalate to high amounts.

4. Methods

In order to demonstrate the effect of data quality on infrastructure planning only the positional accuracy of the schools were taken into account. The rest of the datasets were assumed to be of adequate quality.

4.1 Data

The following spatial data sets had been sourced or created for this project, namely:

1. Schools data as downloaded from the official website of the Department of Basic Education (DBE);
2. Verified schools data;
3. A detailed road dataset of the City of Tshwane;
4. Census 2011 population data per Small Area Layer; and
5. Land use data

Only the public schools were extracted from the schools datasets as obtained from the website of the DBE. Independent schools were excluded from this study since their locations are regarded as market driven. The locations of the public schools were verified and corrected by comparing their locations with Google Maps, Google Earth and website searches of the school addresses. Figure 2 indicates both the locations of the original and the verified locations of the schools in the City of Tshwane. For the purpose of this study, intermediate and combined schools are divided into primary schools and secondary schools. This is done since the criteria such as classroom sizes for primary schools are different to those of secondary schools.

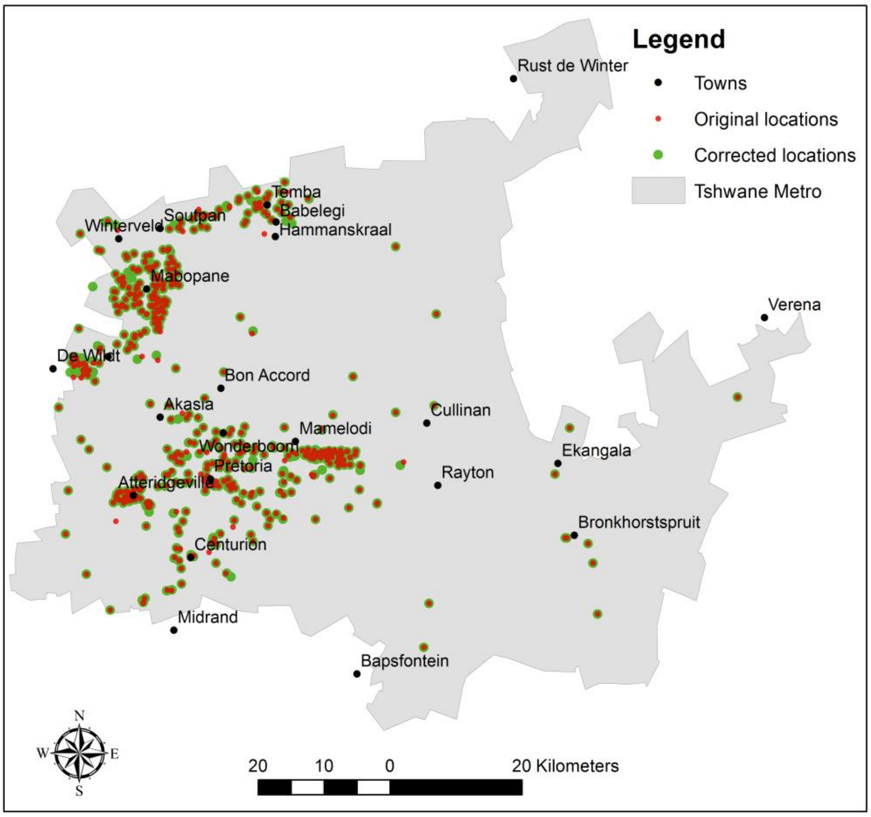

Figure 2: Original and verified locations of the public schools in the City of Tshwane.

To determine the catchment areas for the schools, hexagons for the entire study area were created using Flowmap. The user has the choice in Flowmap to use hexagons, rectangles or polygons which contains the population data in order to determine the catchment areas, accessibility and new locations. Polygons smaller than the Small Area Layer polygons such as the enumerator areas were needed to have a more detailed distribution of the target population. Since the enumerator areas are not available to the general public it was decided to use hexagons. In addition, hexagons cover the study area more effectively. The length of the hexagon's side was set as 350 metres since this was smallest distance that could be used in Flowmap as determined according to the size of the study area. The hexagons thus allowed for areas smaller than the Small Area Layer from Stats SA as well as minimizing the number of schools that could fall in the same polygon. The number of children per hexagon were calculated and allocated to each hexagon based on existing residential land use and the Census 2011 data. The land use was linked to each hexagon and the population data from the Small Area Layer was proportionally allocated to the hexagon. Figure 3 indicates the number of children per hexagon. Hexagons that have no children allocated to them did not contain residential areas. The children were grouped into four groups; the first group is children of school going age namely those between 6 and 18 years old. This group was then divided into a primary schools group, 6 to 12 years old and a secondary school group; those that are 13 to 18 years old. The fourth group are the Grade R, 5 year olds.

A distance matrix that calculated the distances between hexagons using the road network was created in Flowmap. This was used to determine the catchments of the schools as well as to determine an optimal location for new schools.

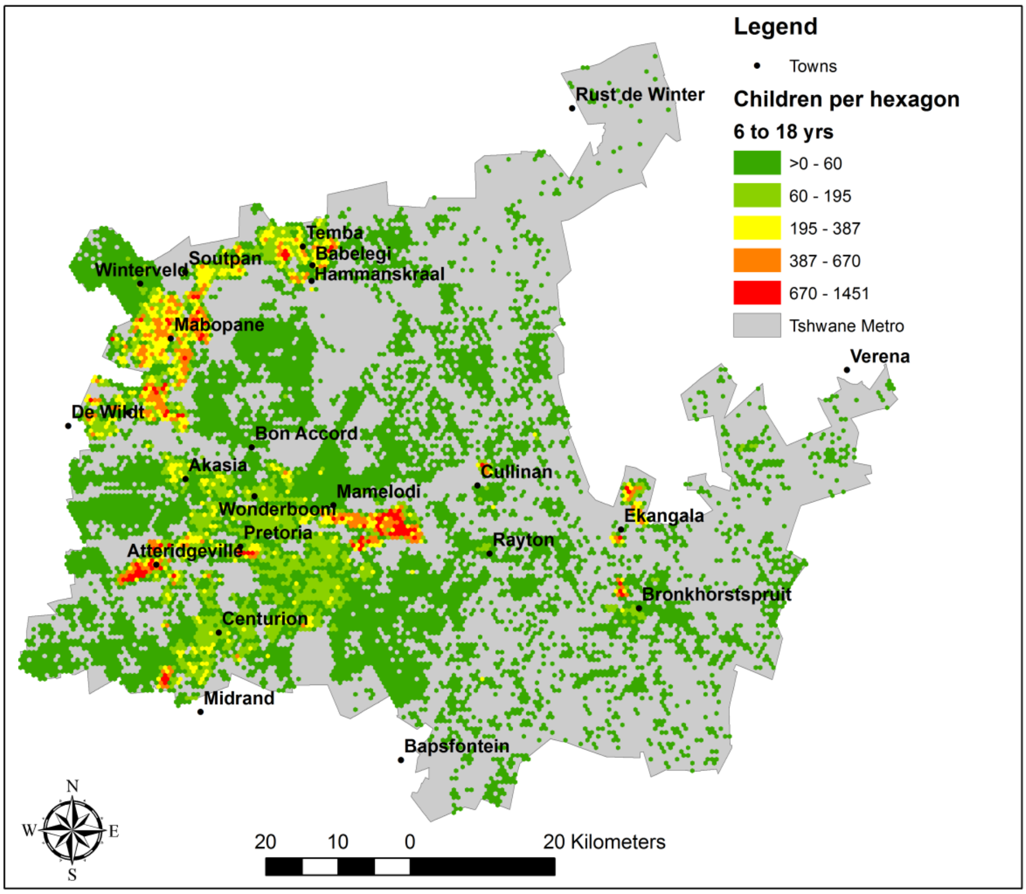

Figure 3: The distribution of the number of children (6 – 18 years) per hexagon in the City of Tshwane**.**

4.2. Analysis

4.2.1. Error distances

The projected coordinates for the original and verified schools spatial data sets were determined. The Euclidian distance between the original location and the verified location was then used to calculate the distance error between the two locations.

4.2.2. Hexagon allocations

For modelling purposes both the original locations and the verified locations were assigned to a hexagon. These hexagons were used to determine the catchment areas of the school and also served as an input to determine the locations of new schools. For the purpose of this study the schools were assumed to have the same capacities as per the norms and standards of the DBE although the DBE has the of number learners enrolled per school available. This assumption was made for ease of modelling in Flowmap. The capacity for primary schools is 930 learners and the capacity for secondary schools is 1000 learners (Gazette, 2008). If the school is a combined or an intermediate school, the school is then listed as a primary school as well as a secondary school, and the learner size is assumed to be that of a small school, thus for primary schools it is 320 and for secondary schools it is 400 learners (Gazette, 2008). The number of class rooms for combined schools is assumed be similar to those of a primary or secondary school. To accommodate both the primary and the secondary learners at these schools it was assumed that the size of these schools were similar to a small school. If a hexagon contains two or more schools, the combined capacity of the schools was used to determine the catchment areas of the schools. In addition the locations of the schools were not captured consequently as some locations were captured at the entrance of the school terrain, in the middle of the school terrain or on top of the building. This could also cause that the school is allocated to a different hexagon.

4.2.3. Catchment analysis – original locations and corrected locations

Using the distribution of children as discussed in Section 4.1 and the hexagons with the original locations of the schools with their capacities, the catchments of each school could be determined. Furthermore, the Schools Act indicates that no learner should travel more than five kilometres to the nearest school (Gazette, 2008). A catchment consists of a number of contiguous hexagons that are located around the hexagon in which the school is situated. The number of hexagons is determined by the capacity and travel distance constraints imposed during the analysis. The travel distance was calculated from an existing road network. The same analysis is conducted with the verified locations and the impact of positional accuracy could then be illustrated.

4.2.4. Locations for new schools

For demonstration purposes to determine the locations of new schools two approaches were followed, namely locating ten new schools using the aforementioned constraints and the original and verified locations of schools; and locating new schools until 90 percent of the learners are within five kilometres from a school. The results were then compared to the verified locations of the schools in order to illustrate the impact of positional accuracy.

5. Results and Discussion

5.1. Error distances

The error distances are direct distances in meters calculated between the original and verified data sets. These distances are summarised in Table 1.

The largest error distance is 111.6 km and the smallest is one metre. **Error! Reference source not found.**Fifty percent of the 501 schools in the study area have no positional error or the errors are less than a metre. Only one school has an error of less than ten meters. A total of 176 schools, (35 %), have a positional error of more than 100 meters. The root mean square error (RMSE) has been calculated with regards to primary and secondary schools. The RMSE is sensitive to large errors in the data and thus a good indicator of the quality of the data (Bolstad, 2008). RSME is calculated based on the difference in distance between the original and verified schools. A large RSME value indicates large errors between the two data sets and a low RSME value shows a higher accuracy with regards to location between the two data sets. The RMSE for primary schools of 6690 meters indicates that the spatial dataset contains large positional errors. The RMSE for secondary schools is lower at 735 meters indicating that secondary schools were more accurately captured than the primary schools.

Table 1: Error distance classes and number of schools.

Distance Class	Frequency
No Error	250
Less than 10m	1
10 to 49m	13
50 to 99m	61
100 to 249m	93
250 to 499m	23
500 to 999m	27
1km to2.49 km	16
2.5 to 5km	8
More than 5 km	9
Total number of schools	501

5.2. Hexagon allocations

Table 2 summarises the impact that positional errors have on allocating schools to a hexagon.

Table 2: Impact of errors on allocation of schools to hexagons.

Distance Class	Frequency	Within same hexagon	Outside the hexagon	Percentage inside original hexagon
No Error	250	250	0	100
Less than 10m	1	1	0	100
10 to 49m	13	12	1	92.3
50 to 99m	61	56	5	91.8
100 to 249m	93	69	24	74.2
250 to 499m	23	4	19	17.4
500 to 999m	27	0	27	0
1km to2.49 km	16	0	16	0
2.5 to 5km	8	0	8	0
More than 5 km	9	0	9	0
Total	501	392	109	

The positional errors contained in the dataset will have an impact on calculating accessibility to a school as well as determining the locations of additional schools. Even with a small positional error that is less than 350m, a school can potentially be placed into a neighbouring hexagon if the school is located close to the hexagon’s boundary. A total of 109 schools were placed in the wrong hexagon and could influence the accessibility of the school as well as the locations of new schools. The placement of schools in a wrong hexagon may lead to a change in the catchment of the school using the 5 km distance and could lead to hexagons that would have been part of the catchment being placed in another school’s catchment.

5.3. Catchments

The catchment area for each school was determined using the school’s capacity and a maximum travelling distance of five kilometres as constraints in the analysis. Flowmap calculates the catchment of a school by adding the number of learners of the nearest hexagons until the capacity of the school or the maximum distance of 5 km has been reached. Four catchment analyses were made, namely original located primary and secondary schools and the verified primary and secondary schools. Three of the catchment areas have changed due to the available number of children allocated in the two applications. To illustrate the impact of unverified schools, Figure 4 shows that if schools have positional errors such as 823 or 2307 meters it will have a major impact on the analysis. In Figure 4 the original schools using the criteria that at least 90 percent of the learners should be within a 5 km distance from a school the model added an extra school when compared against the verified schools. This could lead to the construction of a school that was not required if verified data was used in the modelling process.

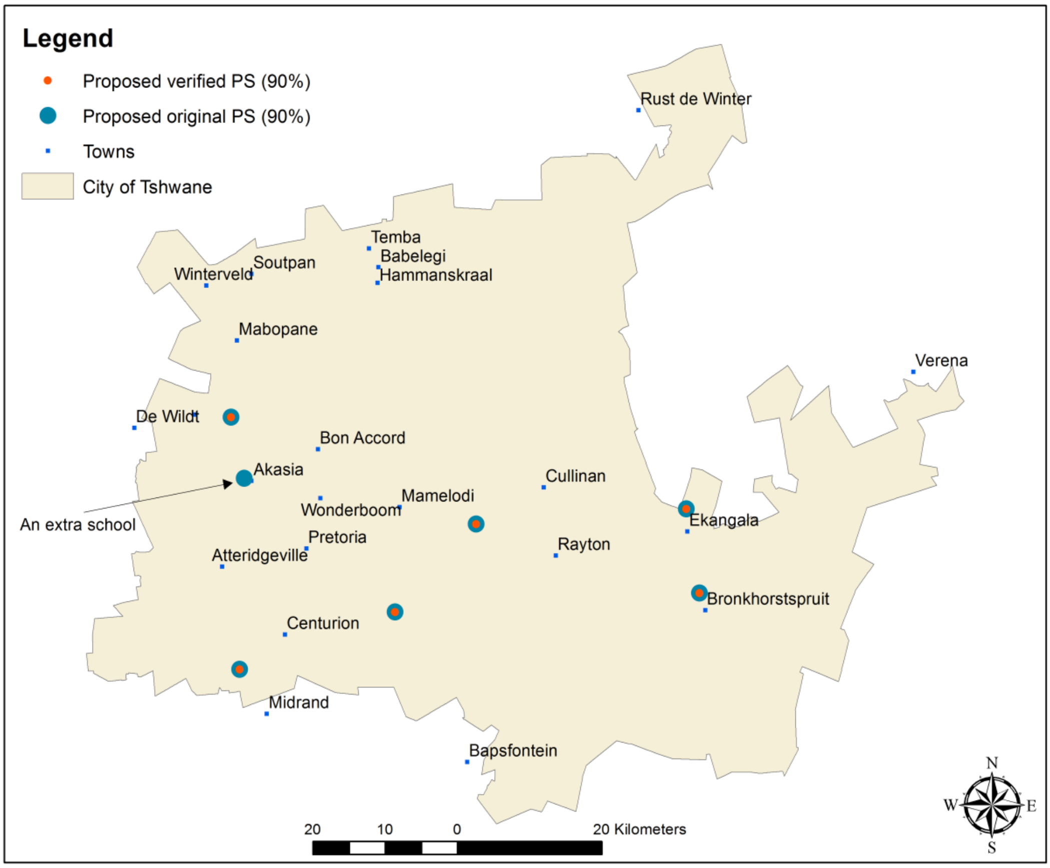

Figure 4: The impact of schools that are not verified with regards to their location.

5.4. Locations for new schools

To demonstrate the impact of unverified locations, the first example is to place an extra ten schools using the criteria of the 5 km travel distance to the school. The selection of ten new schools is purely for demonstrative purposes. The model can be used to determine where these schools should be built to answer the need for access to schools. This was done for primary schools as well as secondary schools. The first analysis compares the locations of the ten new schools using the aforementioned constraints. This was done for both the original and verified schools. Figure 5 shows the locations of the new schools for the original and verified primary schools. For the original schools, two new schools were placed in the vicinity of Mabopane and Hammanskraal in the north.

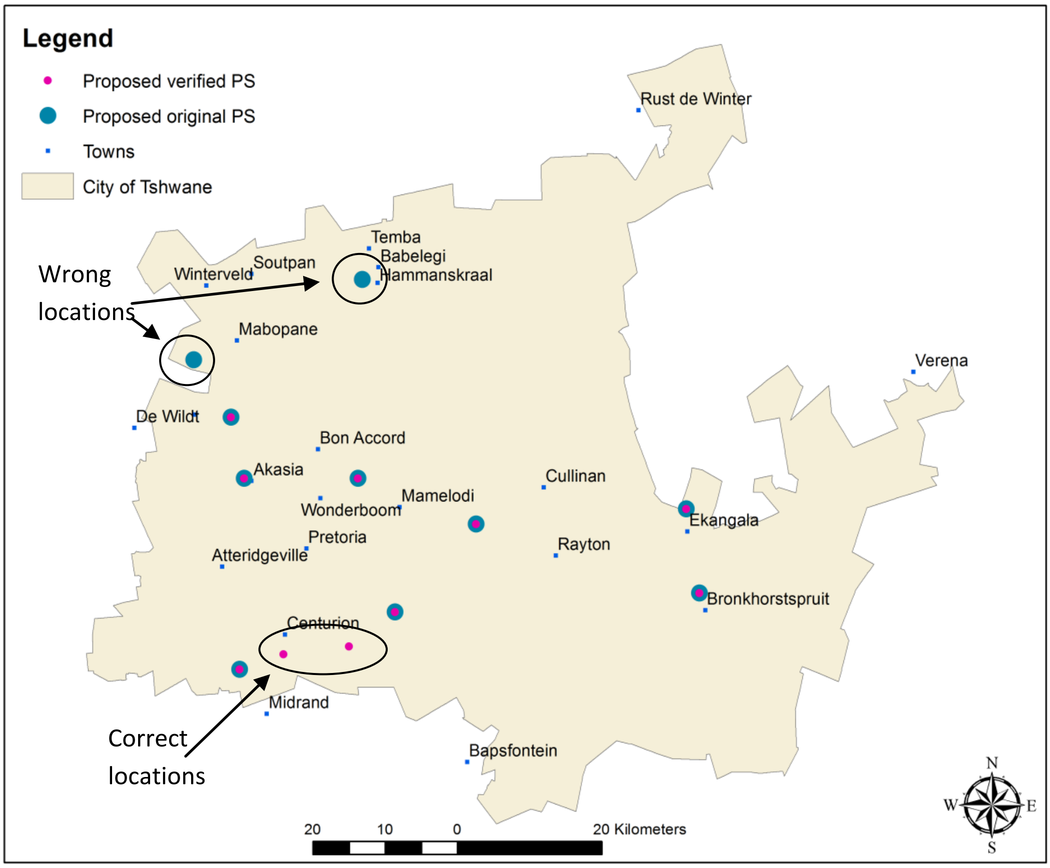

Figure 5: Proposed ten new primary schools.

When the verified school dataset is used, eight of the ten schools were in the same location and two of the new schools were located near Centurion in the south. If the original schools dataset was used the planners might have built two new primary schools in the wrong locations. This would result in spending R80 million on building schools in the wrong location where there actually was no need for new schools.

With regards to the ten new secondary schools, there is little difference in placing the new schools using the original and verified locations. Five of the schools were in the same position and four schools were placed in the neighbouring hexagon, i.e. 600 metres. One school was placed 1km from the original data set location. This corresponds with the fact that secondary schools have a much smaller RMSE than the primary schools. For this study the shifts in locations of secondary schools are insignificant and had no effect on planning, thus there was no danger of placing a new school in a wrong location.

The second form of placing new schools was to determine the number and locations of additional schools so that 90 % of the learners are within five kilometres from a school. The analysis was done for both the original and verified primary and secondary schools. The secondary schools did not show remarkable difference in locations of the proposed schools as shown in Figure 6.

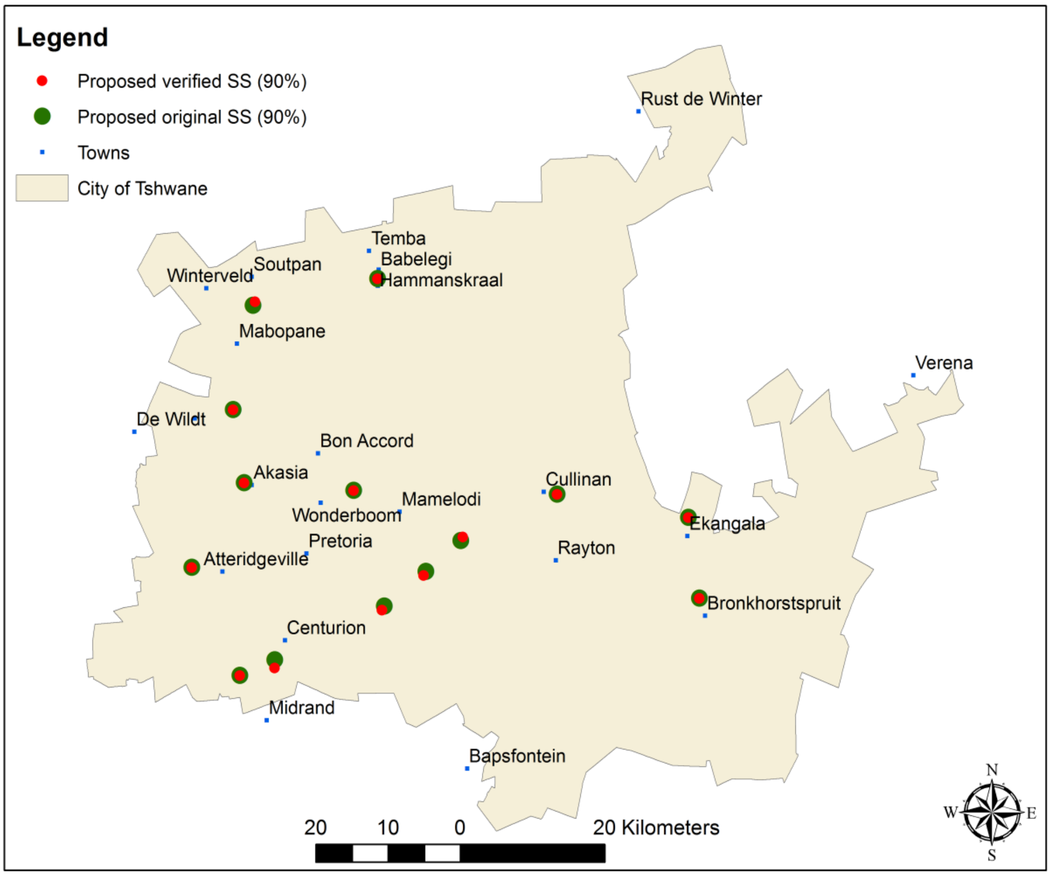

Figure 6: Proposed new secondary schools to give 90 percent coverage.

With regards to the original dataset for primary schools the results indicate that seven new primary schools are required to reach 90 percent of the learners. Using the verified spatial dataset, the results indicated that only six new primary schools are required. Thus using the original spatial dataset could result in the spending of an extra R40 million on a school which is not necessary to build if the verified data was used. Figure 7 indicates the difference in results between the two data sets.

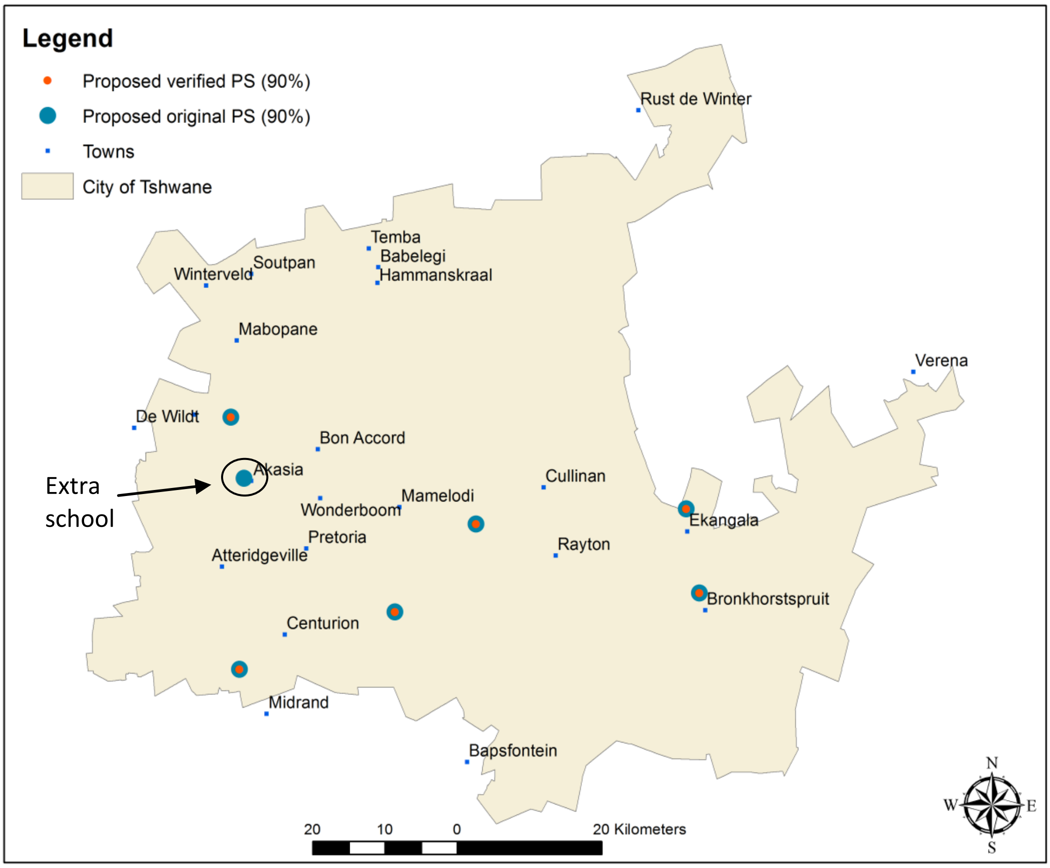

Figure 7: Proposed new primary schools to give 90 percent coverage

5.5 Metadata

The lack of metadata to evaluate the usability of the original dataset has led to a substantial error in the modelling of the placement of additional schools in locations. These errors can be measured in distances and monetary values. The direct distance error between the 2 of the 10 additional schools placed in the north of Tshwane and in the south of Tshwane as discussed in Section 5.4 and shown in Figure 5 is respectively 42 km and 50 km and the monetary value of the wrongly placed schools was R80 million. The error using the 90% coverage an additional school was proposed by the model based on the original data compared to the verified data which led to the spending of an extra R40 million. Due to the lack of metadata it was not possible to evaluate the usability of the original dataset from DBE for this project as no metadata was distributed or made available on the website of the DBE. The metadata should describe the quality of the spatial dataset, including the positional accuracy, to enable potential users to evaluate the usability of a dataset.

6. Conclusion

The authors would again emphasise that this study is not a critique on the good work done by the DBE but was rather to illustrate the impact of data quality and the lack of metadata on decision making. The percentage schools with wrong locations were for both the primary and secondary schools around 54 percent. The wrong locations of schools are due to the errors in the locations of schools in the original data set as well as the inconsistent manner in which the locations of the schools were captured. The positional errors of primary schools were larger than those of secondary schools. These errors lead to differences in the location and number of primary schools that need to be built. If ten new schools needed to be built the error would have been R80 million spent on schools that were not required in an area sufficiently covered by existing schools.

If the original spatial data set was used the DBE would have spent an extra R40 million to reach a 90 % coverage to satisfy the learners' needs.

If the metadata for the schools datasets were available the quality of the data could have been evaluated before using the dataset to determine locations for new schools. This paper illustrates that any error distance causing the verified school to fall in a different hexagon may influence the locations of additional schools. In order to have a finer distribution of learners in Tshwane, it was needed to use a small hexagon size of 350m. Although larger hexagons could have been used, the small hexagons were used to illustrate the impact of errors in the data set.

In this paper only the impact of positional accuracy has been discussed. It did not cover attribute accuracy with regards to addresses, contact details, number of classrooms and number of learners. It is recommended that the research should be expanded to determine the impact of the data quality criteria. These criteria include positional accuracy, attribute accuracy, temporal accuracy, logical consistency, and data completeness.

7. References

Agumya, A., & Hunter, G. J. (2002). Responding to the consequences of uncertainty in geographical data. *International Journal of Geographical Information Science* , 405-417.

Batini, C., Cinzia, C., Francalanci, C., & Maurino, A. (2009). Methodologies for data quality assessment and improvement. *ACM Computing Surveys*, 16:2-52.

Cooper, A. K. (1993). Standards for exchanging digital geo-georeferenced information. MSc, University of Pretoria, South Africa.

CSIR BE. (2012). Geographic accessibility study of social facility and government service points for the metropolitan cities of Johannesburg and eThekwini 2011/12. Pretoria: CSIR Built Environment.

Devilliers, R., Stein, A., Bedard, Y., Chrisman, N., Fisher, P., & Shi, W. (2010). Thirty years of research on spatial data quality: Achievements, failures and opportunities. *Transactions in GIS*, 387 - 400.

Fiske, E. B., & Ladd, H. F. (2005). Elusive Equity: Education Reform in post-apartheid South Africa. Pretoria, South Africa: HSRC Press.

Gazette, G. (2008, November 21). Notice 1439 of 2008: Department of Education, South African Schools Act 84 of 1996 - Call for comments an national uniform norms and standards for school infrastructure. *Government Gazette No 31616* . Pretoria, Gauteng, South Africa: Government of South Africa.

ISO. (2013). ISO/FDIS 19157:2013. Geographic information – Data Quality.

Jacobs, F., & Chase, R. (2011). *Operations and Supply Chain Management, 13th Edition.* New York: McGraw-Hill Irwin.

OBG. (2013, April 3). *South Africa: Reforming education. Retrieved April 3, 2013, from Oxford Business Group (OBG):* http://www.oxfordbusinessgroup.com/economic_updates/south-africa-reforming-education

Olfat, H., Kalantari, M., Rajabifard, A., Senot, H., & Williamson, P. (2012) A GMLS-based approach to automate spatial metadata updating. *International Journal of Geographical Information Science*, 231-250.

SAinfo Reporter. (2013, February 28). *Education in South Africa*. Retrieved March 23, 2013, from South Africa.Info: http://www.southafrica.info/about/education/education.htm#bands

School Realities. (2012, September). *Statistical Publications*. Retrieved March 26, 2013, from Department of Basic Education: http://www.education.gov.za/LinkClick.aspx?fileticket=MMXRVCugRQ4%3d&tabid=462&mid=1327

Spatial Data Infrastructure Act, 2003, no 54 of 2003.

Stats SA. (2012). Census 2011. *Census 2011* . Gauteng, South Africa: Stats SA.