

Ilse Brits

Professional GIS
Geography Division
Statistics South Africa
012 310 8941
ilseb@statssa.gov.za

Abstract

Statistics South Africa (Stats SA) is custodian of many datasets that have to be correctly geo-referenced for reporting purposes. The official place name spatial layer which is used for geo-coding consists of 22 000 sub place names (i.e. suburbs and villages) and 13 000 main place names (i.e. towns and tribal authorities). The layer is based on Enumeration Area (EA) polygons which were created for the purpose of disseminating Census 2001 data.

However, the question of how non EA based datasets should be geo-coded remains a challenge. Part of the challenge is that the Census place name layer does not include place names that cover an area smaller than an EA and does not provide for alternative place names or spelling as one would find on a questionnaire or administrative form.

Based on recent case studies, the paper identifies the merits and limitations of using a place name point dataset as an alternative to geo-reference key datasets.

1. Introduction

Statistics South Africa (Stats SA) is custodian of many datasets that have to be referenced to a geographical area. There is a problem with datasets where the record place name has no link to a Census 2001 place name. The Census 2001 place name database was created for the dissemination of Census 2001 data. The problem is that it is not comprehensive enough for geo coding of place names.

The reason is that the census place name layer was created from EA polygons. One or more EAs were grouped to form a place name polygon. The relationship do not provide for place names that do not match the EA/Place name geographical hierarchy. Alternative names or alternative spelling of names as one would find in a questionnaire or administrative form are not included. As part of the national statistics agency, the Geography Division are responsible for the coding different datasets and needs to stream line the place name coding process.

Based on case studies, this paper will aim to answer the following questions:

1. How useful is the Stats SA Census 2001 place name polygon dataset as a code list in coding non-census data?
2. What are the limitations and advantages of a polygon or a point dataset in coding of place name data?

2. Place name code list

The place name code lists used for coding in the case studies where created from spatial place name datasets. A spatial dataset is not necessary for coding of place names as the place name code is the link to the spatial dataset. The place name code list is used to match to the data record field. If there is an exact match between the place name code list and the data record field the place name code are accepted.

3. The Case Studies

3.1 Place name datasets

The place name datasets that are used as a 'code list' or authority table where:

- The Census 2001 place names that are based on the EA polygons;
- An unofficial place name dataset created by the Geography Division that is based on spatial points.

3.1.1 Census 2001 place name dataset

The official Census 2001 place name layers was created in 2002/2003 and were the most appropriate way to disseminate the Census 2001 information. The place name spatial layers have a total of 21 243 sub place names (suburbs and villages) and 3 109 main place names (cities, towns and tribal authorities). The layers consist of polygons of one or more dissolved EAs. Each EA has a sub place name that fits within a main place name that in turn fit within the Municipal Demarcation Board municipality layer. The set hierarchy means that the place name boundaries are indirectly based on the demarcation of the EA boundaries. For more information on the Census 2001 geographical hierarchy and place names visit the Stats SA website www.statssa.gov.za/census01/html/Geography_Metadata.htm. The polygon place name layer covers the whole of South Africa and is indicated in Figure 3.1.

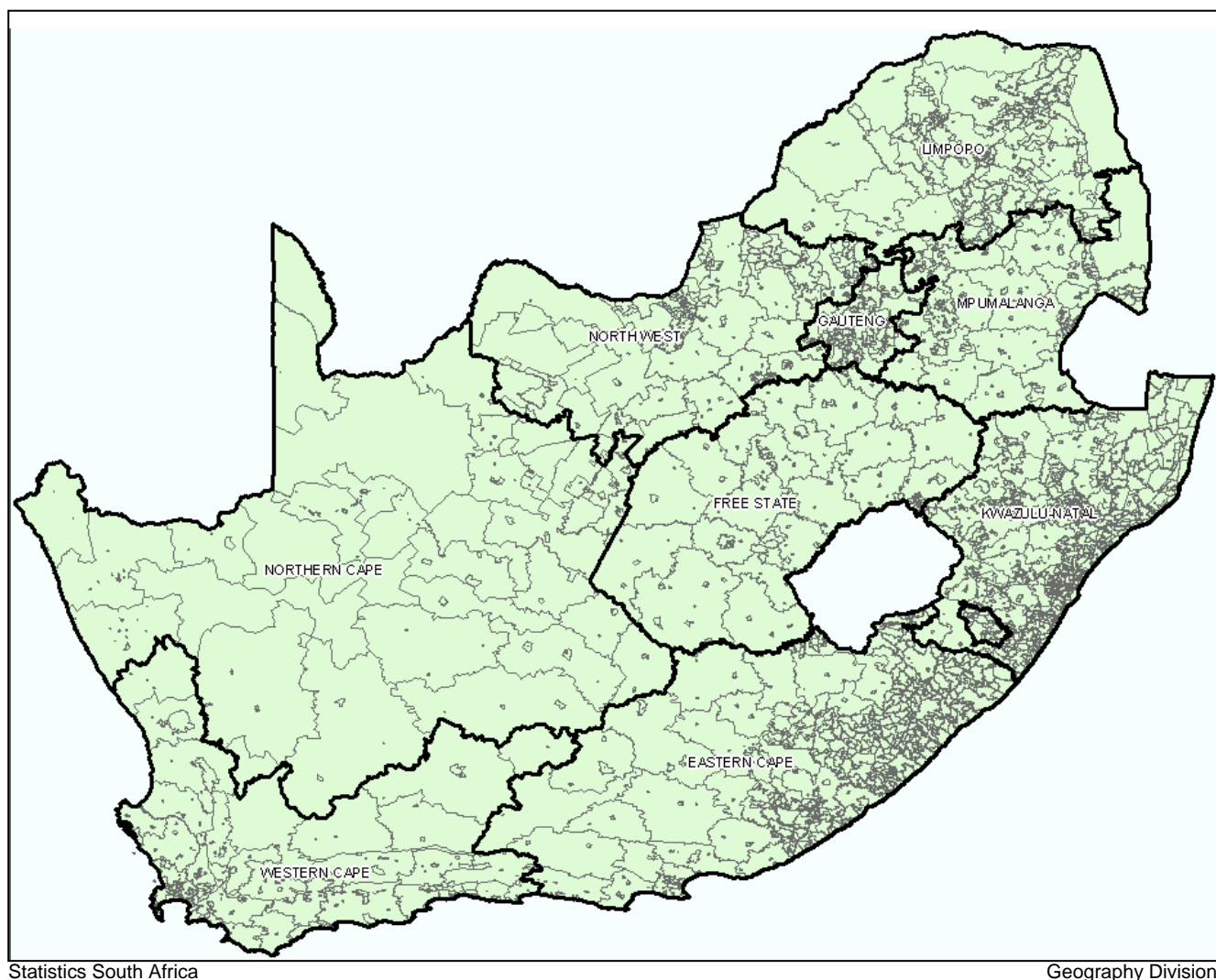


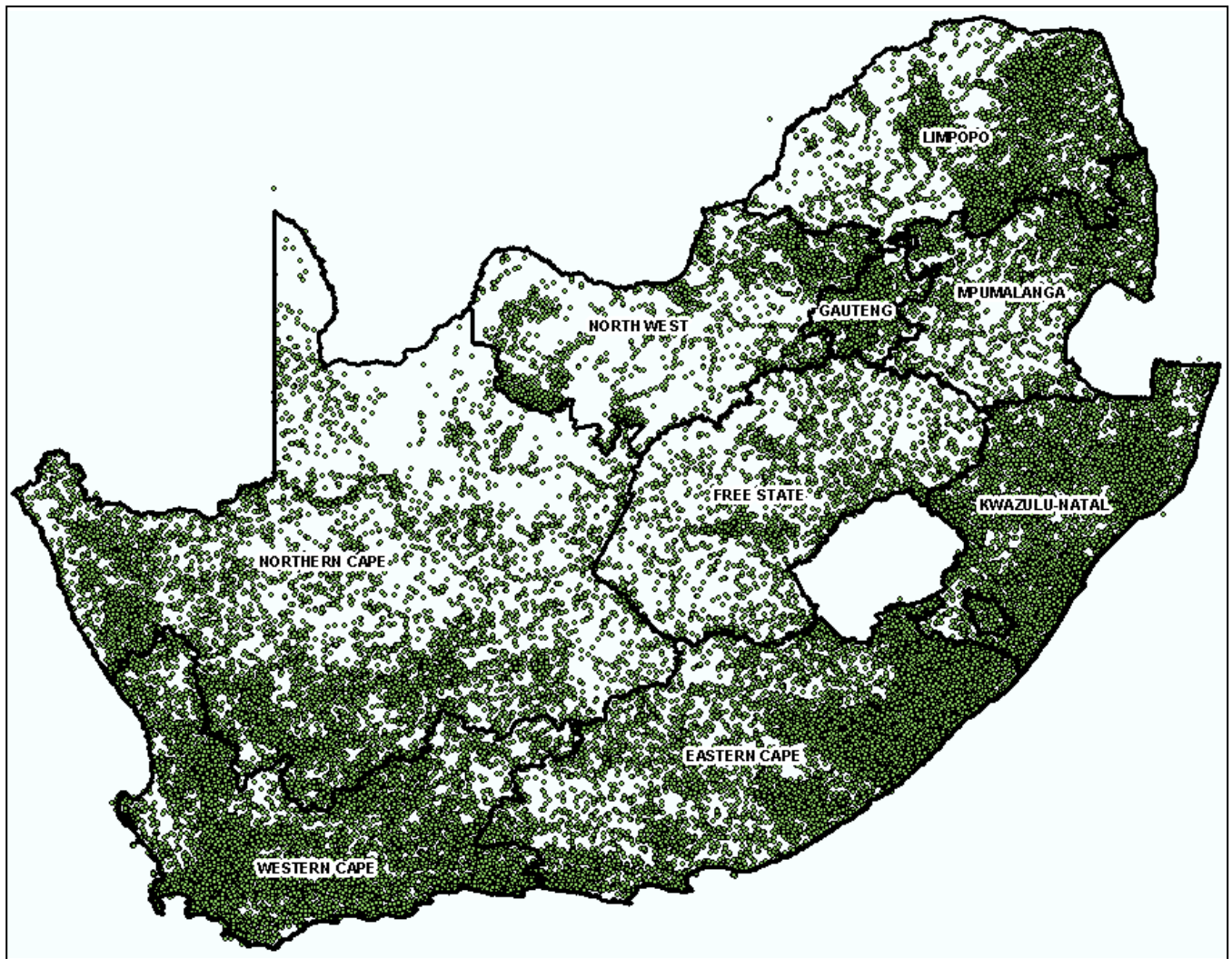
Figure 3.1 100% coverage of place name polygons for South Africa (Sub place name layer)

If the place name did not fit within the Stats SA geography hierarchy the place name was not added to the dataset. As the place name layer is based on a limited number of polygons, the coding list only have 24 352 (21 243 sub place names + 3 109 main place names) records that can be used for coding.

3.1.2 Unofficial place name code list

The unofficial place name code list was created from 56 863 unique place name points. It was formed by merging different place name datasets from external sources. The dataset has no relation to any type of polygon or area as the place name location is indicated by a point as indicated in Figure 3.2. For reference back to the Census 2001 place name dataset a spatial intersection was run on the Census 2001 main place layer and the point place name dataset. The main place code gives a reference to the geographical location of the place name.

The advantage of a code list created from a point dataset is that many place names could be added because there was no polygon structure or hierarchy rules that the place names had to adhere to.



Statistics South Africa

Geography Division

Figure 3.2 Place name points for South Africa

3.2 Case study 1: Geo referencing of Stats SA Business register data to municipality

The Stats SA Business Register (BR) is built on information from administrative data. The BR is used as a basis from which sampling is done for the various economic statistics that are produced by Stats SA. One of the statistical units that are used in the BR is the GEO unit that indicates the physical address from which an economic activity is taken place. The address could either be a street or postal address, suburb or town name or farm name or tribal village name. During 2004 an initiative was launch to match the GEO unit code of the BR records to the respective municipality. In total 2.9 million records was coded.

The first batch of 2 million records was coded by using the Census 2001 place name dataset and the National Address Directory of 2001 (NAD2001). The place name dataset was adequate as coding list but there was still a lot of manual coding. After the first evaluation session, it was clear that a more streamline place name code list was needed if the deadline was to be met. The main problems were:

- Smaller settlements, farms and village names that did not conform to the EA polygons were not in the Census 2001 place name dataset;
- The NAD2001 only covers the larger formal areas in the country;
- Place names that appear in more than one location. For example Middelburg in Western Cape and Middelburg in Mpumalanga. Richmond in Eastern Cape, Richmond in KwaZulu Natal and the suburb Richmond in Johannesburg. Newlands in Tshwane and Newlands in Cape Town. Thembisa in Gauteng and Thembisa near Burgersdorp;
- Spelling and typing errors and
- Inconsistent location descriptions.

The geo-coding process was redesigned and a modified code list was used. In the modified code list, duplicate Census 2001 place names with different place name types were deleted. The other duplicate place names were investigated and a description was added as an attribute to the record. The attributes was used as a quick reference to identify which place name to choose. An example of the duplicates and attributes are indicated in Table 3.1. The new code list eliminated the problem of duplicate place names.

Table 3.1 Duplicate place names and description

| Place name | Description |
|-------------|---|
| Bhongweni | Township of Kokstad |
| Bhongweni | Mine area near Randfontein |
| Boipatong | Settlement near Winburg |
| Boipatong | Near Sharpville |
| Ekuphumleni | Settlement near Sada and Whittlesea |
| Ekuphumleni | Settlement near Keaton on Sea |
| Eluxolweni | Village |
| Eluxolweni | Settlement near Pearly Beach |
| Ikageng | Village |
| Ikageng | Township of Potchefstroom |
| Jamestown | Near Stellenbosch |
| Jamestown | Small town |
| Tembisa | Settlement near Burgersdorp |
| Tembisa | Township of Kempton Park |
| Thokoza | Township of Alberton |
| Thokoza | Settlement near Tweefontein and Vezebuhle |
| Waterval | Near Rustenburg |
| Waterval | Near Newcastle |

3.3 Case study 2: Geo referencing of Causes of Death place names

During 2004 the death notification forms from years 1997 to 2003 were captured and the information was then used to report on mortality in South Africa. (Mortality and causes of death in South Africa, 1997–2003: Findings from death notification (P0309.3) - Published: 2005) The place names on the death notification forms either indicated the place of birth, place of death or place of death registration. The role of the Geography Division was to create a place name code list as well as to search and code 'unknown' place names. 'Unknown' place names were place names that could not be found on the code list.

The experience with the first batch of BR records made it clear that the Census 2001 place name dataset was not inclusive enough for coding the death notification forms. A more comprehensive list was needed and the unofficial place name code list described Section 3.1.2 was created. The reporting level of the publication was on provincial level and that meant the place name had only to be located in the correct province. Only 4% of the total place names captured from the death notification forms were classified as 'unknown' and a weekly batch was sent to the Geography Division to be coded. The process followed to code the 'unknown' place names is indicated in Figure 3.3.

The unknown place name problems included:

- Abbreviations, spellings and typing errors;
- Place names located outside South African borders,
- Hospital names,
- Alternative names;
- Historic names (Sofiatown, Stilton, Lady Selborne; Voortrekkerhoogte)
- The precise location of a place name could sometimes not be determined as there was not enough information available.

Most of the 'unknown' place names were from the place of birth records as place names have changed in the last hundred years. New and historic names were added to the code list which in the subsequent batches left only spelling and typing errors to be corrected.

4. Findings

The polygon place name dataset of Stats SA was created for the dissemination and reporting of the Census 2001 data. The polygon dataset did not perform well as a coding list for mentioned case studies. The unofficial point code list was more effective in coding as it included more place reduced manual input and therefore saved time.

4.1 Polygon place name data as a code list

| Limitations | Advantages |
|---|---|
| <ul style="list-style-type: none">• The number of place name records is limited by the number of polygons.• The place name area is based on the demarcation of EAs and does not necessary take the place name boundary in to account.• If a place name area is not similar with the EA areas, it was not added.• Maintenance and updates are difficult as settlement boundaries change constantly.• Settlement types like traditional settlements boundaries/polygons are usually interpretations and are subjective.• The database can only be updated when the polygons are changed.• If there is a change in the polygon boundary the place name boundary must change. | <ul style="list-style-type: none">• The polygon place names make it possible to report census statistics on small areas like a suburb or small village.• There is a direct relation between the place name area and the reported data. |

4.2 Point place name data as a code list

| Limitations | Advantages |
|--|--|
| <ul style="list-style-type: none">• There is no relation between the place name area and the place name point.• Reporting can only be done on a higher geographical level like municipality or province.• Quality of reporting areas could be compromised because of grouping place names points to a higher geographical level. | <ul style="list-style-type: none">• New place names can be added making updating and maintenance simple;• Spatial layers do not have to change to update the place name dataset;• A large pool of place names can be used as a coding list saving time and improving productivity;• A point dataset can be used for coding of census data or other non census datasets. |

5. Conclusion

A polygon place name dataset or code list is just as good as the function it is created for. No new place names can be added to the official dataset because of the fixed polygon structure. The Census 2001 place name dataset it can only be updated when the EA dataset are reviewed during the pre-enumeration phase of the next census. This means that a new place name polygon code list will only be available after a census which is usually every ten years.

A point place name dataset and code list can be maintained and updated regularly. The large number of place names is more productive and saves time in the coding of large datasets. The important aspect to remember is that the geographical reporting level must be higher than that indicated on the unit records.