

THE DEVELOPMENT OF A SMALL AREA SPATIAL LAYER TO SERVE AS THE MOST DETAILED GEOGRAPHICAL ENTITY FOR THE DISSEMINATION OF CENSUS 2001 DATA:

Nick Grobbelaar,

Manager: Geo-database, Geography Division, Statistics South Africa.

NickG@statssa.gov.za

Abstract

It is a well-know fact that the data collected during the national population census of 1996 was published and disseminated at the Enumeration Area (EA) spatial level. This has allowed all users of the data to aggregate the data based on their specific spatial entities of choice.

Due to concerns in regards to confidentiality, the lowest level on which the Census 2001 data was released was the second tier of the spatial hierarchy namely the sub place, which relates to suburbs, wards, villages, farms or informal settlements.

In an effort to address user concerns in regards to the above mentioned, Statistics South Africa undertook the responsibility of supplying users with custom aggregated data sets, based on the users spatial entity preferences, as long as confidentiality was kept in tact.

Subsequently a project was initiated in 2004, with the objective of developing a spatial layer for the purposes of disseminating data for certain census variables at a level lower than the sub place and as spatially similar to the Enumeration Area, as permitted by confidentiality.

The paper will address the development of an automated spatial process to create a small area layer. It will include the spatial rule set implemented, the geo coding of the spatial layer and a basic spatial synopsis.

1 Introduction

During the launch of the 2001 Census results, in July 2003, Statistics South Africa announced that the community profiles would not be released at the enumeration area level, but only from the then second tier of the spatial hierarchy namely the Sub Place (SP) and upwards, (See Figure 1). This is in contrast with the national censuses of 1991 and 1996, where the basic spatial unit implemented for census information collection and dissemination was the Enumeration Area.

This control was implemented due to the fact that confidentiality as required by section 17(6) of the Statistics act of 1999 may be comprised by cross tabulating 2 (two) or more variables, at the Enumeration Area geographical level. This will result in such small totals that an individual's anonymity will be compromised.

Statistics South Africa has stated from the outset that a user, who requires the population census information at the Enumeration Area level, should approach the organisation to facilitate preparation of the data. Consequently data has been provided to organisations based on customized processing of the Census 2001 Enumeration Area level data.

2 Small Area layer

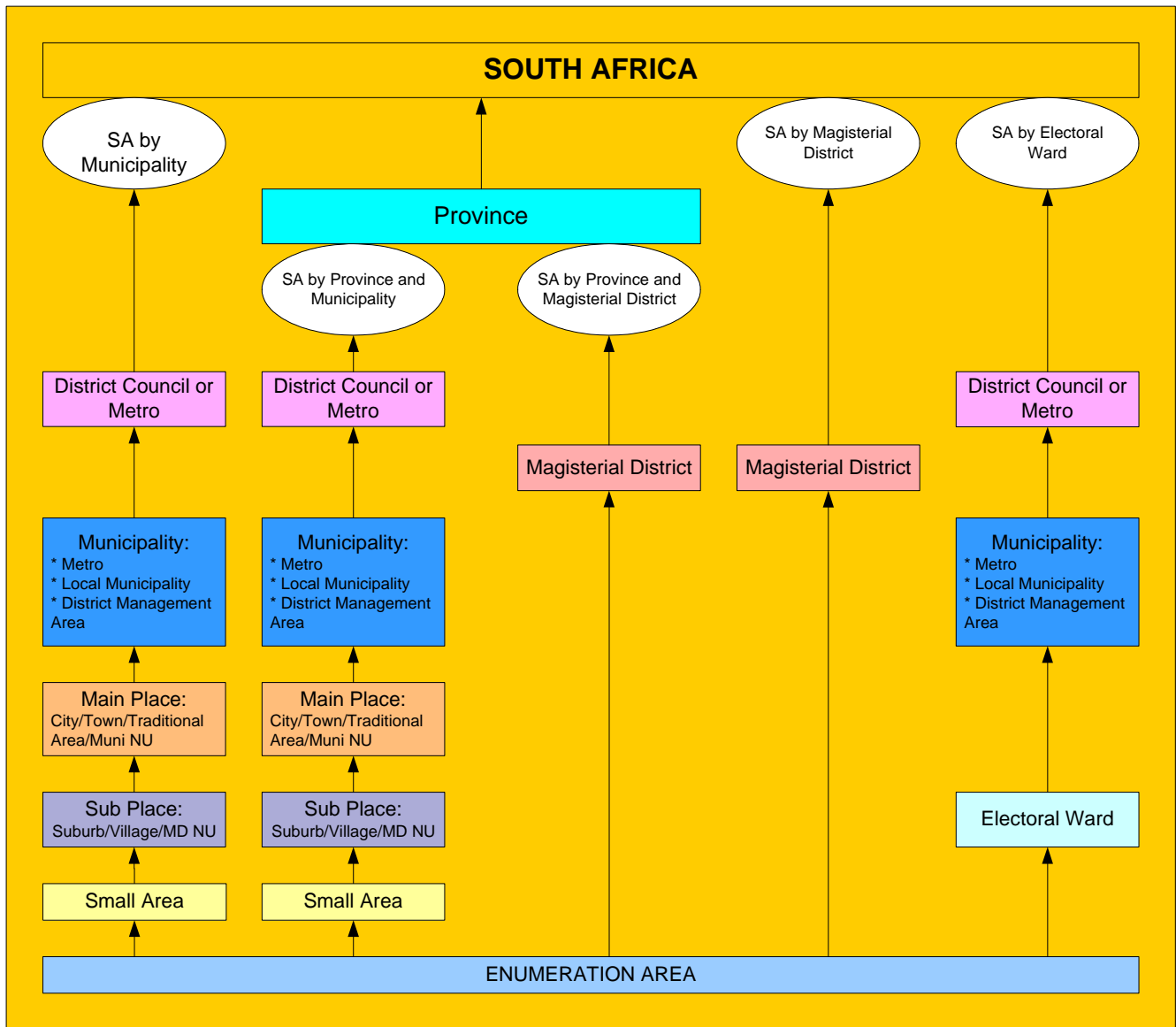
There is currently no firm policy governing the dissemination of the population statistics at the Enumeration Area geographical level. In order to address a continuous demand from users and balance the confidentiality requirement, Statistics South Africa undertook a study to identify a geographical layer that comprises of units that contain a large enough population to reduce the risk of the possible identification of individuals when cross tabulation of variables is done. The fundamental finding of this study applicable to the development of such a

layer was that a minimum population total of 500 per Enumeration Area is required to ensure confidentiality. Inseparably linked to this was the restricting of census variables to be published with the layer.

Based on this, the development of a spatial layer for dissemination purposes, which corresponds as much as possible to the Enumeration Area layer, but with optimal confidentiality, was initiated.

This Small Area spatial layer is located between the Enumeration Area and Sub Place in the geographical frame hierarchy. (Figure 1)

Figure 1: South African Geographical Hierarchy.



3 Spatial creation methodology

ESRI's Arc View 8.3, software acted as host for customised Visual Basic (VB) script, instating an algorithm based on a rule set, which resulted from the abovementioned study.

The automated spatial creation of the Small Area Layer was based on the principle of merging individual Enumeration Areas within the Enumeration Area spatial layer. Merging was based upon a unique code (Flag2) allocated to the Enumeration Area if it adhered to a certain attribute -and spatial requirements rule set. The rule set is as follows:

- The Enumeration Areas can only be merged if they are within same Sub Place.
- The Enumeration Areas can only be merged if they have the same Enumeration Area geography attribute type.
- An Enumeration Area can only be merged if its population is less than 500.
- The resulting Small Area Layer (SAL) polygons must have a population total of 500 and more.

The spatial layers used as base entities for the new dissemination layer were the Enumeration Area and Sub Place spatial layer for the full extent of South Africa. The spatial attributes used as base variables for Enumeration Area aggregation were the adjusted population total of the Enumeration Area, the Enumeration Area Geography type and the Sub Place code linked to the Enumeration Area. The Enumeration Area Geography types are Urban Formal, Urban Informal, Farms and Traditional (Tribal) Areas.

The following principles were used as basis for testing and refining each version of the created spatial dataset:

As the objective of the exercise was the creation of a spatial layer with the optimal confidentiality possible, the maximum count of new Small Area polygons had to have a population total of above 500. This implied that the maximum number of Small Area polygons with a population total of less than 500 must be equal to the number of Sub Places with a population total of less than 500, as the new polygons are to be contained within the existing Sub Place polygons. (The total of which is 7620.)

As Statistics South Africa was striving to provide census data relating to a spatial layer as spatially similar to the existing Enumeration Area spatial layer as possible, a Small Area Layer polygon count as close as possible to 80787 (Enumeration Area count) was required, within the constraints of the rule set. To achieve this, population totals for Small Area Layer polygons exceeding 500, must exceed 500 with the least amount possible. It was also required that the final workflow resulted in the optimum average and standard deviation of population totals per Small Area Layer.

As the level of adherence to the rule set was determined by the spatial and attribute composition of Enumeration Areas within a Sub Place, a more controlled and structured processing workflow was implemented by dividing the national dataset into separate databases, based on their unique spatial characteristics and population attributes.

This implied that through the process of elimination more detailed routines could be specifically developed for more spatially complex Enumeration Area datasets, while Enumeration Areas for which the eventual Small Area Layer status could be accurately predicted could be processed and removed from the national dataset early on.

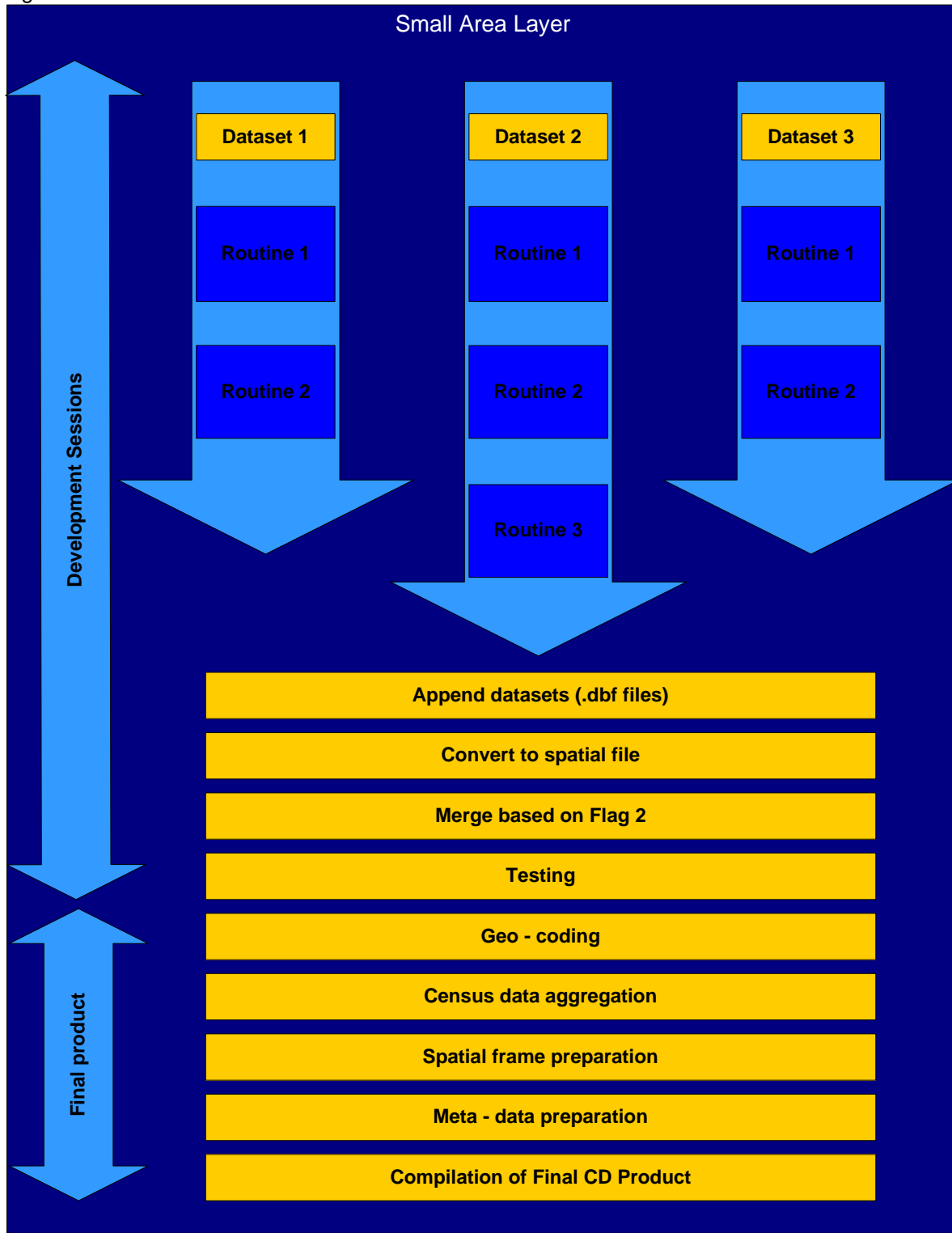
The following datasets were all processed during separate sessions.

- Dataset 1: Dataset 1 consisted out of all the Enumeration Areas in Sub Places (multi and single part) where the inherent population count of the Sub Place is less than 500.
- Dataset 2: Dataset 2 consisted out of all the Enumeration Areas in Sub Places where the Sub Place is spatially a multi polygon and the inherent population count is more than 500.
- Dataset 3: Dataset 3 consisted out of the remainder of the original national Enumeration Area dataset.

4 Detailed Small Area Creation routines:

The following conceptual models and detailed workflows stipulate the detailed routines used to develop the Small Area Layer.

Figure 2: Collective workflow:



The above conceptual model indicates the collective processes required to develop the Small, Area Layer.

4.1 Dataset 1 - Routine 1: (Figure 3)

- From the national spatial dataset select all the Enumeration Areas, which have a 1 to 1 relationship with their relevant Sub Places.
- These Enumeration Areas, where the population total is less than 500 were used as they are for the Small Area layer.
- List all Small Areas created in this routine by Enumeration Area.

4.2 Dataset 1 - Routine 2: (Figure 3)

- Merge all remaining Enumeration Areas within their applicable Sub Places with each other, irrespective of type.
- Flag the merge of different Geography types.
- List all Small Areas created in this routine by Enumeration Area.

Figure 3: Dataset 1 process flow.

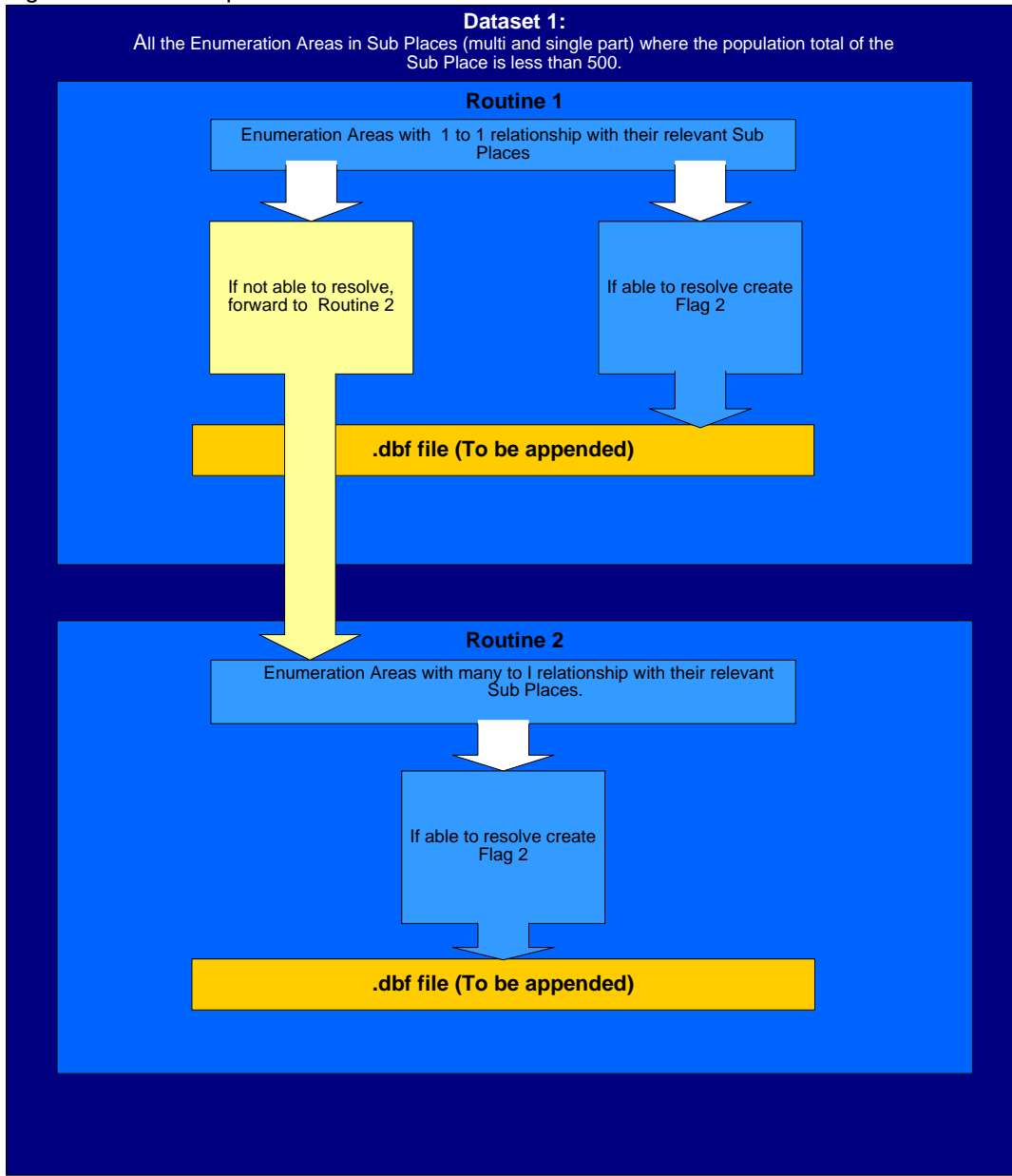
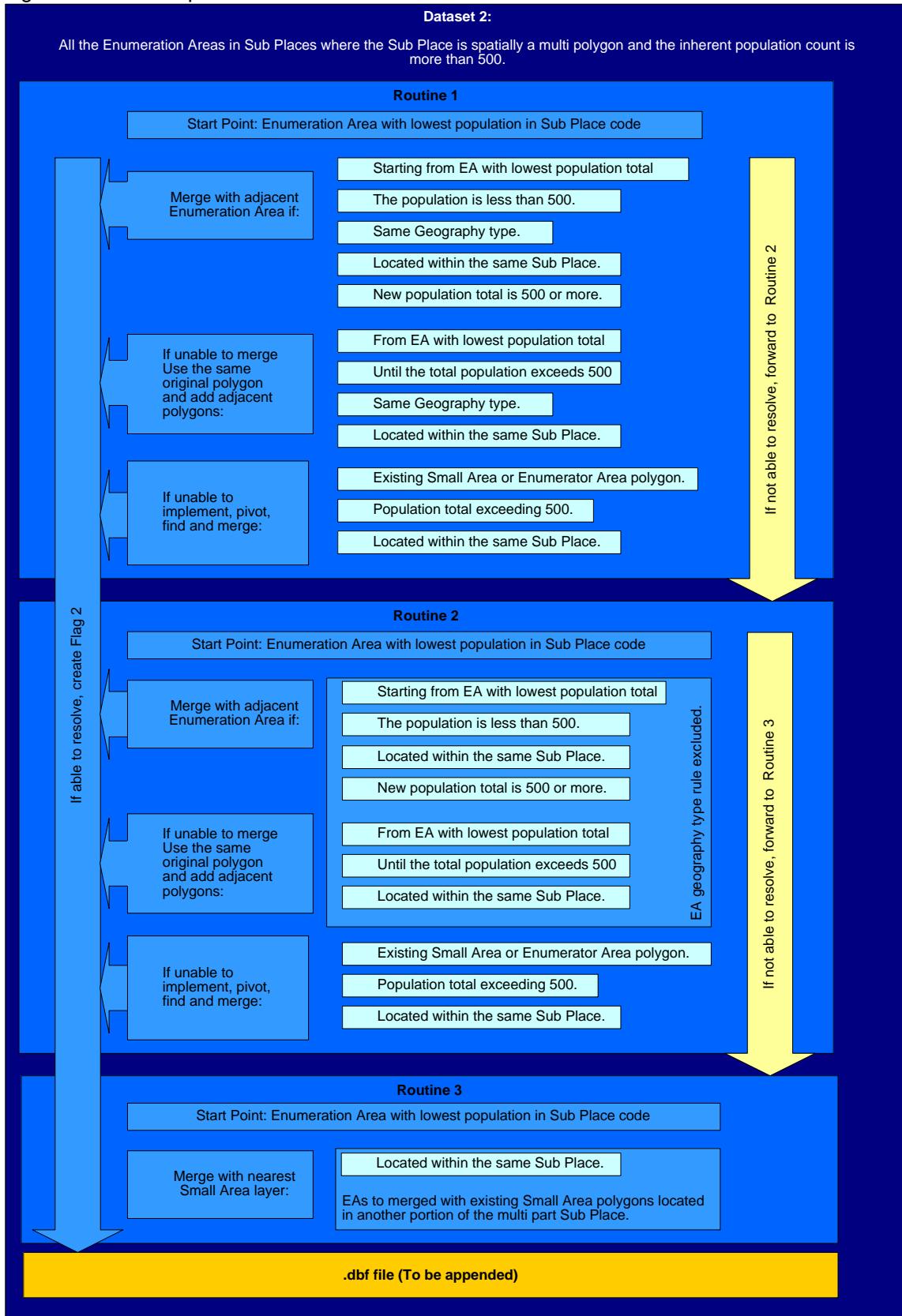


Figure 4: Dataset 2 process flow.



4.3 Dataset 2 - Routine 1: (Figure 4)

- Start Point: Enumeration Area with lowest population in Sub Place code
- Merge with adjacent Enumeration Area if:
 - The population of the adjacent Enumeration Area is less than 500, starting of from the lowest population polygon.
 - The adjacent Enumeration Area has the same Geography type.
 - The adjacent Enumeration Area is located within the same Sub Place.
 - The sum of the population total for the merged Enumeration Areas is 500 or more.
- If unable to merge:
 - Use the same original polygon and add adjacent polygons, which have a population total of < 500 by starting from the lowest adjacent polygon until the total population exceeds 500. (Do not add polygon if pop tot >500).
 - The adjacent Enumeration Area must have the same Geography type.
 - The adjacent Enumeration Area must be located within the same Sub Place
 - The sum of the population total for merged Enumeration Areas must be 500 or more.
- If unable to implement the routine, then pivot and find an already existing Small Area or Enumeration Area polygon with a population total exceeding 500 and merge.
- Repeat until no more merge options are left in the sub place.
- Keep unmerged polygons for next routine.
- List all Small Areas created in this routine by Enumeration Area.

4.4 Dataset 2 - Routine 2: (Figure 4)

- The rule set is similar to that of routine 1 except for the rule on Enumeration Area Geography type, which is now excluded.
- Start Point: Of all unmerged polygons start with the Enumeration Area with the lowest population in an applicable Sub Place.
- Merge with adjacent Enumeration Area if:
 - The population of the adjacent Enumeration Area is less than 500, starting of from the lowest population polygon.
 - The adjacent Enumeration Area is located within the same Sub Place.
 - The sum of the population total for the merged Enumeration Areas is 500 or more.
- If unable to merge:
 - Use the same original polygon and add adjacent polygons, which have a population total of < 500 by starting from the lowest adjacent polygon until the total population exceeds 500. (Do not add polygon if pop tot >500).
 - The adjacent Enumeration Area must be located within the same Sub Place
 - The sum of the population total for merged Enumeration Areas must be 500 or more.
- If unable to implement the routine, then pivot and find an already existing Small Area or Enumeration Area polygon with a population total exceeding 500 and merge.
- Repeat until no more merge options are left in the Sub Place.
- Keep unmerged polygons for next routine.

- List all Small Areas created in this routine by Enumeration Area.

4.5 Dataset 2 - Routine 3: (Figure 4)

- This is the only routine where Enumeration Areas are merged with existing Small Area polygons located in another portion of the multi part sub place.
- Merge all remaining Enumeration Areas in same multi Sub Place polygons with nearest Small Area of same Sub Place.
- List all Small Areas created in this routine by Enumeration Area.
If there are Enumeration Areas that are not solved by the above routines, flag as such. (See paragraph on exceptions under Quality assurance - header 6)

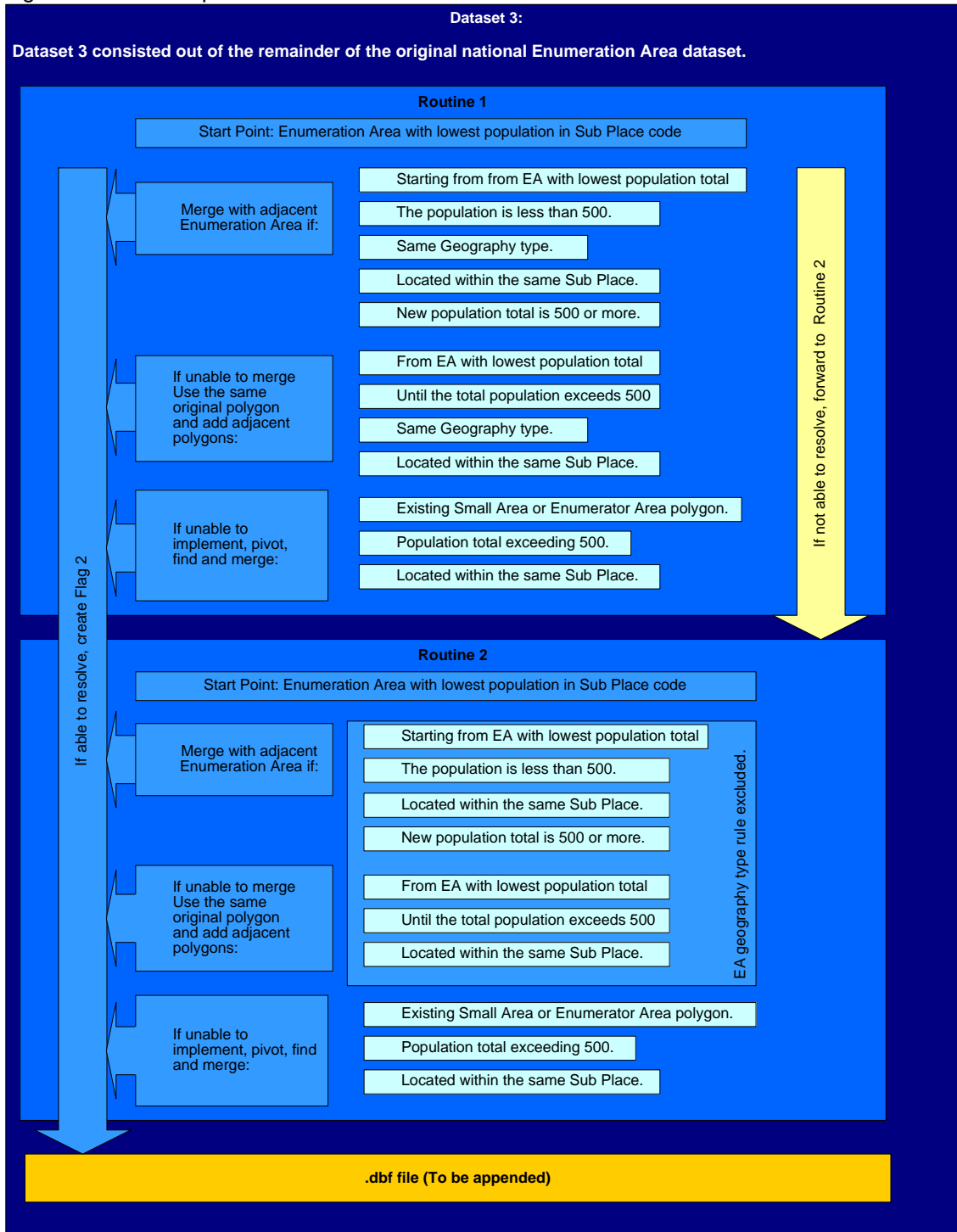
4.6 Dataset 3 - Routine 1: (Figure 5)

- Start Point: Enumeration Area with lowest population in Sub Place code.
- Merge with adjacent Enumeration Area if:
 - The population of the adjacent Enumeration Area is less than 500, starting of from the lowest population polygon.
 - The adjacent Enumeration Area has the same Geography type.
 - The adjacent Enumeration Area is located within the same Sub Place
 - The sum of the population total for the merged Enumeration Areas is 500 or more.
- If unable to merge:
 - Use the same original polygon and add adjacent polygons, which have a population total of < 500 by starting from the lowest adjacent polygon until the total population exceeds 500. (Do not add polygon if pop tot >500).
 - The adjacent Enumeration Area must have the same Geography type.
 - The adjacent Enumeration Area must be located within the same Sub Place
 - The sum of the population total for merged Enumeration Areas must be 500 or more.
- If unable to implement the routine, then pivot and find an already existing Small Area or Enumeration Area polygon with a population total exceeding 500 and merge.
- Repeat until no more merge options are left in the sub place.
- Keep unmerged polygons for next routine.
- List all Small Areas created in this routine by Enumeration Area.

4.7 Dataset 3 - Routine 2: (Figure 5)

- The rule set is similar to that of routine 1 except for the rule on Enumeration Area Geography type, which is now excluded.
- Start Point: Of all unmerged polygons start with the Enumeration Area with the lowest population in an applicable Sub Place.
- Merge with adjacent Enumeration Area if:
 - The population of the adjacent Enumeration Area is less than 500, starting of from the lowest population polygon.
 - The adjacent Enumeration Area is located within the same Sub Place
 - The sum of the population total for the merged Enumeration Areas is 500 or more.

Figure 5: Dataset 3 process flow.



- If unable to merge:
 - Use the same original polygon and add adjacent polygons, which have a population total of < 500 by starting from the lowest adjacent polygon until the total population exceeds 500. (Do not add polygon if pop tot >500).
 - The adjacent Enumeration Area must be located within the same Sub Place
 - The sum of the population total for merged Enumeration Areas must be 500 or more.

- If unable to implement the routine, then pivot and find an already existing Small Area or Enumeration Area polygon with a population total exceeding 500 and merge.
- Repeat until no more merge options are left in the Sub Place.
- List all Small Areas created in this routine by Enumeration Area.
- If there are Enumeration Areas that are not solved by the above routines, flag as such. (See paragraph on exceptions under Quality assurance - header 6)

5 Spatial layer construction

In order to report on which Enumeration Area was included in which Small Area at what stage of the algorithm, a .txt file was generated during each routine.

All the routines followed for the three separate datasets takes shape in a .dbf file containing the original Enumeration Area number, the new Small Area layer code (Flag 2) and an attribute indicating the original dataset.

Appending the datasets and then joining them by Enumeration Area code to an existing spatial Enumeration Area dataset of South Africa created a spatial database created for final spatial integration. This implied merging Enumeration Areas where Flag 2 is the same and keeping record of the total population count for each new Small Area layer polygon. This spatial layer was then subjected to a quality assurance procedure.

6 Quality assurance:

The following automated tests were done to interrogate the database in terms of fixed variables known to Statistics South Africa.

Database queries to confirm that all Enumeration Areas (80787) were included in the Small Area Layer.
Database queries to determine if any of the Enumeration Areas were allocated to more than one Small Area Layer.

Database queries to determine if there were 7620 Small Area polygons for Sub Places with a population total of less than 500.

Database queries to determine if the new Small Area codes (Flag 2) were unique in terms of our whole Geographical hierarchy (Sub Place, Main Place, Municipality, District Council and Province) and if all totals for census variables added up based on the Small Area Layer code for the geographical hierarchy.

Spatial interrogation (manual testing) of all versions of the Small Area layer was done to determine if the prescribed rule set was adhered to.

Through this phase of quality assurance exceptions were encountered. Exceptions are spatial entities, which are spatially and attribute wise so configured that the rule set was unable to yield a Small Area Layer polygon during any of the routines. The routines followed resulted in a total of a hundred and seventy six (176) Sub Places being flagged as exceptions. A process of manual intervention was followed to correct the anomaly. Manual intervention entailed the renaming of the Flag 2 field to the lowest Enumeration Area code within the Sub Place and then merging all the Enumeration Areas within the Sub Place based on the flag code.

A final analysis was done based on basic statistical concepts namely average and standard deviation. Through out all versions of the Small Area Layer the average population count and standard deviation for all Small Area Layer polygons was monitored. This implied that if there was an increase in both or any of the two, for a new version of the Small Area Layer, the process or rule set had to be reviewed.

7 Small Area Layer numbering methodology:

The unique Enumeration Area number was used as basis for the allocation of Flag 2 numbers. Enumeration Areas belonging to a specific Small Area were all flagged with the Enumeration Area number of the first Enumeration Area used in creating the polygon. This is always the Enumeration Area number of the Enumeration Area with the lowest population in the Small Area. (See Detailed Small Area creation routines for datasets 2 - and 3 on pages 6 to 10).

The basis for the unique geo-code for each small area polygon is the code of the Sub Place in which it is located. Three additional digits determined by location within a Sub Place, were added to the Sub Place code to ensure uniqueness.

- Main Place (MP) 10101 1|2|3|4|5|
- Sub Place (SP) 10101003 1|2|3|4|5|6|7|8|
- Small Area Layer (SAL) 10101003123 1|2|3|4|5|6|7|8|9|10|11|

The first digit denotes the province, the second and third digits denote the municipality, the fourth and fifth digits identify the main place, the sixth, seventh and eighth digits identify a unique Sub Place within the main place and the ninth, tenth and eleventh digits identify a unique small area polygon within the Sub Place.

The allocation of unique identifiers within sub places in order to facilitate geo-coding was based upon the location of the Small Area Layer within the Sub Place. The spatial rule set pertaining to Geo-coding is set out as follows:

- The starting point of geo coding was the lowest left (South East) corner of a Sub Place.
- The next number was allocated to the Small Area Layer polygon to the right of the first Small Area Layer code and continued in this fashion until it reached the eastern boundary of the Sub Place.
- The application then looked for the Small Area Layer polygon to the north of it, numbered it, and continued to the left until it reached the western boundary of the Sub Place.
- Repeat until it finishes the Sub Place.

8 Spatial Data synopsis

The final version of the Small Area Layer contains a total of 56255 polygons. A total of 40742 Enumeration Areas have a 1 to 1 relationship with the new Small Area layer and a total of 40045 Enumeration Areas are contained in 15513 Small Areas in a many to 1 relationship. This implies that 40742 Enumeration Areas were used in their original state and a total of 40045 Enumeration Areas were merged into 15513 Small Area polygons.

8.1 Table 1: Enumeration Area confidentiality breached per province.

Province	Total Breached	Total Maintained	Total	% Breached
Western Cape	344	4995	5339	6.44%
Eastern Cape	4556	6458	11014	41.37%
Northern Cape	114	909	1023	11.14%
Free State	195	3268	3463	5.63%
KwaZulu Natal	818	9348	10166	8.05%
North West	342	4109	4451	7.68%
Gauteng	472	9705	10177	4.64%
Mpumalanga	222	3656	3878	5.72%
Limpopo	557	6187	6744	8.26%
TOTAL	7620	48635	56255	13.55%

Confidentiality was maintained for 48635 Small Area Layer polygons (86.45%), while for 7620 Small Area Layer polygons (13.55%) it was breached.

8.2 Table 2: Confidentiality breached per spatial hierarchy.

Spatial Layer	Polygon count: Population < 500	Total entity count	Percentage
Enumeration Area	39997	80787	49.50%
Sub Place	7620	21243	35.87%
Small Area Layer	7620	56255	13.54%

If a comparison based on the confidentiality breached per spatial layer is done, it is clear that there is a definite decline in the percentage of spatial entities where confidentiality is breached when comparing the new Small area layer to the Enumeration Area and Sub Place spatial layers.

8.3 Table 3: Geo-codes per municipality.

Province	Enumeration Area	Sub Place	Small Area Layer
Western Cape	7101	1340	5339
Eastern Cape	18370	7958	11014
Northern Cape	1509	440	1023
Free State	5183	791	3463
KwaZulu Natal	12752	4030	10166
North West	6477	1201	4451
Gauteng	13202	2222	10177
Mpumalanga	5728	891	3878
Limpopo	10465	2370	6744
TOTAL	80787	21243	56255

The comparative totals of the total of geo codes for the Enumeration Area, Sub Place and Small Area Layer makes it clear that the Small Area Layer is much more detailed than the previous lowest level of dissemination namely the Sub Place.

9 Conclusion

From the above mentioned spatial data synopsis it is clear that the Small Area Layer significantly contributed, although not fully, to the improvement of confidentiality through the implementation of the prescribed rule set.

The Small Area Layer therefore cannot be seen as the optimal solution for data dissemination not even to mention the possibility of using it for data collection - It is merely the result of an effort to bridge the gap between user needs and methodological and legislative constraints.

With the eye on Census 2011, refinement of a policy regarding geographical levels of dissemination need to be addressed within the context of the following question: Does the process of formulation of the policy lie with the output area or input areas? Input areas are decisive in determining survey methodology and must take heed of needs that products derived from the census should satisfy, while dissemination areas are the entities through which products are delivered to users (Phirwa: 2004).

It is not only with anticipation that we are looking forward to the demarcation of Enumeration Areas for the next census, but also with some trepidation, as we know that these units will have to satisfy dissemination requirements within methodological and legislative constraints.

10 Reference

- “The creation of a dissemination layer for Census 2001 information”, 2005. Nick Grobbelaar. Statistics South Africa.
- “Small Area Dissemination – Housing and Population Census”, 2004. Motale Phirwa Statistics South Africa.

11 Collaboration:

Pradeep Mahabeer (project manager) and Kashmira Beharee (application developer) from Data World were tasked to implement the VB script for the automated merging of Enumeration Areas based on the development rule set.