TESTING THE SPATIAL ADJACENCY MATCH OF THE INTIENDO ADDRESS MATCHING TOOL FOR GEOCODING OF ADDRESSES WITH MISLEADING SUBURB OR PLACE NAMES

Serena Coetzee
Dept Computer Science, University of Pretoria
Pretoria, 0002, South Africa
scoetzee@cs.up.ac.za

Magnus Rademeyer
AfriGIS, PO Box 14134, Hatfield
Pretoria, 0028, South Africa
magnus@afrigis.co.za

Abstract

Geocoding refers to the process of assigning geographic identifiers and/or geographic coordinates to the description of a feature location. Address matching is the specific case of geocoding where the description of a feature location is an address, to which an address from a reference dataset is assigned. Address matching is complicated by an incomplete or inaccurate input address, or one that includes a misleading address element, such as an incorrect suburb or place name. The cause of such an input address is often due to the ambiguities originating from uncertainties regarding suburb and/or place name boundaries. Relaxing the requirement to match the suburb accurately, adds more addresses to the list of potential matches to choose from, but with the increased risk of geocoding the input address to a spatially removed suburb or place with the same street name and number as the input address.

The Intiendo (Spanish for 'I understand') address matching tool is a software toolset that is used to structure and geocode addresses, i.e. to understand and interpret addresses. Intiendo employs multiple search algorithms during geocoding. Alphanumeric string matching between the input address and addresses in the reference dataset is done based on the edit distance and numeric distance algorithms. This matching resolves spelling variations between different sources of addresses very well, but does not address the problem of misleading suburb or place names in an input address. Intiendo implements a novel spatial adjacency match that probes the area around a potential address match when it is suspected that the input address contains a misleading suburb or place name. In this way, Intiendo, improves the chances of matching the input address to the correct address in an adjacent suburb or place, instead of an address in a remote suburb with the same street name and number.

In this paper we present the methodology and results of a geocoding test during which a sample dataset is geocoded twice: once with the novel spatial adjacency match enabled

and once with the spatial adjacency match disabled. We present the results of the geocoding test and provide an analysis and discussion of the results. The results confirm that address geocoding with the novel spatial adjacency match allows more input addresses to be matched more accurately to addresses in the reference dataset, and thus improves the quality of the results. In other words, more addresses are geocoded more accurately. The results further reveal that the spatial adjacency match delivers the best results when geocoding input addresses that include misleading address elements, such as an incorrect suburb or place name. Our discussion of the results points out current shortcomings of the spatial adjacency match for which we provide suggestions for improvement and we give ideas for future work.

The objectives of this paper are to present the Intiendo address matching algorithm; to describe the methodology of the geocoding test for the comparison of address geocoding with and without the spatial adjacency match; to discuss and analyze the results of the geocoding test; and to present our conclusions from the results of the geocoding test.

1. Introduction

Geocoding refers to the process of assigning geographic identifiers and/or geographic coordinates to the description of a feature location. Address matching is the specific case of geocoding where the description of a feature location is an address, to which an address from a reference dataset is assigned. A geocoding algorithm takes an address as input and returns either a matching address from a reference dataset, or returns false, indicating that the address could not be matched. Typically, coordinates for the matching address are also returned.

There are similarities between address geocoding and entity resolution for data integration. Sehgal *et al.* (2006), Fu *et al.* (2005) report on entity resolution, which is the process of determining a single consolidated collection of ‘true’ locations from a collection of sources referring to geospatial locations. The target of entity resolution is to match locations from different sources that correspond to the same real-world location. For this, the location name, the coordinates, the location type and the hierarchy of location names can be used to match pairs from different sources. Geocoding can be seen as a special case of entity resolution where only two sources are involved, the location type is always ‘address’ and the input address in most cases lacks coordinates. There is similar overlap with research on the conflation of gazetteers (Hastings, 2008).

Addresses are often structured into a spatial hierarchy that describes a location with increasing levels of detail. In the address ’14 Richmond Road, Mowbray, Cape Town, South Africa’, the levels of detail increase from country (South Africa) to city (Cape Town) to suburb (Mowbray) to street (Richmond Road) to street number (14). The international standard, ISO 19112:2003, *Geographic Information – Spatial referencing by geographic identifiers*, provides a general model for spatial referencing using

geographic identifiers and defines the components of a spatial reference system. This general model is applicable to an address structured into a spatial hierarchy. In a previous paper (Rahed *et al.* 2008), we compared the ISO 19112 and Intiendo data models.

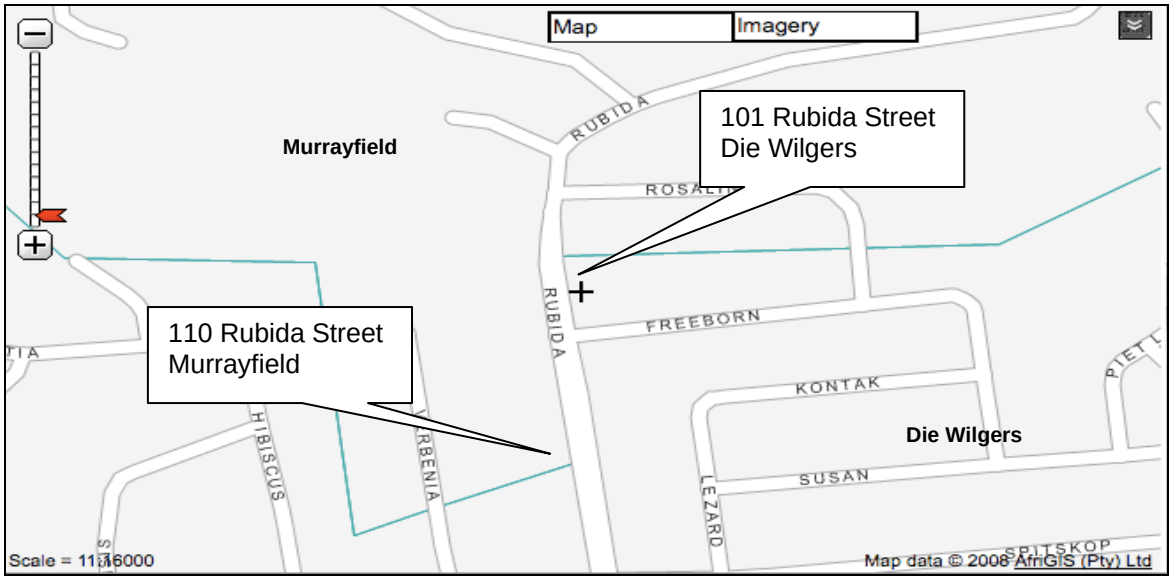

Figure 1. Addresses on the boundary between 'Murrayfield' and 'Die Wilgers

The process of address matching is complicated by an input address that is incomplete or inaccurate, or one that contains a misleading geographic identifier in its location type hierarchy. The cause of such an input address is often due to the ambiguities originating from uncertainties regarding suburb boundaries. While a single set of official place name boundaries for a country can reduce such ambiguities, Coetzee and Cooper (2007) point out that one will never get rid of these ambiguities because suburb or place name boundaries are not the type of boundary to be physically fenced off and hence obvious for all to see. There is also always the human ego factor that sees a person, living near the boundary of a more prestigious suburb, use the name of that suburb in their address. '101 Rubida Street, Murrayfield' is an example of an input address with a misleading suburb or place name. A geocoding algorithm that employs an alphanumeric matching approach will incorrectly match this address to '110 Rubida Street, Murrayfield', and not to the more accurate '101 Rubida Street, Die Wilgers' on the opposite side of the road, albeit in a different suburb. Refer to Figure 1.

Relaxing the requirement to match the suburb adds '101 Rubida Street, Die Wilgers', to a list of potential matches, as well as '101 Rubida Street, Rondebosch' and '101 Rubida Street, Wilgenhof', which are from distant places in the country. A geocoding algorithm has to determine which one of these potential addresses is the best match. While 'Die Wilgers' and 'Wilgenhof' are closer in terms of string matching, they are spatially much further apart than 'Murrayfield' and 'Die Wilgers'. In their survey on the field of geocoding, Goldberg et al. (2007) list attribute relaxation as one of the common causes

of error in the matching stage of the geocoding process. Davis and Fonseca (2007) propose a so-called geocoding certainty indicator (GCI) that takes into consideration the spatial transformations that an address record goes through during the matching, and the approximations used to match the input address with an existing address in the reference dataset. This indicator considers alphanumeric proximity of suburbs (based on string matching) but not spatial proximity.

The Intiendo address matching tool employs two types of matching in its geocoding algorithm. Firstly, alphanumeric string matching based on the edit distance algorithm considers alphanumeric proximity and is therefore comparable to the CGI that Davis and Fonseca (2007) propose. Secondly, in certain cases Intiendo considers addresses in spatial proximity to a potential address match that might be 'misled' by misleading information in the source address. In our previous paper (Rahed *et al.* 2008), we presented the implementation of the spatial adjacency match based on a kd-tree (k-dimensional tree).

The objectives of this paper are to present the Intiendo address matching algorithm; to describe the methodology of the geocoding test for the comparison of address geocoding with and without the spatial adjacency match; to discuss and analyze the results of the geocoding test; and to present our conclusions from the results of the geocoding test.

2. Address matching in Intiendo

Intiendo (Spanish for 'I understand') is a software toolset that is used to structure and geocode addresses, i.e. to understand and interpret addresses. Intiendo has been used to successfully geocode large volumes (5 million address records or more per dataset) of diverse address datasets for a number of clients in both the private and the public sector.

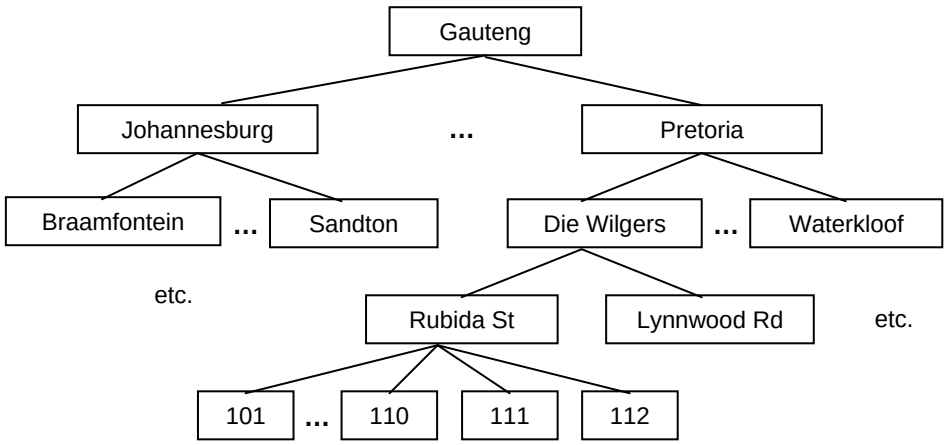

Figure 2. Data items from a SANS 1883 Intiendo HDB

Intiendo is based on the principle that an address has a hierarchical pattern, i.e. the street number belongs to a street, the street to a suburb, town, province, etc. An Intiendo hierarchy consists of a number of levels, forming a hierarchy. Before address matching can take place, an address reference dataset is converted into a so-called Intiendo hierarchy database (HDB. Input addresses, either in free format or in a number of address lines, are parsed, structured and matched against the Intiendo HDB. For example, the Street Address type in the South African address standard (SANS 1883:2009) has levels for the Province, Town, PlaceName, CompleteStreetName, and CompleteStreetNumber. Sample items are shown in Figure 2.

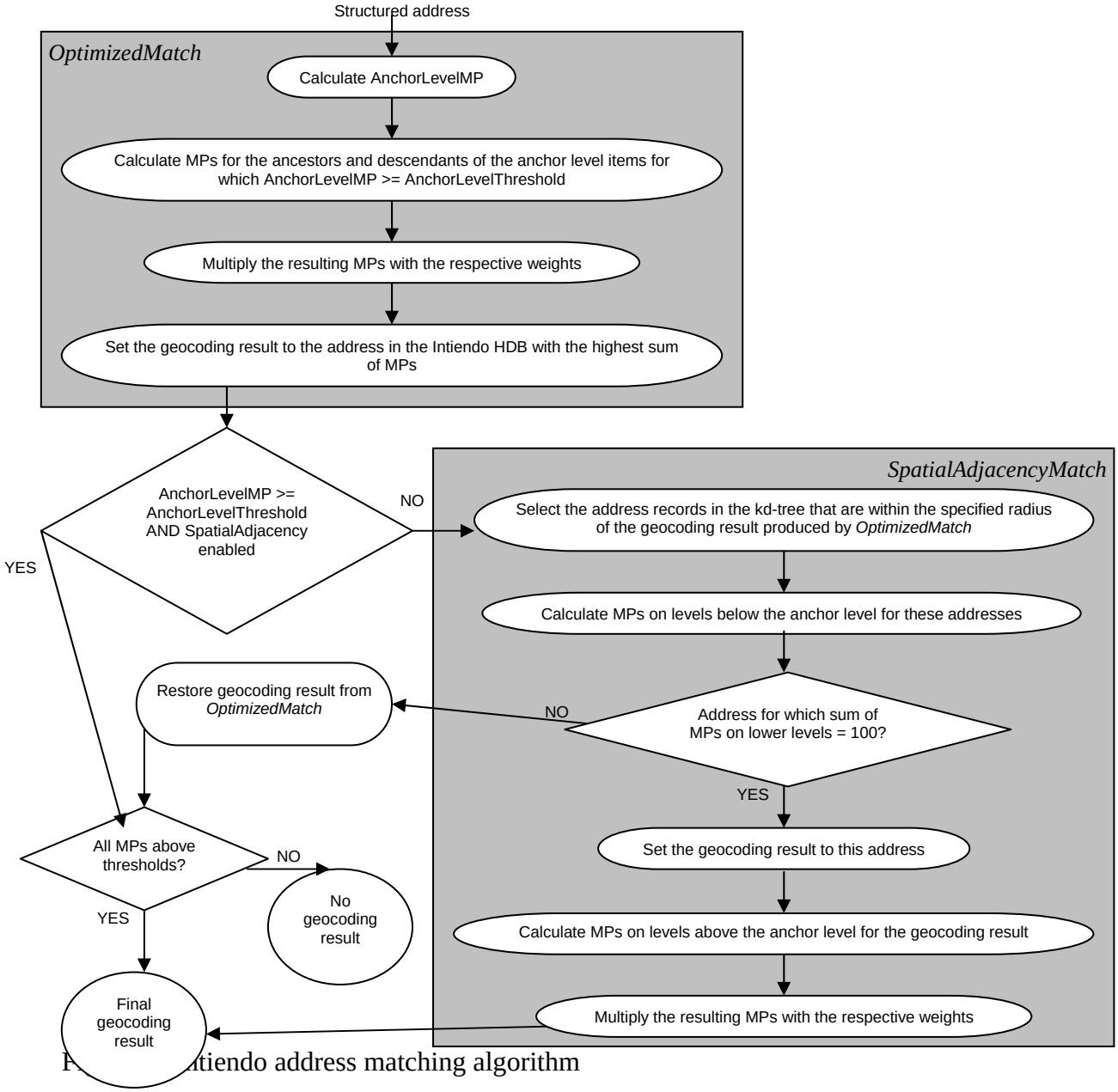

Figure 2: Intiendo address matching algorithm

Before the actual Intiendo address matching starts, input addresses are parsed and structured to fit the Intiendo HDB against which matching will take place. The Intiendo address matching algorithm is illustrated in Figure 3. The first step of the address matching process is the *OptimizedMatch*, and depending on the outcome of this match, the algorithm either ends with a geocoded output address, or a further *SpatialAdjacencyMatch* is attempted in order to find a more suitable match.

OptimizedMatch will first try to match the information on the anchor level of an input address with items of the corresponding level in the Intiendo HDB. The matching percentage (MP) between an input address element and an item in the Intiendo HDB is calculated based on the edit distance algorithm. Only the ancestors and descendants of those items in the HDB for which the AnchorLevelMP is above the AnchorLevelThreshold are included in the subsequent search refinement. *OptimizedMatch* discards geocoding results if the MP on of the levels is below the threshold set for that level.

SpatialAdjacencyMatch builds searches addresses in the Intiendo HDB that are within a specified radius from the geocoding result provided by the *OptimizedMatch*. If a match is found on the levels lower than the anchor level, this address is set as the geocoding result and returned. For example, if the anchor level is set to the suburb, it searches for a matching street name and street number. This increases the chances of finding a better address match in an adjacent suburb. If no match is found, the geocoding result from the *OptimizedMatch* is restored and returned as the final geocoding result. The details of the minimum bounding rectangle (MBR) and kd-tree implementation in the Intiendo *SpatialAdjacencyMatch* search are discussed in Rahed *et al.* (2008).

3. Methodology

We conducted a geocoding test in Intiendo in order to evaluate geocoding results that incorporate Intiendo's *SpatialAdjacencyMatch*. For this test we geocoded a sample dataset of addresses with misleading suburb names.

Province	Town	Suburb	Street name	Street number
Gauteng	Johannesburg	Saxonwold	Engelwold Road	19
Gauteng	Pretoria	Atteridgeville	Sekukuni Street	104
Gauteng	Midrand	Noordwyk	Sagewood Avenue	637

Table 1. A few addresses from the sample dataset

The sample dataset comprises 14,670 address records from South Africa. Sample address records from the dataset are listed in 0. The quality of address reference data in urban areas of South Africa is generally more accurate and complete. We wanted to isolate the test to misleading suburb names and eliminate problems resulting from

incomplete input addresses and/or incomplete reference data. Therefore, we selected the addresses mostly from urban areas and included only addresses for which at least the province, suburb, street name and street number are populated.

Variable	Value	Note
Anchor level	2	Suburb level
Radius	2km	Used to determine which addresses are included in the kd-tree durin the *SpatialAdjacencyMatch*.
Province threshold	85%	Allows limited variation between the province names in the customer address and the Intiendo hierarchy.
Town threshold	0%	Allows any variation between the town names. This is acceptable because town boundaries do not have any legal status, thus ambiguity is omnipresent.
Suburb threshold	87%	For the SR of our gecoding test, an AnchorLevelMP less than this threshold, also referred to as the AnchorLevelThreshold, causes the *SpatialAdjacencyMatch* to be executed.
Street name threshold	87%	Allows limited variation between the street number in the customer address and the street number in the Intiendo hierarchy.
Street number threshold	100%	Allows no variation between the street number in the customer address and the street number in the Intiendo hierarchy.
Weights	1	All weights are set to 1 so that they do not have any effect.
Ignore street type	True	The street type is not included when the MP is calculated between the street name in the customer address and the street name in the Intiendo hierarchy.

Table 2. Intiendo settings for the geocoding test

The selected addresses are a subset of a larger real-world dataset of customer addresses. The combination of province and suburb was validated when the customer addresses were captured, albeit against a different reference dataset than the one we used for this geocoding test. Therefore, there is variation between the suburb names in the customer dataset and the suburb names in the Intiendo HDB that was used for the geocoding in Intiendo: some suburb names exist in the dataset and not in the HDB, and vice versa, and there are spelling variations between the dataset and the HDB. The combination of suburb with street name and number was not validated when the customer addresses were captured. This implies that customers could have entered incorrect suburb names for their respective street names and street numbers. The customer's street name and number were entered in free format, resulting in spelling errors in the street name, as well as possibly invalid street names and numbers. Thus, the chosen sample dataset of customer addresses includes misleading suburb names.

The Intiendo HDB used for the geocoding test comprises the AfriGIS national address dataset, a dataset that is compiled by aggregating address data from multiple municipalities in South Africa. The hierarchy has five levels: province, town, suburb, street name and street number, with data items similar to the ones the displayed in Figure 2. The sample dataset was geocoded twice with Intiendo: once with the *SpatialAdjacencyMatch* disabled and once with the *SpatialAdjacencyMatch* enabled. We refer to these two runs as the non-spatial run (NSR) and the spatial run (SR) respectively. For the NSR, only the *OptimizedMatch* is executed, while for the NSR, the

SpatialAdjacencyMatch is executed if AnchorLevelMP <= AnchorLevelThreshold. Refer to the Intiendo address matching algorithm in Figure 3. For both runs the anchor level was set to 2, i.e. the suburb level. The complete list of Intiendo settings for the geocoding test is presented in 0 above.

4. Discussion of results

The results for the SR and NSR are listed in 0 below. From this we can see that SR produced 8,905 geocoding results, i.e. 8,905 or 61% of the customer addresses could be matched with sufficicient confidence to an address in the Intiendo HDB. For the NSR this figure is 3% less at 8,514. Thus, the SR has produced more results.

	SR	**NSR**
Customer address records	14,670	14,670
Matched address records (geocoding results)	8,905 (61%)	8,514 (58%)
Non-matched address records	5,765 (39%)	6,156 (42%)

Table 3. Geocoding results for SR and NSR

We use a specific example from the gecooding test to show why SR produced more results. The NSR geocoding result listed in line 2 of the 0 was discarded because of the 100% threshold value on the street number. MPs are shown in brackets. Line 3 shows the result produced by the SR run. Ignoring the suburb level MP (with the *OptimizedMatch*) would have produced the same result, but the *SpatialAdjacencyMatch* increases the confidence level because the matched address is now known to be in proximity of New Redruth. The adjacent suburbs are shown in Figure 4. Thus, this example confirms that *SpatialAdjacencyMatch* increases the total number of geocoding results.

	Source	**Province**	**Town**	**Suburb**	**Street name**	**Street number**
1	Input	Gauteng	Alberton	New Redruth	Voortrekker Road	16
2	NSR	Gauteng (100%)	Alberton (100%)	New Redruth (100%)	Voortrekker Road (100%)	35 (96%)
3	SR	Gauteng (100%)	Alberton (100%)	South Crest (44%)	Voortrekker Road (100%)	35 (100%)

Table 4. A few addresses from the sample dataset

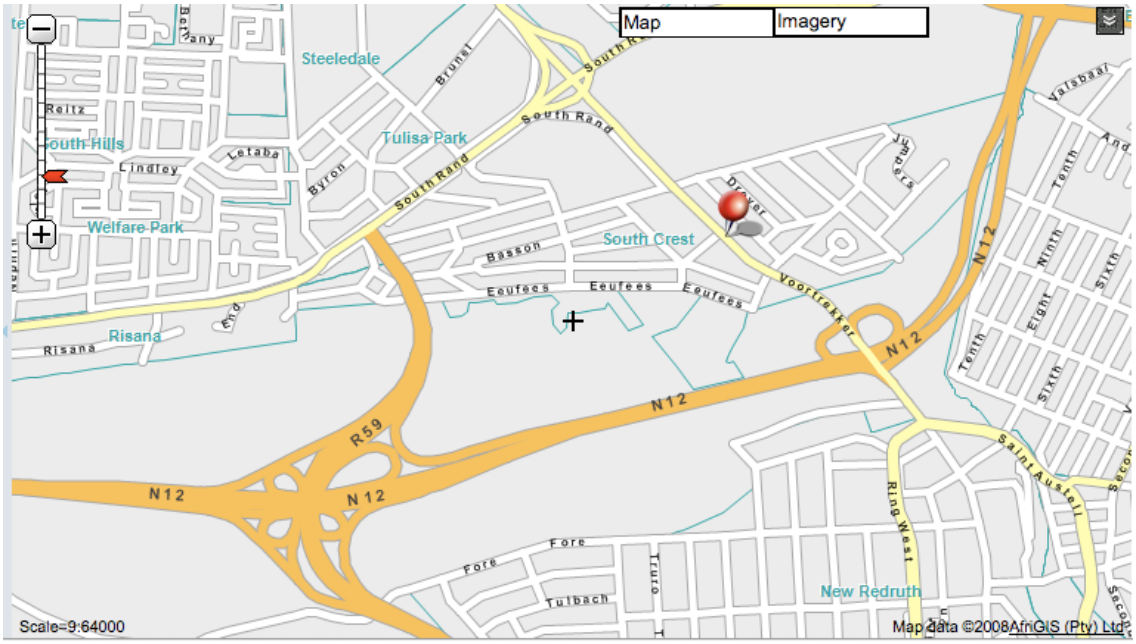

Figure 4. 'South Crest' and 'New Redruth' suburbs in Alberton

We conducted the geocoding test with a threshold value of 100% for the street number level. Had we relaxed this threshold to 90%, which is still very high, the NSR geocoding result of Line 2 of 0 would not have been discarded. The SR geocoding result would still be the same, thus the *SpatialAdjacencyMatch* has the potential to improve the quality of geocoding results.

SpatialAdjacencyMatch. The percentage improvement resulting from the *SpatialAdjacencyMatch* is relatively low at 3%, but depending on the purpose and quality requirements of the geocoding, 3% can constitute a significant improvement. For example, a high way separates the adjacent suburbs of 'South Crest' and 'New Redruth' in Figure 4. A highway sometimes acts as a natural barrier and the demographics of the two suburbs could differ significantly. In future, one should conduct more geocoding tests, similar to the one we reported on here, in order to get an idea of the average percentage improvement resulting from the *SpatialAdjacencyMatch.*

4. Conclusion

In this paper we presented the Intiendo address matching algorithm to show that it employs both alphanumeric matching of the strings in an address, as well as a spatial adjacency match. The latter also matches addresses within a specified radius from an alphanumerically matched address, if it is suspected that the input address includes misleading information. We described the methodology used for a geocoding test to compare address geocoding results with and without the *SpatialAdjacencyMatch* and discussed and analyzed the results of this test.

The results of the two runs, SR and NSR, in the geocoding test confirm that the

SpatialAdjacencyMatch of Intiendo improves the geocoding results in two ways. Firstly, the total number of geocoding results increases and secondly, the quality of the geocoding result is better. We suggested improving the *SpatialAdjacencyMatch* by dynamically resizing the area in which to search, and recommend that more tests are run to get a better idea of the average percentage improvement resulting from the *SpatialAdjacencyMatch*. The spatial adjacency matching technique described in this paper could also be tested for use in entity resolution and conflation of gazetteers.

Acknowledgements

The authors wish to thank Christopher Ueckermann from AfriGIS for running the geocoding test with Intiendo. Serena Coetzee's research on this work is supported in part by the South African Department of Trade and Industry (dti) and AfriGIS (Pty) Ltd. The Intiendo address matching toolset is owned and developed by AfriGIS.

References

Coetzee S and Cooper AK, What is an address in South Africa?, *South African Journal of Science (SAJS)*, Nov/Dec 2007, vol. 103, no. 11/12, pp449-458.

Davis AD Jr and Fonseca FT, 2007, Assessing the certainty of locations produced by an address geocoding system, *Geoinformatica*, **11**:103-129.

Fu G, Jones C and Abdelmoty AI, 2005, Building a geographical ontology for intelligent spatial search on the web, *Proceedings of IASTED International Conference on Databases and Applications*, Innsbruck, Austria, 2005, pp. 167–172.

Goldberg DW, Wilson JP and Knoblock CA, 2007, From text to geographic coordinates: The current state of geocoding, *Journal of the Urban and Regional Information Systems Association (URISA)*, vol. 19, no. 1, pp33-46.

Hastings JT, 2008, Automated conflatsion of digital gazetteer data, *International Journal of Geographical Information Science*, **22**(10), pp1109-1127.

ISO 19112:2003, *Geographic information – Spatial referencing by geographic identifiers*, 2003, International Organization for Standardization (ISO), Geneva, Switzerland.

SANS/CD 1883:2009. *Geographic Information – Address Standard* (draft standard), 2009, South African Bureau of Standards (SABS), Pretoria, South Africa.

Rahed AA, Coetzee S and Rademeyer M, 2008, A data model for efficient address data representation - Lessons learnt from the Intiendo address matching tool, *Proceedings of the academic track of the 2008 FOSS4G Conference, incorporating the GISSA 2008 Conference*, 29 September - 3 October 2008, Cape Town, South Africa.

Sehgal V, Getoor L and Viechnicki PD, 2006, Entity resolution in geospatial data integration, *Proceedings of the ACM-GIS '06*, 10-11 November 2006, Arlington, Virginia, USA.