A Workflow for Geocoding South African Addresses

Alexandria van Rensburg

Minor Dissertation presented in partial fulfilment of the requirements for the degree of Master of Philosophy in the Department of Computer Science

University of Cape Town

January 2015

Supervisor: Sonia Berman

Plagiarism Declaration

I know the meaning of plagiarism and declare that all of the work in this thesis, save for that which is properly acknowledged, is my own.

Signature: _____

Abstract

There are many industries that have long been utilizing Geographical Information Systems (GIS) for spatial analysis. In many parts of the world, it has gained less popularity because of inaccurate geocoding methods and a lack of data standardization. Commercial services can also be expensive and as such, smaller businesses have been reluctant to make a financial commitment to spatial analytics. This thesis discusses the challenges specific to South Africa as well as the challenges inherent in bad address data. The main goal of this research is to highlight the potential error rates of geocoded user-captured address data and to provide a workflow that can be followed to reduce the error rate without intensive manual data cleansing.

We developed a six step workflow and software package to prepare address data for spatial analysis and determine the potential error rate. We used three methods of geocoding: a gazetteer postal code file, a free web API and an international commercial product. To protect the privacy of the clients and the businesses, addresses were aggregated with precision to a postcode or suburb centroid. Geocoding results were analysed before and after each step. Two businesses were analysed, a mid-large scale business with a large structured client address database and a small private business with a 20 year old unstructured client address database. The companies are from two completely different industries, the larger being in the financial industry and the smaller company an independent magazine in publishing.

Both businesses were subject to address elementising, cleansing and standardization. Their data was mapped and displayed for all three geocoding methods, with all three showing significant error rates. Discrepancies between the three methods were quantified using the Great Circle Distance formula. We found in both instances that an acceptance threshold of 22.2km (0.2^{o}) is recommended to distinguish usable results from those requiring human intervention.

When the data was cleansed, and once there was a proven confidence level in the data, it could be mapped in various ways for the businesses to utilize. Demographics and marketing segments could be added to enhance the understanding of the existing client base.

Acknowledgements

I would like to thank my family and friends for their inspiration and never-ending faith in me. I would also like to express my gratitude to my supervisor, Sonia Berman, for her advice, her resolve, and her patience.

My sincere thanks also goes out to my colleagues and incredibly supportive company who have encouraged me every step of the way.

Lastly, my amazing friends Sharna and Heidi have been invaluably helpful in this process and are always there with moral support. I am whole heartedly appreciative.

Table of Contents

List of figures

List of tables

1 Introduction

As spatial analytics, geo-informatics, geo-demographics and other geography based analytics are starting to play a greater role in traditional analytics, many companies find themselves confronting the challenge of how to implement these new technologies with their current data set. IT departments often find themselves faced with the task of cleansing, standardising and managing these new systems. For many companies with sub-standard data, this task can be a limitation that prevents them from continuing further.

In order for any geographical analysis to be performed on a data set, its data needs to be geocoded, i.e., associated with latitude and longitude values giving its location. For most non-spatial or conventional data sets, there is only address data which can be used as a basis for geocoding. Geocoding presents a number of issues including varying precision levels in different parts of the world. Some international geocoders do not support South African address data at all, while other products do have support, but not at the precision level desired. Commercial geocoders can fill this gap in some areas, but can also be expensive to use.

We found South African address data to be inherently dirty, although it still is usable for business analytics if one understands the error rate and possible issues associated with the data. There is a growing trend that believes analytics can still have benefit even when the data isn't fully clean (Shacklett, 2014).

The list of some common problems that organisations currently face when attempting to geocode address data are:

a. Manually captured address that have incorrect or dirty data
b. Lack of capturing standards
c. Addresses coming from multiple sources
d. No singular source of address truth, i.e., postal delivery vs. utility service
e. Geocoding databases that have out of date, incomplete, or incorrect information for South Africa
f. Geocoders can provide conflicting information
g. Commercial geocoding services can be very expensive.
h. Privacy concerns when releasing address data to be cleaned or geocoded

The aim of this work is to provide a workflow and toolset that enables organisations to automatically geocode and display South African address data while tackling the issues of poor data quality, inconsistent geocoders, expense, and privacy concerns. The work seeks to show that organisations of any size can enhance traditional analytics with geospatial data regardless of these common entry barriers. The workflow allows organisations to tailor the degree of precision required for their business and identify records which can be improved with manual intervention.

The workflow is applied using actual data from two organisations of differing sizes and in different industries, with one employing rigorous data capture standards and the other not, in order to show that the process can be applied across various business types.

1.1 Scope

This work focuses on relatively coarse-grained geocoding of data sets, namely down to the level of postal code region and suburb or town, rather than delivery point or street addresses. In our study we make use of the data of two organisations which specifically asked to remain anonymous. The larger organisation is a financial institution that requested a non-disclosure agreement (NDA) for any identifying information for both clients and brokers. The smaller organisation is a publishing company with data from subscriptions as well as advertisers. Although they did not require an NDA, they did request, for privacy reasons, that the specific details of their subscribers and advertisers remain unidentifiable.

For both companies, decisions are based on analysis and understanding of clusters rather than individuals. Postal code region is then adequate for the purposes of both privacy and for the study of business behaviour. Mapping individual addresses onto their exact latitude-longitude values is thus not within the scope of this thesis.

User interface design in this field warrants an extensive separate study of its own. Thus interaction design and end-user evaluation of workflow usage are also beyond the scope of this thesis and left for future work.

1.2 Thesis outline

Chapter2: Background

Chapter 2 discusses the various components required to embark on spatial analytics using address data. It discusses data quality and the issues surrounding it and defines geocoding accuracy. This chapter also covers some of the many uses for spatial analytics around the world and looks at current practices and challenges in South Africa.

Chapter 3: Workflow Developed

Chapter 3 discusses the workflow used in order to understand the potential error rates and steps to achieve a higher level of confidence in the final data. It covers workflow components for address elementising, address

cleansing, aggregation, geocoding, comparison of the geocoded results, and the extraction and mapping of acceptable records.

Chapter 4: Workflow application and results

Chapter 4 looks at the end results for each company as well as the results from each of the geocoders.

Chapter 5: Spatial analytics examples

Chapter 5 visually illustrates some insights into the client data of the two companies that became apparent from the maps produced as the final product of the workflow.

Chapter 6: Discussion

Chapter 6 discusses the results, future studies and the implications of using the data after workflow implementation.

Chapter 7: Conclusion

Chapter 7 describes the conclusions of the study

2 Background

2.1 Data quality

According to the Data Warehousing Institute in America, poor data quality costs US businesses six hundred billion dollars annually (Geiger, 2004). Geiger explains that many organizations are in denial about the extent of their data quality problems. He states that at the beginning of a project, many clients insist that they do not have issues with data integrity. By the end of a project, he has never had a client say they have no data quality issues. Often it takes a major set-back to the project or a business catastrophe before they act to correct these issues.

A recent survey estimated that about half of the respondents, mostly IT professionals and IT leaders, found that the quality of their organizations' data should be questioned (Shacklett, 2014).

Data quality issues are often referred to as 'dirty data'. According to Kim *et al.* dirty data can be described as "missing data, not missing but wrong data, and not missing and not wrong, but unusable" (Kim, Choi, Hong, Kim, & Lee, 2003). Data can be seen to be dirty if an application or end user is unable to retrieve a result because of issues with the data or if the result is incorrect.

Examples of dirty data also include misspellings, duplicates, illegal values, contradictory records, abbreviations and incorrect references (Rahm & Do, 2000).

Data governance generally sits with IT because the data is viewed as an IT asset. According to Khatri and Brown (2010), the decisions on how the data is accessed, interpreted, retained, as well as the standards for the data quality play key roles in how IT infrastructure and prioritization are set.

2.2 Improving address data quality

Variations naturally occur in manually captured data, regardless of capturing standards. In environments where there is no standardization, this variance is greater. In one large South African organisation, Coetzee and Cooper (2007a) describe the town of Witbank as being captured in over 200 different ways. In South Africa there are tools designed to assist with standardisation and cleansing for the South African Post Office (SAPO) certification that allows businesses to receive discounts when they achieve an accepted level of postal data quality. The Postal Address Management Service Suppliers (PAMSS) provide checking against South African address data models as a service in order to receive a certificate. The official SAPO standard is SANS-

1883. South Africa is also compliant with the international UPU-S42 standard. (Rossouw, 2009) SANS-1883 defines twelve address types and utilises a data model that includes 60 elements. The standard can be categorised into four address groups:

- Traditional formalised
- Composite
- SA Post Office
- Descriptive

(Coetzee & Cooper, 2007b)

This standard aims to, "enable interoperability in address data, which in turn will facilitate developing a national address database" (Coetzee & Cooper, 2008). At this time, the standard is not supported in software, but aims to establish a common terminology.

2.2.1 Categories of address cleansing

Address cleansing, according to Maletic and Marcus can be categorised in 6 steps:

- Elementising
- Standardising
- Verifying
- Matching
- Householding
- Documenting.

(Maletic & Marcus, 2000)

2.2.1.1 Elementising

Elementising, also sometimes referred to as parsing, is achieved by breaking apart a record into distinct elements. The first stage of address cleansing, it can also be described as separating the addresses into fields, or removing superfluous fields or text, so that there is no ambiguity when it is cleansed. This does not need to be strict, as most address cleansing software can handle common and logical variations.

The address cleansing engine can have difficulty with certain edge cases if a client has not defined their address fields correctly, or has mixed address data with other business fields. These errors may need to be catered for manually. For instance, an address box could have been a free text input where an end user typed the following:

> Joe's Restaurant
> Deliver on Sunday
> Knock 2 times at back door
> On Cincinnati, Pleasantville, 0001

This example becomes very difficult for the software to interpret. It can misunderstand '2 times' for a building or street, while the actual street name 'Cincinnati', could be misinterpreted as a suburb. Some CRM (Customer Relationship Management) systems have the ability to configure and validate for this step. Front-end validation and structured input fields are generally required in order to avoid the necessity for post-capture elementising.

2.2.1.2 Standardising

Standardising removes capture variations and enforces conformity of regularly used address elements. Language variations should also be configured. In South Africa, most address cleansing software supports English and Afrikaans inputs. This step also includes a choice in text case of only uppercase, lowercase or camel case.

Differences in structure such as street, unit, box, suite, or bag number captured in the wrong order are standardised, for example, 'P. O. Box' vs 'PO Box'. These variations are important to standardise as many new regulations require the business to know if the address is a physical address or only a mailing address.

2.2.1.3 Verifying

Verification, where possible, is also handled by the address cleansing engine. Where possible, delivery points, and suburb/postal code combinations can be verified. For the records where the addresses are not verified, a set of error codes are returned.

2.2.1.4 Matching and Householding

Data matching and householding can be handled at point of capture or post capture. Both attempt to eliminate duplicate or similar records. Matching does this by searching for exact or similar child records for the same parent record, such as a contact address captured twice with a slightly different spelling. It can include the automated merging of records, reporting, or a warning on input.

Householding is similar to matching, but instead attempts to identify related parent records by finding matching child records. A number of methods involve using fuzzy logic matches to identify clients who may share similar attributes and location, including address, email and phone numbers. An example is a husband and a wife captured separately who have the same postal address.

2.3 Geocoding Accuracy

A geocode can be described as the coordinates of an address defined by a precise latitude and longitude (Trillium Software, 2009). There are many levels of precision that a business can choose when deciding to

geocode their data. This ranges from regional or country levels to rooftop or exact delivery point. The following levels are described by Thompson (2011), as the Association for Cooperative Operations Research and Development (ACORD) Simple Address Standard.

Address Levels:

- Point
- Building or Landmark
- High Resolution: Parcel
- Street-Segment Imputed
- Medium Resolution: Block
- Street Centroid
- Postal Centroid/Micro Zone
- Administrative Region: 3rd Order
- Administrative Region: 2nd Order
- Administrative Region: 1st Order.

For privacy reasons it is important to consider precision levels used in analysis. A precise address level can pinpoint a single address, which can be linked back to an individual. This is similar to reverse engineering point location data to locate actual residents.

Traditional geocoding uses the full address to geocode a coordinate. This is the most accurate way to geocode an address as erroneous postal codes or administrative regions can usually be ignored by using the street address.

In the example below, both addresses are incorrect; however, the second address will geocode to the street level in most geocoders as the street name matches the postal code. The first address will default to the most precise administrative region, which in this case is Cape Town as both the suburb of Claremont and 7441 are found within Cape Town, but not associated directly to each other.

- Claremont, Cape Town 7441
- 17 Blouberg Road, Claremont, Cape Town 7441

In order to avoid privacy violations, it is becoming more common for businesses to share geographic data that is geocoded to a less precise level of accuracy such as postal code or suburb. An example is the transportation company, Uber. They have recently engaged in a partnership with the city of Boston. Uber is

providing aggregated geographic transportation data identified only by postcode. (Graham, 2015) The city of Boston hopes to use this data to gain insight on travel patterns and travel times.

2.4 Geocoding Methods Explored

Before a geographic information system (GIS) can be used with any data set, its records need to be geocoded, i.e., to be associated with latitude and longitude values. Geocoding information containing address data can either be done using an existing (free or commercial) geocoding service, or by loading a directory associating place names with geocodes (called a gazetteer) into the database and using SQL.

Swift *et al.* (2008, p. 2) reviewed eight common geocoders in their study for the Centres for Disease Control and Prevention (CDC). These can be seen in Table 1 below.

Table 1: Geocoders evaluated by Swift *et al.* (2008)

Name	Application	Commercial/Open Source	Coverage
Centrus US Street Point Data	PC-based	Commercial	United States (Worldwide cities)
ESRI Address Locator	PC-based	Commercial	Worldwide (User Defined)
Geocoder.us	Web-based	Commercial	United States
Geolytics GeocodeDVD	PC-based	Commercial	United States
Google Earth	Web-based	Commercial	Worldwide
Google Maps API	Web-based	Commercial	Worldwide
Yahoo Maps API	Web-based	Commercial	Worldwide
USC Geocoding Platform	PC/Web-based	Open Source	United States

In addition to the geocoders in the Swift *et al.* study, we also evaluated an additional three geocoders. These were selected to be evaluated as they are commonly used geocoders and were within an acceptable price range or free of use. They were Bing Maps API, MapQuest, and gazetteer data dumps. The full list can be seen in Table 2 below, with the geocoders selected for the study in red text.

MapQuest only supports geocoding for the United States, so it was eliminated. Google, which had the most accurate results worldwide, had limitations on how the data can be displayed. According to the Maps API Terms of Service Licence restrictions, the "Geocoding API may only be used in conjunction with a Google map" (Google, 2013). The restriction means that the raw data would not be available for study and import back into the database. As such, Google can be used only as a reference to check other geocoders.

There are many gazetteer data sources available. Most use data primarily from the National Geospatial-Intelligence Agency (NGA) as well as other international sources. We acquired our gazetteer postal code data dumps from Geonames.org. Geonames is considered to be one of the leading geographical databases and is

used by large international corporations such as Ubuntu, Adidas, the New York Times and the BBC. The work is licenced under a Creative Commons Attribution 3.0 Licence and is updated on a daily basis for countries all over the world. The data dumps are free, however, according to Geonames, “is provided "as is" without warranty or any representation of accuracy, timeliness or completeness” (Geonames.org, 2014). A gazetteer, according to Hill and Zheng (1998, p. 1) is defined as, “a list of geographic names, together with their geographic locations.” The gazetteer can also contain other place name information such as political and administrative areas, manmade structures and natural features. The precision varies according to the place name and are defined in the file by the field ‘accuracy’. Field names and formats can be seen in Table 3.

Table 2: Investigated geocoding tools

Geocoding Tool	Address Accuracy (South African)	Conversion Process	Geocode Output	Usability	Cost	Additional Info	Used in Study
Bing Maps API	Street Centroid	Upload CSV to a browser based Geocode Dataflow API	Geographic Coordinate	Simple, but connection often times out	Free: Up to 10,000 Geocodes per 24 hour period. Paid services available	Requires free registered API key	**YES**
Centrus	Not for South Africa	N/A	Geographic Coordinate	N/A	N/A		NO
Esri (World Geocode Service)	Point or Street-Segment	DB connection or Excel/CSV upload to application with connection to central geocode server	Geographic Coordinate	Simple. Both raw and mapped data available	Free: 2500 Geocodes; Entry level cost: R29,000 ~R1 a geocode	Desktop Software requires licence /geocoding subscription	**YES**
Gazetteer	Postal Centroid	Import flat files to a database; Match suburb and town	Geographic Coordinate	Time intensive, often produces duplicates	Free	Not always up to date	**YES**
Geocoder. us	Not for South Africa	N/A	Geographic Coordinate	N/A	N/A		NO
Geolytics GeocodeD VD	Not for South Africa	N/A	Geographic Coordinate	N/A	N/A		NO
Google Earth Pro	Not for South Africa	N/A	Geographic Coordinate	N/A	N/A		NO
Google Maps API	Building or Landmark	Client Side or Server Side browser based JavaScript API	Geographic Coordinate on Google Maps only	Raw data is not available in bulk	Free: 100,000 requests per day	Requires free registered API key.	Only 2nd-ary
MapQuest	Not for South Africa	N/A	Geographic Coordinate	N/A	Free		NO
Yahoo Maps API	Service no longer operating	N/A	N/A	N/A	Free		NO
USC Geocoding Platform	Not for South Africa	N/A	Geographic Coordinate	N/A	Free		NO

For South Africa, admin codes and names were not listed in the postal code file. Updates rely on user feedback. As the technology becomes more widely used in a country, the updates generally become more frequent. The precision levels provided are returned with a numerical value and description of '1=estimated' to '6=centroid'. (Geonames.org, 2014)

Table 3: Data fields from Geonames gazetteer postal code file

Data Field	Format
country code	iso country code, 2 characters
postal code	varchar(20)
place name	varchar(180)
admin name1	1. order subdivision (state) varchar(100)
admin code1	1. order subdivision (state) varchar(20)
admin name2	2. order subdivision (county/province) varchar(100)
admin code2	2. order subdivision (county/province) varchar(20)
admin name3	3. order subdivision (community) varchar(100)
admin code3	3. order subdivision (community) varchar(20)
latitude	estimated latitude (wgs84)
longitude	estimated longitude (wgs84)
accuracy	accuracy of lat/lng from 1=estimated to 6=centroid

Bing Maps is a free API service that has full worldwide capabilities with precision to the street level. It has a limit of 10,000 geocodes a day and requires a user to have an API key in order to geocode. The recommendation for business services that require more than this is to apply for a business subscription. Match codes are returned with the following values; 'Good', 'Ambiguous' and 'UpHierarchy'. Below are the descriptions provided by the Bing Maps Location Resource.

- ***Good**: The location has only one match or all returned matches are considered strong matches. For example, a query for New York returns several Good matches.*
- ***Ambiguous**: The location is one of a set of possible matches. For example, when you query for the street address 128 Main St., the response may return two locations for 128 North Main St. and 128 South Main St. because there is not enough information to determine which option to choose.*
- ***UpHierarchy**: The location represents a move up the geographic hierarchy. This occurs when a match for the location request was not found, so a less precise result is returned. For example, if a match for the requested address cannot be found, then a match code of UpHierarchy with a RoadBlock entity type may be returned.* (Microsoft, 2015)

Esri is one of the leading companies in the world for GIS software products and Esri's World Geocode Service is a popular commercial geocoding service. Esri requires a subscription to do any geocoding, although a 30 day trial account can be created that allows the user to use up to 200 credits towards geocoding. Esri has an interactive credit estimator that will calculate how many credits a user needs to purchase based on their geocoding needs (ArcGIS Online Credits Estimator, 2014). An entry level subscription in South Africa is R29,000 (excl. VAT), which equals 2500 credits. At 80 credits for 1000 geocodes, geocoding comes out to about R1 a geocode.

The supported level for South Africa is level 2, which Esri considers to be a "Good quality geocoding experience." (Esri, 2013) . When one is concerned with suburb precision, level 2 is adequate.

The levels are described by Esri below:

- ***Level 1*** *(darkest-shaded countries in map): Highest quality geocoding experience. Address searches are likely to result in accurate matches to PointAddress and StreetAddress levels.*
- ***Level 2*** *(medium-shaded countries in map): Good quality geocoding experience. Address searches often result in PointAddress and StreetAddress level matches, but sometimes match to StreetName and Admin levels*
- ***Level 3*** *(lightest-shaded countries in map): Fair quality geocoding experience. Address searches sometimes result in street-level matches, but more often match to StreetName and Admin levels.*
- ***Level 4*** *(unshaded countries in map): Street-level geocoding is not supported at all, and address searches will return no match. Only admin place geocoding is available* (Esri, 2013).

Admin levels or admin places can be described as a place name such as a locality, municipality, city, or even state or province in that order of precision from highest to lowest. The administration levels can also be mixed with postal boundaries to create a hybrid 'PostalLoc' or postal locality.

Figure 1 below shows that although South Africa does not have level one geocoding quality, it is significantly better than the surrounding African countries that are mostly at level three or four.

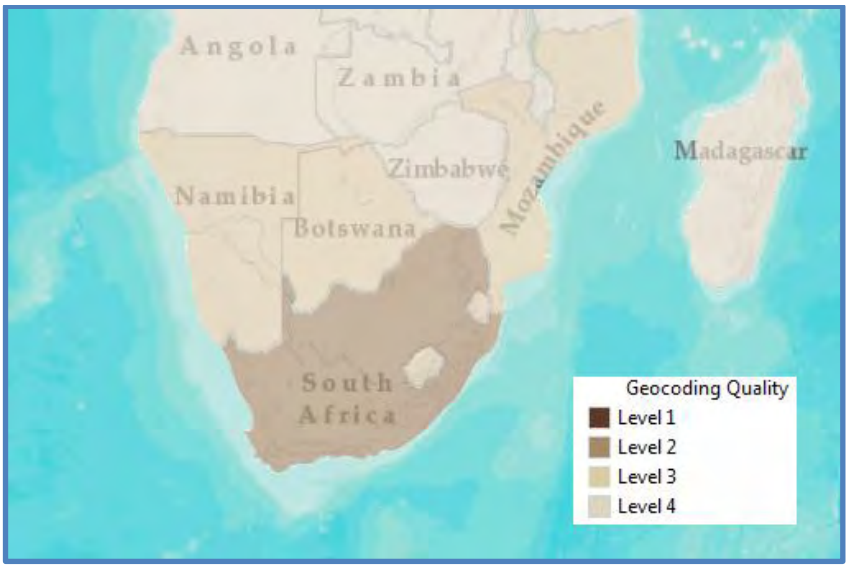


Figure 1: ArcGIS Geocoding quality for South Africa

2.4.1 Verifying geocoded results

A technique known as "interactive geocoding" (Goldberg, Wilson, Knoblock, Ritz, & Cockburn, 2008) uses multiple geocoding APIs or services to achieve the best result. They describe using a free or readily available method for the first pass of geocoding. If a geocoder is unable to match any results, it will either return an error or a blank row. For any inaccurate or missing results, a smaller subset can then be sent for processing via a commercial method or manually retrieved. This technique can help to defer initial costs of a large GIS project.

2.5 GIS and Spatial Analytics

2.5.1.1 What is GIS?

A GIS is described as "A collection of hardware, software, and data and liveware which operates in an institutional context" (Maguire, 1991, p. 17). As this definition could describe any number of technical systems, Maguire goes on to describe it as having three main views: the map, the database, and spatial analysis. The main focus is on spatial entities and relationships in each view. Essentially, a GIS has the ability to organise data sets by geography.

Spatial analysis can either be tightly or loosely coupled with GIS systems (Hongjian, Fei, & Chunxiu, 2011). Hongjian *et al.* describe loose coupling as lower risk and easier to realize. However, it is not seamlessly

integrated. The systems stand alone and data is either extracted manually from one system to the other or there is a data transformation interface between the spatial data base and the data documentations.

The other model they describe is a tight coupling of spatial analysis with the GIS system. In this model there is a seamless connection, with one system being dominant and the other acting as a layer added in. In this model, the advantages are graphic demonstration and spatial analysis computation.

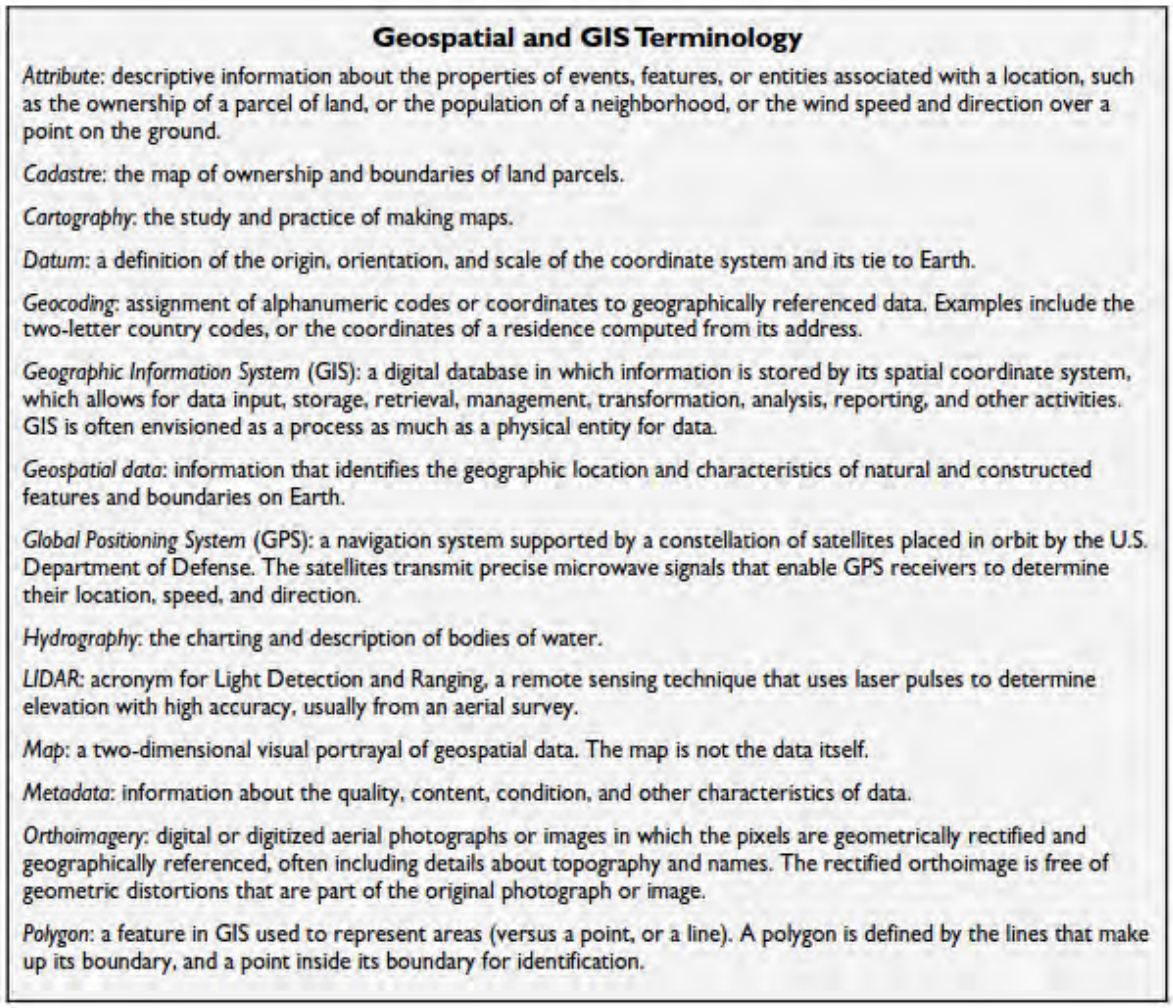
GIS uses a set terminology across the three different functions of the map, the database and spatial analysis. These can be seen in Figure 2 below from Folger (2009). Regardless of the industry, the terminology remains the same which makes a GIS adaptable to a multitude of industries around the world.

2.6 The Importance of Place

Is place important in analytics? It has been said by many in marketing that the three secrets to success are 'location, location, and location' (Jones & Simmons, 1987). At first, this may not seem to apply to businesses that do not rely on direct face to face contact with their clients. However, understanding the demographics of one's client base can be vitally important to the decisions made for the business.

Understanding place allows us to ask the question; 'Where?' It allows us to visually see a pattern across a location or multiple locations as well as do comparisons in areas that we might not have seen before. Location enables us to see trends in surrounding areas and find resemblances in variables across population groups that may not have otherwise been grouped similarly (Gregory, 2005).

According to Lombaard (2010), thinking spatially is rarely considered in most organisations when making decisions about organisational structure and customer service. She continues by explaining that it is even rarer for organisations to consider using the location of their existing client base to assist in the acquisition of new clients.



Geospatial and GIS Terminology

Attribute: descriptive information about the properties of events, features, or entities associated with a location, such as the ownership of a parcel of land, or the population of a neighborhood, or the wind speed and direction over a point on the ground.

Cadastre: the map of ownership and boundaries of land parcels.

Cartography: the study and practice of making maps.

Datum: a definition of the origin, orientation, and scale of the coordinate system and its tie to Earth.

Geocoding: assignment of alphanumeric codes or coordinates to geographically referenced data. Examples include the two-letter country codes, or the coordinates of a residence computed from its address.

Geographic Information System (GIS): a digital database in which information is stored by its spatial coordinate system, which allows for data input, storage, retrieval, management, transformation, analysis, reporting, and other activities. GIS is often envisioned as a process as much as a physical entity for data.

Geospatial data: information that identifies the geographic location and characteristics of natural and constructed features and boundaries on Earth.

Global Positioning System (GPS): a navigation system supported by a constellation of satellites placed in orbit by the U.S. Department of Defense. The satellites transmit precise microwave signals that enable GPS receivers to determine their location, speed, and direction.

Hydrography: the charting and description of bodies of water.

LIDAR: acronym for Light Detection and Ranging, a remote sensing technique that uses laser pulses to determine elevation with high accuracy, usually from an aerial survey.

Map: a two-dimensional visual portrayal of geospatial data. The map is not the data itself.

Metadata: information about the quality, content, condition, and other characteristics of data.

Orthoimagery: digital or digitized aerial photographs or images in which the pixels are geometrically rectified and geographically referenced, often including details about topography and names. The rectified orthoimage is free of geometric distortions that are part of the original photograph or image.

Polygon: a feature in GIS used to represent areas (versus a point, or a line). A polygon is defined by the lines that make up its boundary, and a point inside its boundary for identification.

Figure 2: GIS terminology (Folger, 2009)

2.6.1 Definition of place

There are many different sizes of 'place' and many ways in which it can be defined. Some definitions are political, such as countries or city boundaries, and others are more personal to the person describing it. The difference in the idea of place can also vary widely by profession. A geographer will generally associate more with the size of a city, town or settlement, while an architect will generally think on the scale of a building or group of buildings (Tuan, 1979).

The implications of adding geographic location can change the way we look at data. Adding a geographic coordinate to standard analysis techniques, allows us to take a large data set and reduce it to a smaller more meaningful sub-area. We can visually check our assumptions, discover outliers and vary the models in

multiple relationships types (Fotheringham, Brunsdon, & Charlton, 2000). A visual model allows us to observe patterns that would have otherwise seemed random. It allows us to test and build upon these patterns to aid in our predictions.

Visual geographical analysis can be added to environments that aren't necessarily geographically. Neves and Camara (1999) describe this level of interaction as a 'virtual environment'. It is defined as the human cognitive level of mental maps in combination with audio and visual images in a computer. It allows the user to visualize the data while analysing it on multiple planes or layers. This sort of abstraction is called 'spatializiation' (Goodchild & Janelle, 2010, p. 9). It is described as, "The construction of abstract spaces of knowledge that can aid in visualization, pattern detection, and the accumulation of scientific insight". This allows non-traditionally geographic topics to also be studied spatially with graphical representation.

In the past, it was argued that GIS would not be critically needed because distance would no longer be an issue in the digital world (Cairncross, 1997). This argument can be contradicted by distribution issues alone; there is an essential need to know where clients are in order to deliver to them (Grimshaw, 2001). Once client location is known, geographical knowledge can then be used to forward one's business strategies.

2.7 The Various Uses for GIS

There are many industries around the world that currently utilize spatial analytics and GIS. The following examples are listed and discussed below, although this list is far from exhaustive.

- Agriculture
- Banking
- CRM Systems
- Civil Government
    - Utilities
    - Public Safety
    - Disaster Management
- Healthcare
- Insurance
- Postal and Delivery Services
- Retail Marketing.

2.7.1 Agriculture

GIS has become essential for analysis on crop dynamics, crop hazards, crop insurance, and knowledge sharing. At the United States Department of Agriculture (USDA) numerous GIS applications have been developed across the seventeen agencies of the department (Marshall, 2013). In the event of a flood, for

instance, GIS is used to plan municipal impact, to estimate damage, to negate the threat of insurance fraud and in planning future uses for the land affected. It also acts as a knowledge base for crop estimates, previous harvests, ground conditions, conservation efforts and farmer surveys amongst other uses. The USDA has a national resource website by the name of CropScape that hosts the Cropland Data Layer (CDL). "The CDL is a raster, geo-referenced, crop-specific land cover data layer created annually for the continental United States using moderate resolution satellite imagery and extensive agricultural ground truth" (U.S. Department of Agriculture, 2014).

The USDA is not the only the only organisation employing GIS for agriculture, it is quickly becoming a standard worldwide. Johnson and Yespolov (2014) are of the belief that "GIS technology has the potential to revolutionize global agriculture". They argue that precision farming based on the use of new modelling systems has the ability to predict more accurate yields for regions employing the technology. It also allows organisations to anticipate shortages and take subsequent action against it.

2.7.2 Banking and finance

A number of uses have been identified in the non-traditional GIS sector of banking. The most frequent uses of GIS are for banking and ATM locations as well as in risk management. However, there is also a new trend to use GIS for marketing purposes in the financial industry. By using demographic information and mapping customer behaviour, companies can develop business strategies and calculate risk when marketing to new user groups (Prasad & Ramakrishna, 2011).

Lifestyle/life-mode groups are a new way in which GIS is affecting analysis in the financial industry (Parish, 2009). Parish looks at these groups to show new ways to penetrate the market and to determine site selection. This type of analysis is becoming increasingly popular and is dictated by marketing segmentations or 'tapestry segments'. These are more thoroughly examined in section 2.8.

Two interesting investment observations that were made possible by the use of GIS are what Brown, Ivkovich, Smith & Weisbenner (2004) describe as the 'community effect' and the 'local firm effect'. In the community effect, the premise is that stock market investors in communities are more likely to invest if other investors exist in households within 50 miles of them. If the household is 'less financially sophisticated' the effect has been reported to be stronger.

The second factor described by Brown *et al*. is the 'local firm effect'. In this factor, the actual presence of the firm in the local community increases the investor's desire to purchase stocks from that firm. They have

surmised that this could be due to a number of reasons including word of mouth, social interaction or 'observational learning'. They also surmise that 'familiarity' could have a strong influence.

2.7.3 CRM systems

Most companies employ some version of a Customer Relationship Management (CRM) system. These systems play a vital role with regard to organising and tracking client data. The use of GIS can expand relationships within the CRM system by using existing data to make visual and spatial connections that might have otherwise been missed. This creates the ability to do spatial analysis on the client, client groups or the full client database.

2.7.4 Government

Governments are perhaps the largest consumer of GIS systems and have typically been seen as the early adopters of the technology. As such, the applications in government are numerous and contain sophisticated solutions.

The US government is one of the largest consumers of geospatial information, so much so that the cost of it has become a fairly large concern. According to Folger (2009), the release of Google Earth in 2005 shifted the perception of how people use and understand geospatial information. He states that the percentage of geospatial components in US government information has been said to be as much as 80% to 90%. It has offered users a free and easy to use multi-scale visualization of places all over the globe.

The US has gone to the extent of creating a committee that focuses entirely on spatial data. The Federal Geographic Data Committee (FGDC) is a US based interagency government committee that, "promotes the coordinated development, use, sharing, and dissemination of geospatial data on a national basis" (Federal Geographic Data Committee, 2013).

In an effort to enable sharing and development of spatial data, the FGDC created the National Spatial Data Infrastructure (NSDI) The purpose of the NSDI is as a, "physical, organizational, and virtual network designed to enable the development and sharing of this nation's digital geographic information resources" (Federal Geographic Data Committee, 2013).

This virtual network can be accessed via the Geospatial One-Stop portal. Folger describes this portal as, "… the official means of accessing metadata resources, which are published through the National Spatial Data Clearinghouse and which are managed in NSDI" (Folger, 2009).

In 2010, the US president in his 2011 budget speech announced a roadmap to develop what is now known as the 'Geospatial Platform' (FDGC, 2011). As a result of the budget speech, it was stated, "The Geospatial Platform will explore opportunities for increased collaboration with Data.gov, with an emphasis on reuse of architectural standards and technology, ultimately increasing access to geospatial data" (US Office of Management & Budget, 2010).

2.7.4.1 Utilities and services

Cape Town endeavoured to integrate enterprise resource planning (ERP) systems in 2001 and did so by creating "one of the largest ERP systems ever implemented by a local government" (Esri, 2007/2008). They implemented an ArcGIS enterprise platform that integrated and consolidated utility and property databases into the GIS. An example of the implementation can be seen in the city's water services. The GIS allows a spatial data model for tracking issues such as outages, burst pipes or meters as well as allowing the ability to track consumption and change tariff structures for high usage areas. It also allows for an alarm system to be integrated into the infrastructure to facilitate faster responses to any problem areas.

2.7.4.2 Public safety

Policing and crime analysis have also benefited from GIS systems in recent years. Because most crimes are associated with a place, it lends itself well to the technology. Crime patterns and crime indicators can be studied and used to predict future crime as well as prevent it by increasing the presence of law enforcement in high risk areas (Ferreira, Joao, & Martins, 2012). Ferreira *et al.* describe the usage of hot spot analysis as being the most common method of analysis in crime detection. This analysis has shown that initial spots that are considered to be 'moderate' can show increases in crime over time, with the severity of the crime often worsening.

Geographic profiling can also be used to determine possible crime areas. In South Africa, census data was used to define 20 categories based on social-economic factors including over "250 census variable and 74 crime variables" (Breetzke, 2006). The categories were then used to prioritize intervention in each police station. Aside from creating awareness for the individual stations, it also provides insight into the factors that cause these crimes.

2.7.4.3 Disaster management

GIS systems can play a critical role in disaster management. Johnson (2000), describes a GIS as having the ability to centrally house and display essential information in the event of an emergency. He divides emergency management into 5 phases: Planning, mitigation, preparedness, response and recovery. The key

to successful mitigation of risk in these disasters is the ability to model potential disasters. The models can then be used for training and preparation or for actual resource organisation should a disaster occur.

There have been a few major disasters that have recently used the final two phases for response and recovery, although reactively. The earthquakes that hit Haiti in 2010 and Japan in 2011, are remarkable examples showing how the combination of GIS and volunteered information from ground efforts can help the disaster relief effort (Ortmann, Limbu, Wang, & Kauppinen, 2011). In combination with Twitter, shortages of food, water and medication could be reported and mapped. Makeshift hospitals or treatment areas could also be mapped and communicated to those in need. The relief efforts made use of Google and OpenStreetMap in conjunction with Ushahidi's Crowdmap.

2.7.5 Healthcare

Epidemiology and geography have partnered in healthcare since 1854 when John Snow tracked a cholera outbreak in London to a water pump. The case, now known as the 'Broad Street' epidemic of 1854, found that 700 deaths had occurred in a 250 yard radius (Newsom, 2006). John Snow, a local doctor in the area who had heard of the epidemic, mapped the deaths surrounding the water pump and established a correlation to the water pump. He then had it removed and the outbreak subsided. He is now considered to be the father of geographical epidemiology.

Since then GIS has been widely used to track disease outbreaks in healthcare. Cases similar to John Snow's cholera outbreak can now be tracked in multiple overlay layers using not only location, but also by identifying elevation, vegetation, water patterns, and other factors such as mosquito control. Preventative measures can be taken in areas mapped with high risk variables (Clarke, 1996). Clarke also describes how effective these visualizations can be when mapped over time. When the AIDS epidemic hit the United States, a series of animated maps were able to show the movement of the virus through major cities.

Mobile technologies in conjunction with GIS have added to these efforts. On the Thai-Cambodian border mobile devices were used in an effort to follow up and track cases of malaria. These individual cases were then mapped and spatial analysis was used for preventative measures, containment and to allocate resources (Meankaew, Kaewkungwal, Khamsiriwatchara, Khunthong, Singhasivanon, & Satimai, 2010).

Search engines and trending key word query data have enabled early detection of epidemics and diseases by mapping them to the locations where they originate. Google employees in collaboration with the CDC were able to track influenza epidemics based on 'influenza-like illness' searches (Ginsberg, Mohebbi, Patel,

Brammer, Smolinski, & Brilliant, 2009). This project is part of Google's 'Predict and Prevent Initiative' which includes grants to organisations like HealthMap. HealthMap, which is a project out of the Boston Children's Hospital, is a, "multistream real-time surveillance platform that continually aggregates reports on new and ongoing infectious disease outbreaks" (Brownstein, Freifeld, Reis, & Mandl, 2008).

HealthMap has most recently been in the media for its timeline following the outbreak of the Ebola virus in West Africa. The founders detected what was then known as a 'mystery haemorrhagic fever' over a week before it was confirmed as Ebola (Gilpin, 2014). The timeline can be seen on their website and includes findings from March 14th, 2014. The Ebola outbreak was only confirmed by the World Health Organisation on March 22nd (HealthMap, 2014).

2.7.6 Insurance

GIS has become a major contributor in the insurance industry in both assessments and claims. In risk assessment, a GIS can be a key tool for identifying high risk zones, areas of loss potential, mapping historical claims, and determining loss and backup plans. In assessments for coverage for natural disasters, Prasad and Ramakrishna (2011) describe how a GIS can effectively map areas for frequencies of events that are not regular occurrences. Historical earthquakes, floods, hurricanes, fires, etc. can be mapped and the surrounding areas can be marked as high risk and charged for appropriately.

Areas with high associations for crime or for previous crime related claims can also be mapped and used to determine high risk areas.

GIS can be used to make solid investment and risk decisions when looking at large scale changes over geographical areas. For instance, Dubai is a very good example of rapid growth and can be shown by comparing development at different moments in time (Nassar, Blackburn, & Whyatt, 2012). It can also be used to look at environmental changes because of rapid growth.

Cluster heat maps are another way that insurance companies are utilizing big data. Sampath (2012) describes how heat maps can be used to analyse sensor information from automobile tracking systems. These systems track braking information as well as violations for 'good driver' benefits. A premium can be determined by comparing the mean number of violations in the area to the violations the driver has accrued. Areas with a high concentration of violations will be clustered and colour coded by intensity.

2.8 Geo Demographics/Tapestry Segments

Segmentation is the ability to group markets geographically in order to target for marketing. It is a practice that has been in effect for over 30 years. It has the ability to describe lifestyle, diversity and economic variables for different areas. The main theory behind studying this information is generally to understand consumer markets (Esri, 2012). Esri is a corporation based out of Redlands, California. They are considered to be leaders in the GIS marketplace and have over 30 years of experience in market intelligence. (Esri, 2014)

Location-based marketing has been increasing in the past year in both small and medium enterprises (SMEs) and in the corporate environment. Etienne Louw, chief executive officer of mapIT, in a recent interview with Insurance Chat, stated that, "Digital mapping is proving to be the hidden secret weapon of South African business." His research has shown that 76% of South African corporations and 38% of SMEs are currently using mapping services. Of that percentage, 35% of SMEs and 41% of corporations are using these services for location based marketing (Jonckie, 2013).

In the United States, Esri has developed what they call 'Tapestry Segments'. The most recent segments are the culmination of a fourth generation effort to classify and distinguish market groups based on government census data and other annual surveys such as the American Community Survey. These are compiled together with list-based data and a ZIP+4 postal code designation grouped together by addresses. Some of the list-based data sources include purchase information, public real estate records, surveys, publications, directories and registrations (Esri, 2012).

Esri's Tapestry Segments are grouped into eight areas of information:

- Population by Age and Sex
- Household Composition and Marital and Living Arrangements
- Patterns of Migration, Mobility, and Commutation
- General Characteristics of Housing
- Economic Characteristics of Housing
- Educational Enrolment and Attainment
- Employment, Occupations, and Industrial Classifications
- Household, Family and Personal Incomes.

  (Esri, 2012)

These tapestry segments are so precise that different streets in the same suburb can have different profiles. This is beneficial when setting up retail locations, a supermarket, banking branches or service centres, or maybe a clothing shop or mall. There are 65 different segments classified in the US by Esri (2012). These segments are broken into summary groups. There are 12 LifeMode Summary groups and 11 Urbanization Summary groups. Two examples of how this is classified can be seen in Table 4.

Table 4: Example Tapestry Segment classification (Esri, 2012)

Segment Group	Lifemode Group	Urbanization Group
08 Laptops and Lattes	L4 Solo Acts	U1 Principle Urban Centers
57 Simple Living	L5 Senior Styles	U6 Urban Outskirts 11

There are currently no Tapestry Segments for South Africa although SAARF publishes the Living Standards Measure (LSM®) which is similar. It contains ten segments with high, medium and low classifications (SAARF, 2012). In order to move away from the classifications of the past a set of variable was chosen to formulate the questions in the survey for classification. These variables are amended year on year as technology and tastes change. For instance, in 2001, A VCR player was retained on the list and a sewing machine was added. The sewing machine was dropped later in 2008 and a home theatre system was added. In 2011, the VCR player was dropped. (SAARF, 2012). The variables used in South Africa for 2011 can be found in Appendix 0.

A big difference between the overseas marketing classifications and those in South Africa is the basis on a fixed geographical area. These segments can generally be classified to suburb. Many marketing companies base their business off these suburb classifications. In the US, however, geographical segmentation is frequently based on postal code. According to Lombaard (2010), "In South Africa postcode geography has never been used as an output form for demographic or census data".

2.9 South African Postcode Geography

South Africa's postal code system was launched on the 8th of October, 1973 (Rossow, 2008). It is a 4 digit code that relates to the region, specific to post office distribution, postal sorting centre (HUB), and to some extent, province (Lombaard, 2010). It is ordered numerically according to distribution area and as such, the sorting ranges do not directly relate to province. These can be seen in Table 5.

Table 5: South African post office sorting lines (Lombaard, 2010)

Province	HUB	Range	
		Start	End
Gauteng	Pretoria 1	0001	0204
Mpumalanga	Pretoria 2	0205	0698
Limpopo	Polokwane (Pietersburg)	0699	0999
Mpumalanga	Pretoria 3	1000	1199
	Nelspruit	1200	1399
Gauteng	Germiston	1400	1699
	Heidelberg	1438	1444
	Krugersdorp	1700	1799
	KDP/Soweto	1800	1870
	Vanderbijlpark	1871	1990
	Witspos (Johannesburg)	2000	2199
Mpumalanga	Pretoria 4	2200	2494
North West	Krugersdorp	2495	2519
	Potchefstroom	2520	2709
	Mafikeng	2710	2899
KwaZulu Natal	Ladysmith	2900	3199
	Pietermaritzburg	3200	3309
	Ladysmith	3310	3599
	Durmail 2	3600	3799
	Richards Bay	3800	3990
	Durmail 1	3991	4179
	Port Shepstone	4180	4299
	Durmail 2	4300	4641
	Port Shepstone	4642	4730
Eastern Cape	Umtata	4735	4739
KwaZulu Natal	Port Shepstone	4740	4799
Eastern Cape	Umtata	4800	4899
	East London	4920	5049
	Umtata	5050	5199
	East London	5200	5750
	Port Elizabeth	5751	6499
Western Cape	George	6500	6699
	Worcester	6700	6899
	Beaufort West	6900	7099
	Cape Mail	7100	8179
Northern Cape	Upington	8180	8299
	Kimberley	8300	8799
	Upington	8800	8999
Free State	Bloemfontein 1	9300	9409
	Welkom	9410	9699
	Bloemfontein 2	9700	9999

Often common names are duplicated across provinces. In the example in Table 6, the suburb of 'Mountain View' is listed in eight ways, including one partial match. The name is found in six different postal HUBs.

Table 6: Suburb name with multiple postcodes and in different provinces

SUBURB	PCODE	AREA	HUB
MOUNTAIN VIEW	0082	HERCULES	PRETORIA 1
MOUNTAIN VIEW	1055	MIDDELBURG	PRETORIA 3
MOUNTAIN VIEW	2192	JOHANNESBURG	WITSPOS
MOUNTAIN VIEW	2470	VOLKSRUST	PRETORIA 4
MOUNTAIN VIEW	5880	CRADOCK	PORT ELIZABETH
MOUNTAIN VIEW	6229	UITENHAGE	PORT ELIZABETH
MOUNTAIN VIEW	7646	PAARL	CAPE MAIL
MOUNTAIN VIEW VILLAGE	7945	RETREAT	CAPE MAIL

3 Workflow Design

This chapter describes a workflow for geocoding data containing South African addresses. The process followed in order to reach an acceptable confidence level and understand the potential error rates in ones data, can be broken down into six main steps. Below in Figure 3, is a high level overview of the full workflow diagrammed in Figure 6.

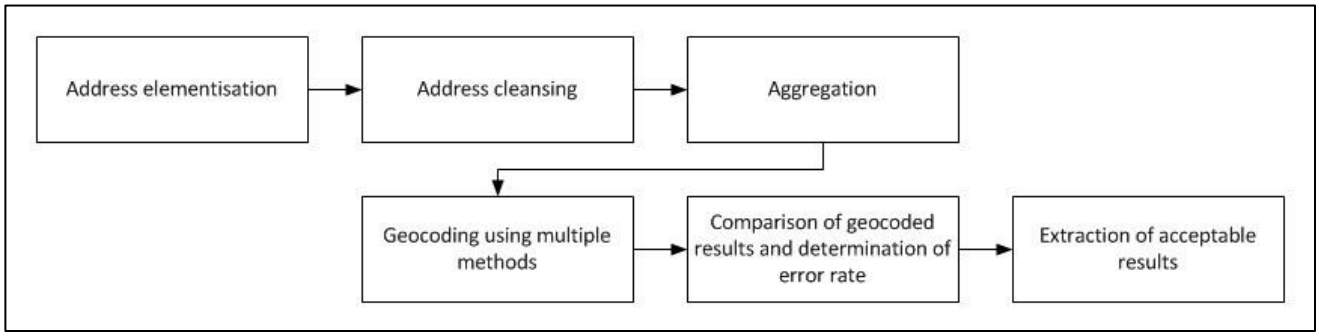


Figure 3: Workflow for geocoding and determining an error rate

Once these have been completed, mapping and analytics can begin so as to reap the benefits of GIS which have been illustrated in Chapter 5. The workflow is a semi-automatic process; hence each step not only produces a revised data set but also creates categorised and colour-coded result reports for the user, who can then manually correct any errors if desired.

3.1 Workflow prerequisites

A number of steps were executed as precursor to the final workflow above. We began by researching the tools and level of maturity in South Africa. Prior to signing up for a commercial service, we tested the free public tools on the data from the financial company. Goldberg *et al.* suggest augmenting the results from free geocoders with commercial geocoding results for any questionable or erroneous records. A subscription to Esri's World Geocoding Service was then made. All three geocoders were used to compare the results and determine what level of error was acceptable to the businesses.

We then re-evaluated the workflow as a semi-automated solution in order to allow for user intervention to reduce the error rate. We built a reporting solution with a colour categorisation system for invalid results.

The workflow was applied to the two data sets for the financial company and then later to the two data sets for the publishing company. Using this method allowed us to test that the workflow worked for differing data sets and different industries.

The data was mapped for both organisations and combined with business variables to determine if both companies gained useful insights from mapping their data geographically.

3.2 Address Elementising

Address elementising, is the first step in the workflow and in certain instances can be handled by sophisticated address cleansing software. However, depending on the client database quality, a portion of the elementising effort often needs to be done manually. The elementising component in the workflow assists by automatically identifying commonly occurring address substrings and patterns (such as 4-digit postal codes) using fuzzy logic and wildcard searches. All addresses are converted to uppercase before uploading although the address cleansing software does have the capability of converting to camel case. Basic checks are run in SQL to ensure the data had been categorised correctly. For instance, wildcard operator searches are performed in the suburb and city fields to check for street addresses or postal box data. An example of some of the wildcard combinations can be seen in Table 7 below.

Table 7: Example wildcard searches

Example Wildcard Searches	
%P O BOX%'	%STREET%
%PRIVATE BAG%'	%DRIVE%
%POSTNET%'	%LANE%
%PO BOX%'	%CIRCLE%
%BOX%'	%CRESCENT%
%BAG%'	%ST%
%POSBUS%'	%DR%
	%LN%
	%CR%
	%CRES%

3.3 Address Cleansing

In order to compensate for other capture errors, address cleansing software is used. The application runs on a Glassfish webserver and uses windows services to run the address cleaning engine.

Depending on the data set size and time available, the returned addresses can either be kept as-is or suggestions from the address cleansing software can be verified manually.

The address cleansing engine is able to standardize case, spelling, language, and match street/box, suburb and postal codes within South Africa. It returns both the old and the amended addresses with codes of what changes are made.

Client data is extracted from the client data system and interfaced via a SOAP API. Any web service automation tool, such as SoapUI, can be used or a custom interface can be written. We automated the web service with the option of either writing the results directly to the database or as an output to Excel. This process is illustrated in Figure 4 below.

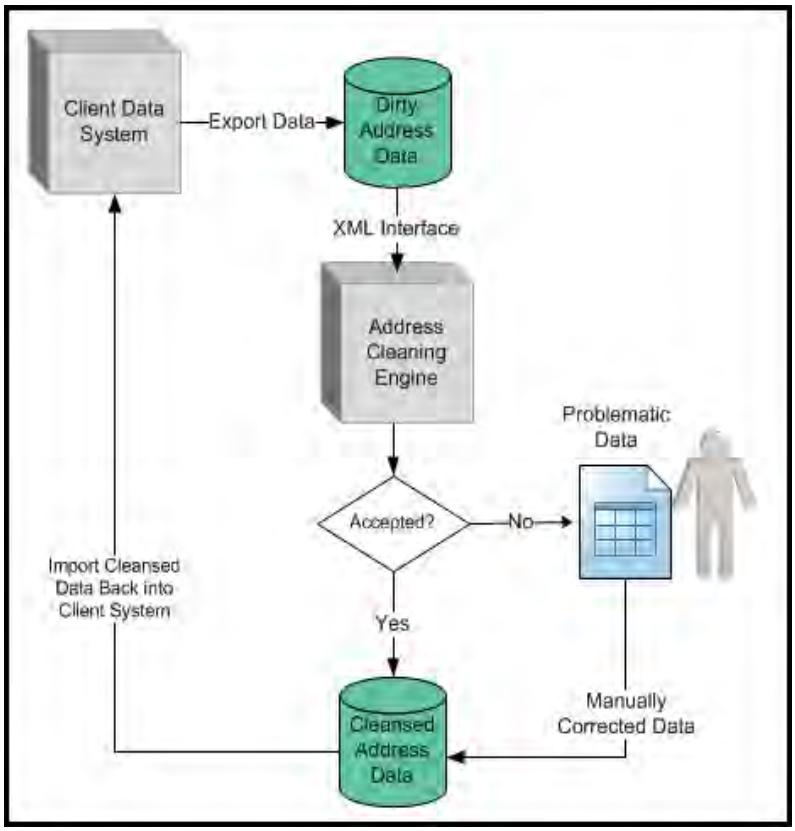


Figure 4: Address cleansing process

The complete address is used when cleansing. This assists with verification where more than one match is possible. If cleansing remains in-house, it is possible to use the full address without risking privacy concerns. In instances where the data is from a third party, or there are privacy concerns, then only the suburb, town and postal code can be used.

3.3.1 Interpreting address cleansing codes

About one hundred variations of change codes can be returned by the address cleansing software. A stored procedure using SQL categorizes these return codes more broadly, to facilitate subsequent human intervention. Returned codes as well as addresses before and after cleansing are inserted in a separate relation for analysis. Codes are grouped into seven levels (0-6) and three basic colour bands (green, yellow and red). This can be seen below in Table 8. All yellow and red address groups need major changes or are not found in the postal data tables.

Table 8: Acceptance levels for address cleansing

Level	Status Colour	Reason Message
1		Accepted Address
2		Accepted Address, Changes Made
3		Address Corrected, Confirm Accuracy
4		Foreign Address, Confirm Accuracy
5		Unauthorised Address, Revalidate or Confirm
6		Rejected address, Keeping Original
0		Invalid Request Structure

A sample of the SQL for categorizing the results for an accepted level can be seen below.

```
UPDATE @ResultSummary
SET    CodeLevel = '2_GREEN',
       ReasonMsg = 'Accepted Address, Changes Made'
WHERE  DelStatCode IN ('B')
       OR (DelStatCode IN ('C')
              AND (ChangeCode IN ('A01', 'B01', 'B03')
                     AND ValidityCode IN ('V04')))
```

The codes are returned with delivery status, validity, and details of the change. The delivery status is according to official PAMSS postal tables. While these would be important in a full cleansing exercise, for our purposes, the validity and details of change are more important. For instance, a validity status of V14 indicates an error where there are multiple unrelated suburbs, an example being Table View, Johannesburg or Claremont, Milnerton. This would require manual intervention if the street address and postal code were also ambiguous.

The change code indicates what was corrected on the record if a change was proposed. 'B03', for example indicates that a misspelling was corrected, such as 'CENTURIAN' to 'CENTURION'. These types of changes were accepted and marked with a code level of either '1_GREEN' or '2_GREEN'. Examples of the corrected addresses and code levels are listed in Table 9.

Table 9: Examples of old and new addresses after cleansing as well as associated levels

Code Level	Types of Changes	Reason Message		Address Line 1	Address Line 2	Address Line 3	Suburb	City	Postcode
1_GREEN	Standardize case, suite, bag and box format, correct spelling, language	Accepted Address	***OLD***	*Suite 123*	*Private Bag X 123*		*MTHATHA*		*5099*
			NEW	POSTNET SUITE 123	PRIVATE BAG X123		UMTATA	UMTATA	5099
1_GREEN	Standardize case, suite, bag and box format, correct spelling, language	Accepted Address	***OLD***	*P O Box 123*			*Derdepoortpark*		*0035*
			NEW	PO BOX 123			DERDEPARK	PRETORIA	0035
2_GREEN	Standardize case, suite, bag and box format, correct spelling, language and postal code	Accepted Address, Changes Made	***OLD***	*123 Somewhere Drive*	*Becon Bay*		*EAST LONDON*		*5241*
			NEW	123 SOMEWHERE DRIVE	BECON BAY		EAST LONDON	EAST LONDON	5201
3_YELLOW	Standardize case, suite, bag and box format, correct spelling, language and postal code and suburb. Needs to be manually confirmed	Address Corrected, Confirm Accuracy	***OLD***	*123 SOMEWHERE PLACE*	*SOMEWHERE CLOSE*	*MORNINGSIDE*	*SANDTON*		*2165*
			NEW	*123 SOMEWHERE PLACE*	*SOMEWHERE CLOSE*		*MORNINGSIDE*	*JOHANNESBURG*	2196
4_YELLOW	Foreign Address; Standardize case, suite, bag and box format, correct spelling. Needs to be manually confirmed	Foreign Address, Confirm Accuracy	***OLD***	P O BOX 123	TANGA		TANGA		*2165*
			NEW	PO BOX 123	UNITED REPUBLIC OF TANZANIA		TANGA	TANGA	
5_RED	Address need to be checked. Street/Box, postal code and suburb do not match.	Unauthorised Address, Revalidate or Confirm	***OLD***	*123 SOMEWHERE*	*STELLENBERG*		*BELLVILLE*		7530
			NEW				STELLENBERG	DURBANVILLE	7550
6_REJECT	Reject address to original. No match, returns error. Generally foreign.	Address Rejected, Keeping Original	***OLD***	*123 SOMEWHERE AVENUE APT 123*		*USA*	*DALLAS*		*75219*
			NEW	123 SOMEWHERE	UNITED STATES OF AMERICA STREET		DALLAS		

3.3.2 SSRS reports and originating table updates

A SQL Server Report Services (SSRS) report is generated using a stored procedure to translate return codes into business friendly output levels. The report can be automated and emailed to users should the process be scheduled nightly or weekly. It can also be run on an ad-hoc basis. A snippet of the code for the reporting can be seen below.

```
CASE WHEN CodeLevel IN ('3_YELLOW', '4_YELLOW', '5_RED',
                                '6_REJECT', '0')
     THEN ISNULL(OldAddrLine1, '')
     ELSE AddressLine1
 END AS AddressLine1,
 CASE WHEN CodeLevel IN ('3_YELLOW', '4_YELLOW', '5_RED',
                                '6_REJECT', '0')
     THEN ISNULL(OldAddrLine2, '')
     ELSE AddressLine2
 END AS AddressLine2,
 CASE WHEN CodeLevel IN ('3_YELLOW', '4_YELLOW', '5_RED',
                                '6_REJECT', '0')
     THEN ISNULL(OldAddrLine3, '')
     ELSE AddressLine3
 END AS Suburb,
 CASE WHEN CodeLevel IN ('3_YELLOW', '4_YELLOW', '5_RED',
                                '6_REJECT', '0')
     THEN ISNULL(OldCity, '')
     ELSE City
 END AS City,
 CASE WHEN CodeLevel IN ('3_YELLOW', '4_YELLOW', '5_RED',
                                '6_REJECT', '0')
     THEN ISNULL(OldPostcode, '')
     ELSE Postcode
 END AS Postcode
```

The stored procedure also updates the new relation to the business defined level of acceptability. Updates to the original address are made to a new cleansed address master, while retaining the old address with the return codes that were not acceptable.

For our purposes, we did not go to the level of matching and householding although future development for this may be possible.

Once the addresses have been standardized and cleansed, the process of geocoding can begin. The cleansed addresses are re-integrated back into the database and the new cleansed address database is used exclusively going forward.

3.4 Aggregation

In order to save on geocoding costs financially, and for efficiency reasons, only unique combinations of suburb, town, and postcode are geocoded. This is accomplished by aggregating on suburb, town and postcode after the addresses have been cleansed and standardized.

This aggregation is also done in SQL. Distinct combinations of suburb, town, and postcode are returned with a count of the number of records in each.

3.5 Geocoding

To reduce the amount of geocoding required for the sake of affordability and efficiency, and to retain client anonymity, we use a combination of postal centroids and administrative regions. Because hundreds or thousands of individuals can live within a postal code, it allows one to study client groupings without disclosing exact details of the client. This is in line with the 2013 POPI legislation that dictates personal information is excluded from the Act as long as it has, "been de-identified to the extent that it cannot be re-identified again" (Republic of South Africa, 2013).

The workflow geocodes using a gazetteer, a free API and commercial desktop product.. The cleansed address data is uploaded to Bing Maps and ArcGIS World Geocode Service. These services match the addresses data returning a confidence rating percentage in the case of ArcGIS and precision level in the case of Bing Maps and the gazetteer files.

The precision levels are descriptions of size such as 'neighbourhood', 'city/town', or 'country'. For our purposes the input used was the cleansed address format of: 'Suburb', 'Town', 'Postal Code', 'Country'. The output returned for Bing and gazetteer files can be seen below in Table 10.

Table 10: Geocoded output fields for Bing and gazetteer files

	Output in order returned						
BING	latitude	longitude	elevation	name	desc	source	precision
Gazetteer	country code	postal code	place name	latitude	longitude	accuracy	

The gazetteer postal files are imported into the SQL database. The postal codes are matched to the aggregated address data using the script below:

```
SELECT DISTINCT
                a.Suburb,
                a.Town,
                a.Postcode,
                a.SuburbCount,
```

```
                za.LAT,
                za.LONG
INTO GAZ_ADDR_CODED
FROM ADDR_COUNT a
LEFT OUTER JOIN POSTCODE_GAZ za on
       (a.Suburb = UPPER(za.PLACENAME) or a.Town = UPPER(za.PLACENAME))
       AND za.PCODE = a.Postcode
```

ArcGIS returns a more detailed version of the fields above including minimum X, maximum Y, display and standard deviation of the returned latitude and longitude. It also returns other descriptive identifiers such as 'Subregion', 'Region', and 'ARC_Neighb'. Examples in our data are 'Cape Winelands', 'Western Cape', and 'Stellenbosch', respectively.

3.5.1 Final geocoded dataset

Once the data is cleansed and geocoded as above, a new relation is created, containing:

- Suburb
- Town
- Postcode
- Longitude
- Latitude.

This can then be joined with the client data using the Suburb/Town/Postcode combination as a key. Essentially, this relation can then act as the geocoding reference table. It can be integrated into live capturing and updated when new unique combinations are created. However, usable entries first need to be distinguished from inaccurate or unreliable results.

3.6 Comparison

The next step after using the three geocoding sources is to compare the results of all three methods, and calculate the distance between the points returned by each method. When the 3 points match up within an accepted degree of accuracy, the geocoding is considered to be correct. The relationship between the sources for comparison can be seen below in Figure 5.

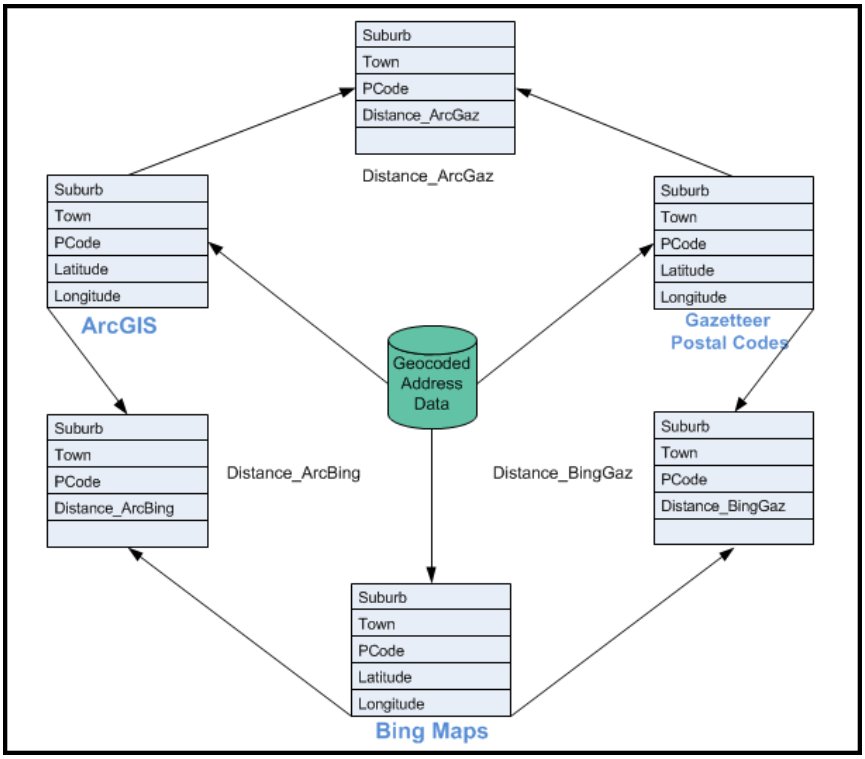


Figure 5: Comparing returned geocoded data

The formula to calculate distances in kilometres between coordinates is based on the 'Great Circle Distance' formula:

$$\Delta\hat{\sigma} = 2\arcsin\left(\sqrt{\sin^2\left(\frac{\Delta\phi}{2}\right) + \cos\phi_s\cos\phi_f\sin^2\left(\frac{\Delta\lambda}{2}\right)}\right)$$

Equation 1: 'Great Circle Distance' formula (BlueMM, 2007)

This calculation can be done in SQL using the Haversine formula as seen above in Equation 1. The radius is set to 6371, which is the average radius of the Earth in kilometres. A relation of unique Suburb, Town, and Postcode combinations is created with distance values for ArcGIS-vs-Bing, Bing-vs-gazetteer and ArcGIS-vs-gazetteer.

Once the three geocoding results have been obtained and discrepancies quantified as above, the client's acceptance threshold is employed to distinguish usable results from results requiring user intervention. The conversion for accuracy in kilometres vs degree in geographic coordinates can be seen in Table 11 below.

Table 11: Accuracy vs Degree (Thompson, 2011)

Accuracy vs Degree in Geographic Coordinates

1 sec arc = 1.68 feet = 51.233 cm ~ ½ meter

Decimal Places	Degrees	Distance
0	1	111 km
1	0.1	11.1 km
2	0.01	1.11 km
3	0.001	111 m
4	0.0001	11.1 m
5	0.00001	1.11 m
6	0.000001	0.111 m
7	0.0000001	1.11 cm
8	0.00000001	1.11 mm

3.7 Result Extraction

The final step in the workflow is presentation of the results of Geocoding. After matching between geocoders and applying the error threshold the businesses are comfortable with, an SSRS report is produced for the business users. They can then correct any erroneous data and update the CRM system. The final output is in the form of a map. The GIS application connects directly to the CRM database and uses the geocoded information to create a shapefile to map their client data. This data can be queried and combined with other variables from the client database. This allows the business to analyse their client data visually in new ways.

3.8 Workflow Overview

Figure 6 shows the final workflow for the entire geocoding process using aggregated data. This workflow provides the following benefits:

- Standardisation of address data
- Reporting for addresses that have been identified as erroneous by the address cleansing software
- Reduced geocoding costs by aggregating address data
- Adherence to POPI and privacy laws by aggregating address data
- Confirmation of geocoded results by cross checking multiple geocoders
- Reporting for invalid geocodes results
- Visual geographic analytics for client data

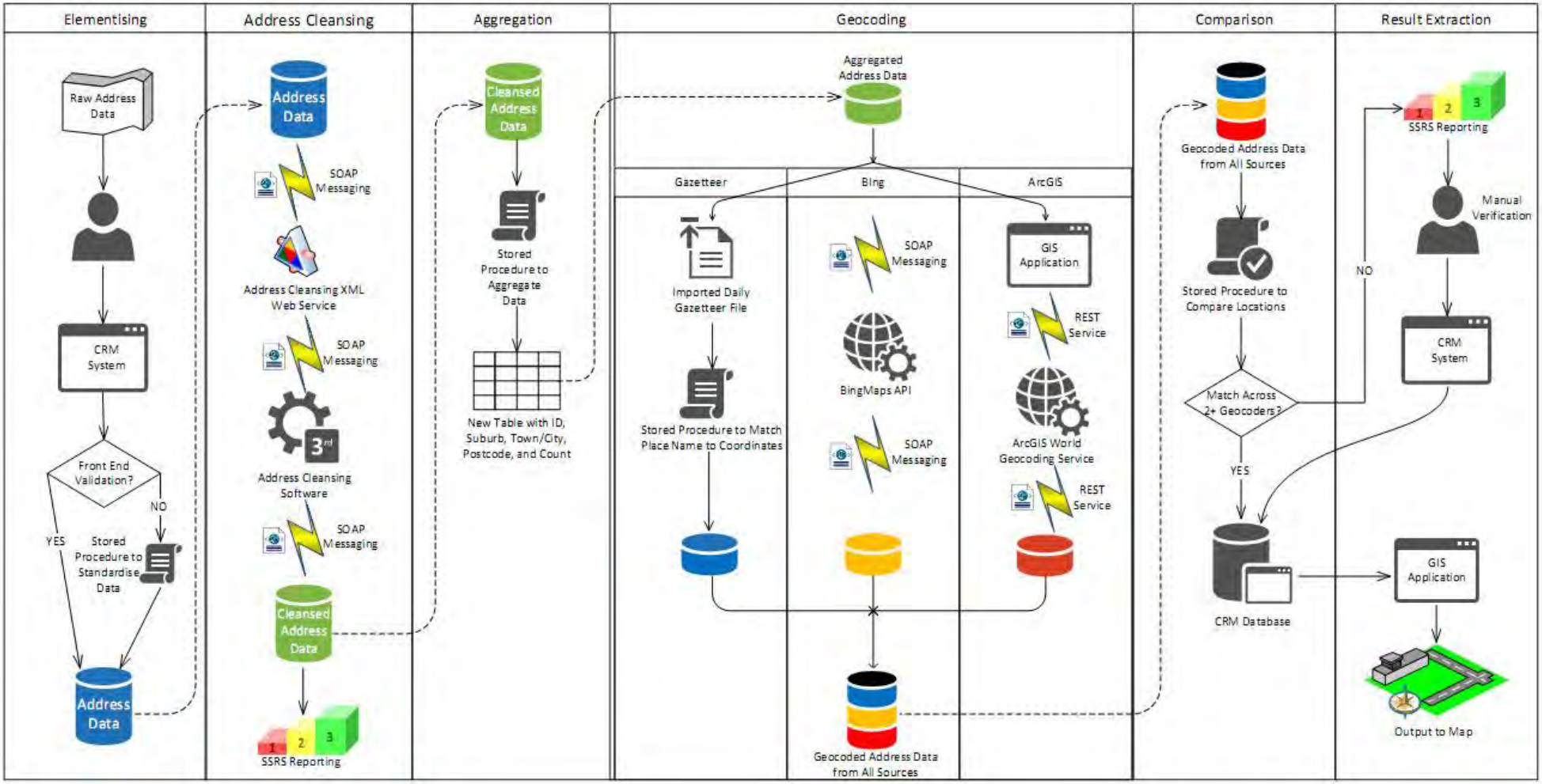


Figure 6: Final Workflow for Geocoding with Aggregated Data

4 Workflow Application and Results

4.1 Available Data

After the workflow was designed and developed, it was applied to two different datasets, the one belonging to a medium-sized financial institution and the other to a small magazine publishing company. The financial institution has address sets relating to clients and to brokers. The clients have residential addresses whereas the brokers have business addresses. The business addresses for brokers tend to be located near urban centres while the residential addresses can be urban, suburban, or rural. The publishing company similarly has addresses relating to two distinct categories, advertisers and subscribers. The advertisers tend to be businesses also located in or near urban centres. The subscribers can have business or residential addresses. The publishing company did not distinguish between the two subscriber types. The address can be urban, suburban, or rural. This chapter describes experiences using this address data in each workflow component during these two studies.

4.2 Raw Address Data

Raw address data can be collected in many forms prior to being entered into a CRM system. The address may or may not have been validated or come from a reliable source. Often an address is collected telephonically and subject to the interpretation of the capturer. This adds to the difficulties associated with accurate address data. Below in Table 12 are some of the most common origins of address data for the two companies.

Table 12: Origins of raw address data

Origin	Description	Capture method	Interface to CRM system	Financial Company	Publishing Company
Application Forms	The address is received on paper or static electronic forms. There is very loose structure and no validation.	Employee or consultant manual capture	NO	YES	YES
Proof of Address	A bank statement, cell phone bill, utility bill or other approved address verification document.	Employee or consultant manual capture	NO	YES	NO
Telephonic	The address is received verbally from the client. There is no verification or validation.	Employee or consultant manual capture	NO	YES	YES

Online	The address is received via the company's website. An online form has been configured with some validation for form and structure.	Client captured.	Yes - Financial No - Publishing	YES	YES
Third Party Sources	Marketing lists or addresses captured by outside sources. Validation unknown.	Third party	Possible imports or manual data capture.	YES	YES

4.3 Elementising

The financial institution uses a sophisticated CRM system which requires consultants to enter address data in a standardized way. The company has employed strict capturing standards for the past 5 years. The database itself goes back over 10 years, although a greater number of clients were captured in recent years.

The addresses of the financial institution were clean in terms of elements. The capturing inputs have been structured and named, which contributes to this. The inputs used were Address Line 1, Address Line 2, Address Line 3 (Suburb), City(Town), Postal Code and Country. Postal code is required when the country is South Africa. The company has a central SQL database as well as a data mart that address data can be mined from.

There was some confusion in the financial business with regards to the definition of 'Suburb' versus 'City' and these were used interchangeably at times by users capturing the data. For SAPO mailing standards, only 'Suburb' and 'Postcode' are used and other data is considered to be extraneous.

The publishing company uses a less sophisticated CRM system. It has the ability to store information as above, in address lines 1-3, city and postal code. However, there is no validation for any of the lines and no minimum requirements. The database is about twenty years old and has never been properly maintained for data quality. The system runs on an Access database. The address records were exported into CSV and reimported into a SQL database for processing and analysis.

The publishing company had not been following a specific capture method and as a result the data needed formatting to correctly separate the addresses into the field groupings above. The data had to be standardised using wildcard searches for known issues by the elementising component. These updates were applied directly to the company data.

4.4 Address Cleansing

Although the financial company has adhered to strict capturing standards, variations in address structure, spelling, and incorrect suburb, town, and postcode associations were found in almost all of the residential addresses. Only 298 addresses of the over two hundred thousand residential addresses processed were in the approved format and coded level '1_GREEN'. The financial company had standardization of case, as only uppercase letters were used. However, there were variations in how the addresses were captured as those fields were free-text. This led to differences in structure such as street, unit, box, suite, or bag number captured in the wrong order. In total, these slight changes amounted to 84% of the residential addresses processed. These were primarily found in code level '2_GREEN' and the database was updated to the suggested format without user intervention. Addresses coded yellow or red were not manually corrected or confirmed due to time constraints at the client, and hence were excluded from further processing.

In the smaller publishing company, there were no capturing standards, and as such, many contacts were missing postal codes. In this case, unless it is a very specific street name, it is almost impossible to correctly cleanse an address. It will return as an error. It is generally imperative that postal codes are captured to prevent this happening. Even when codes are incorrectly captured, there is a better chance of matching the area than when there is no code at all.

The larger financial company did not have this issue as postal codes are compulsory on all captured addresses. Validation has been built into the capturing software to prevent the user from leaving this field blank when the address is in South Africa.

Because the address cleansing software is designed for postal compliance, the levels of acceptance can appear very differently when using residential vs business or postal addresses. It is worth noting that while an address may comply with SAPO standards, the address details may not be correct. The SAPO standard focuses specifically on the formatting of the address and a correct postal code/suburb combination. Conversely, an address could be deliverable by courier, but does not comply with SAPO standards. For instance, subscription clients are posted regularly and receive a publication, but do not necessarily conform to SAPO postal standards.

Overall, the financial company had an 84% SAPO acceptance rate for client residential address, and a 97% acceptance rate for broker business addresses. The publishing company had only a 75% SAPO

acceptance rate for their business clients and 80% acceptance rate for their subscription clients which could be either residential or postal address, as shown in Table 13.

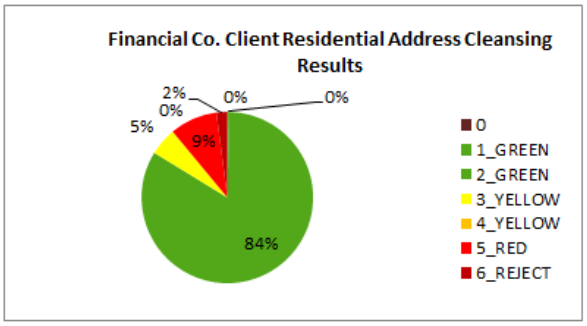
Table 13: Acceptance rate for address cleansing (SAPO Standards)

	Acceptance Rate	Manual Intervention
Publishing Business	75%	25%
Publishing Subscriptions	80%	20%
Financial Broker (Business)	97%	3%
Financial Client (Residential)	84%	16%

4.4.1.1 Cleansing results comparison

Figure 7 and Figure 8 below show comparisons of the address cleansing results between the two companies as well as the difference in address type. While the business broker addresses for the financial company had the highest acceptance rate, it is worth noting that the total numbers are much smaller than those of the residential company. The business addresses also are predominately found in common urban centres that are often located in densely populated areas. The financial company also requires registered company information for their brokers which tends to be more accurate than the proof of residence varieties that are provided by clients.

Financial Co. Client Residential Addresses	
Acceptance Level	Count of AddrID
0	18
1_GREEN	698
2_GREEN	168755
3_YELLOW	10542
4_YELLOW	32
5_RED	18115
6_REJECT	4167
Grand Total	202327



Financial Co. Broker Business Addresses	
Acceptance Level	Count of AddrID
1_GREEN	2408
2_GREEN	333
3_YELLOW	10
5_RED	63
6_REJECT	22
Grand Total	2836

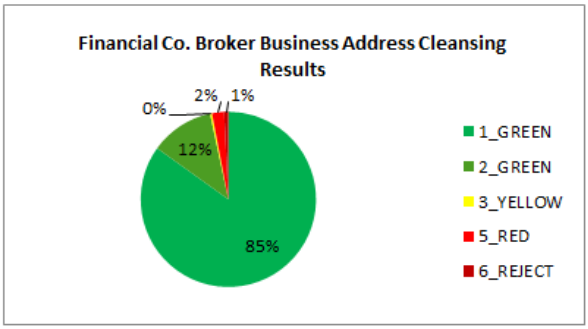
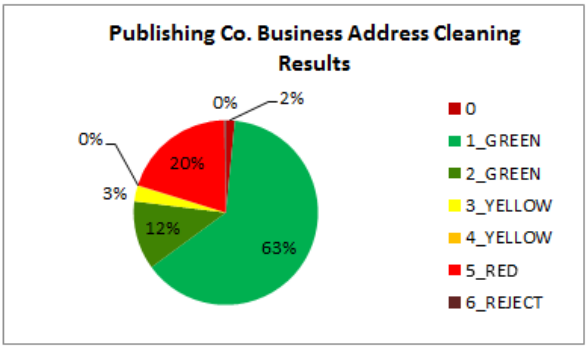


Figure 7: Financial co. address cleansing results

Publishing Co. Business Addresses	
Acceptance Level	Count of ContactID
0	36
1_GREEN	1449
2_GREEN	275
3_YELLOW	62
4_YELLOW	3
5_RED	456
6_REJECT	6
Grand Total	2287



Publishing Co. Subscriptions Addresses	
Acceptance Level	Count of ContactID
0	28
1_GREEN	1647
2_GREEN	385
3_YELLOW	93
4_YELLOW	3
5_RED	390
6_REJECT	5
Grand Total	2551

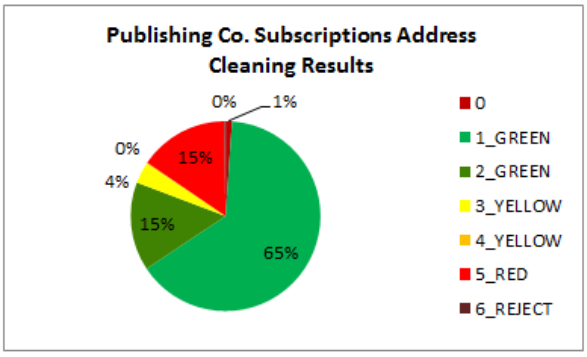


Figure 8: Publishing co. address cleansing results

4.5 Aggregation

Prior to cleansing, in the financial institution's data, there were 29366 unique combinations of suburb, town, and postal code. After cleansing, this number fell to 6464. (See Table 14 below)

The difference in numbers is due to a number of reasons. Amongst the top reasons were misspellings, suburbs set as towns and vice versa, incorrect postal codes, postal boxes with street addresses, and language differentiations.

Table 14: Count of unique address combinations cleansed data vs dirty data

Unique Combination	Uncleansed	Cleansed
Suburb, Town and Postal Code	29266	6464
Town and Postal Code	23203	4431
Postal Code	3256	2011

Table 15 below is a detailed breakdown of the top 10 postal codes encountered in our data. There were over six hundred percent more suburb and town variations in the top postal code before cleansing than after. In 8001, Cape Town central, there were over sixteen hundred percent additional variations. This is most likely due to confusion between the definition of the 'greater Cape Town area' versus the actual designated Cape Town city boundaries.

Table 15: Top 10 postal code ranks before and after cleansing

Postcode Unclean	Rank Before	Rank After	Count of Suburb & Town Variations	Count After	Percent More Before
2191	1	2	661	107	618%
0157	2	1	614	126	487%
2196	3	4	550	78	705%
0081	4	15	363	59	615%
2194	5	12	347	61	569%
0181	6	10	326	63	517%
7441	7	25	321	45	713%
7530	8	16	289	57	507%
8001	9	89	270	16	1688%
1724	10	7	261	72	363%

Rank	Postcode Clean	Count of Suburb & Town Variations
1	0157	126
2	2191	107
3	1459	81
4	2196	78
5	3201	76
6	1709	72
7	1724	72
8	1501	68
9	1619	63
10	0181	63

An example in Table 16 is the top postal code, 0157 in the Centurion area in the Gauteng province near Pretoria. In postcode 0157, prior to cleansing the financial institution's data, there were 614 unique combinations of town/suburb to postcode, with 338 unique town to postcode combinations. After cleansing, this fell to 126 unique combinations of town/suburb to postcode and 9 town to postcode combinations. Centurion itself was spelled in 9 different ways prior to cleansing. This sort of finding will be common in any organisation where addresses are captured telephonically or even captured by the clients themselves.

If left uncleansed, the geocoding process also becomes more difficult as a correct match cannot be made to a set of coordinates. Table 17 shows how many variations of suburb and town could be found in a single postcode.

Table 16: Comparison of towns in the 0157 postal code before and after cleansing

Towns in 0157 Uncleansed	Count of Suburb
Total Count in Postcode 0157	614
CENTURION	164
PRETORIA	20
IRENE	16
HIGHVELD	11
LYTTELTON	10
KOSMOSDAL	10
CELTISDAL	8
WIERDAPARK	7
… 329 other combinations	…

Towns in 0157 Cleansed	Count of Suburb
Total count in Postcode 0157	126
CENTURION	75
PRETORIA	13
ROOIHUISKRAAL	12
WIERDAPARK	10
THE REEDS	6
HIGHVELD	6
KOSMOSDAL	2
CORNWALL HILL	1
RIETVALLEI PARK	1

4.6 Geocoding

The data for both the financial company and the publishing company were analysed using the three selected methods. There was also a distinction made between the address types for geocoding. The addresses were geocoded in groups according to company and address type. The groups were described as follows:

- Financial company, business addresses (brokers)
- Financial company, residential addresses (clients)
- Publishing company, business addresses (advertisers)
- Publishing company, residential addresses (subscribers)

4.6.1 Gazetteer postal code geocoding

The gazetteer data dumps were uploaded to the client database and address data matched based on and postcode, or town and postcode. In our file most accuracy levels were defined as level '4'. Accuracy in the file is defined as, "1=estimated to 6=centroid". (Geonames.org, 2014)

The unique key within the file is 'postal code' and 'place name' which can be a suburb, a city, a town, or administrative zone. One of the main issues identified with the postal code files was that many postal codes, as well as many place names, mapped to multiple coordinate points, as shown in table 17. This is because town/suburbs/cites can encompass more than one postal code. This means that a 'greater city area' such as Milnerton or Cape Town, can have dozens of associated postal codes and geographic points. The script also returned about 10% of records with no match to the gazetteer files.

Table 17: Duplicated Suburbs/Postcodes in the gazetteer postal code files

country code	post code	place name	admin name1	admin code1	admin name2	admin code2	admin name3	admin code3	latitude	longitude	accur.
ZA	7440	Melkbosstrand							-33.7306	18.4361	4
ZA	7441	Melkbosstrand							-33.7306	18.4361	4
ZA	7441	Milnerton							-33.8213	18.4454	
ZA	7441	Table View							-33.8167	18.4833	4
ZA	7441	Bloubergrant							-33.8213	18.4454	
ZA	7441	Cape Town							-33.9167	18.4167	4
ZA	7441	Bloubergstrand							-33.8213	18.4454	
ZA	7442	Milnerton							-33.8667	18.4833	4
ZA	7443	Milnerton							-33.9166	18.4166	4
ZA	7444	Bloubergrant							-33.8326	18.4632	
ZA	7446	Milnerton							-33.8667	18.4833	
ZA	7447	Milnerton							-33.8667	18.4833	4
ZA	7449	Milnerton							-33.8667	18.4833	4
ZA	7449	Bloubergrant							-33.8667	18.4833	

In the analysis for the gazetteer postal code files for clients in the financial company (residential addresses), 152 of 5468 suburbs or postal codes were associated with more than one geocode location.

4.6.2 Bing Maps

For the residential client database, Bing Maps brought back the worst match rate, with 461 of 5395 entries returning a zero. If Bing Maps was being used alone for GIS study, it would be advisable to retrieve those missing values from other sources.

Even after missing results were excluded a number of other outliers were discovered. The maximum range between Bing Maps and the other sources was 7757.54km. This distance is greater than the entire length of South Africa. On investigation, it was discovered that two of the values were returned with a positive latitude point instead of a negative latitude point. (i.e., 33.9833 vs -33.9833)

4.7 Financial company results

4.7.1 ArcGIS broker results for the financial company

4.7.1.1 Broker data

ArcGIS was determined to be the most accurate of the three methods for the financial company. Broker data in particular had a higher accuracy rate using ArcGIS. In the first pass of geocoding, the results returned all points within South Africa and were clustered in expected groupings corresponding to larger urban areas. Below, in Figure 9, clustering is based on uniquely returned locations, not on the total number of brokers. The clustering corresponds relatively to the size of the matching urban area.

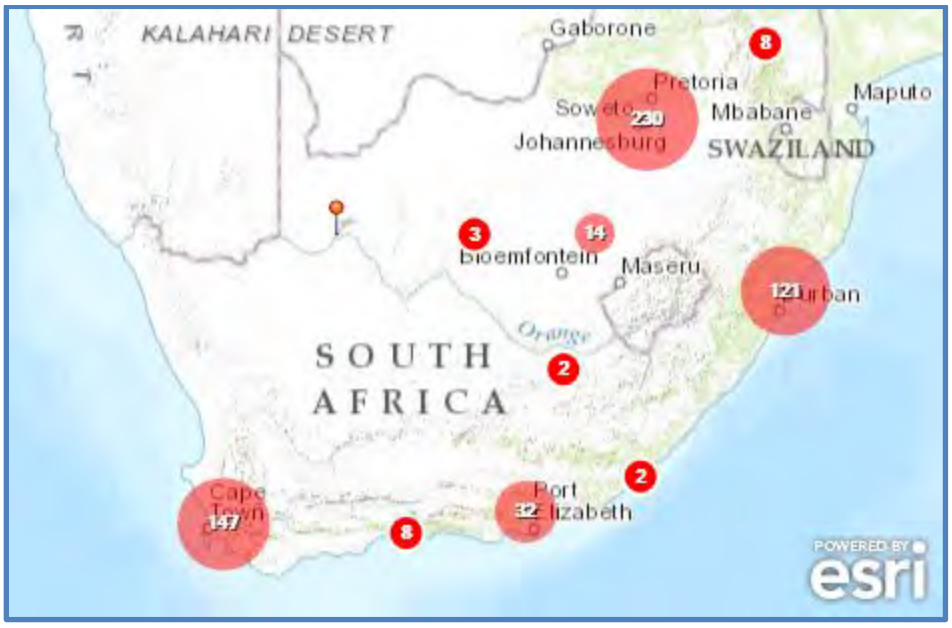


Figure 9: ArcGIS broker (business address) clustering

4.7.1.2 Gazetteer broker results compared to ArcGIS

If gazetteer uploads had been used alone, a total of 1162 unique suburb, town and postal code combinations would have differed by more than 11.1km from their ArcGIS geocoding. This counts for 30460 of 177808 records, or 17.13% margin of error in the individual records. This can be seen by comparing the numbers in Figure 9 above with Figure 10 below. In all clusters excluding the greater Johannesburg area and the Garden Route, the total numbers are different.

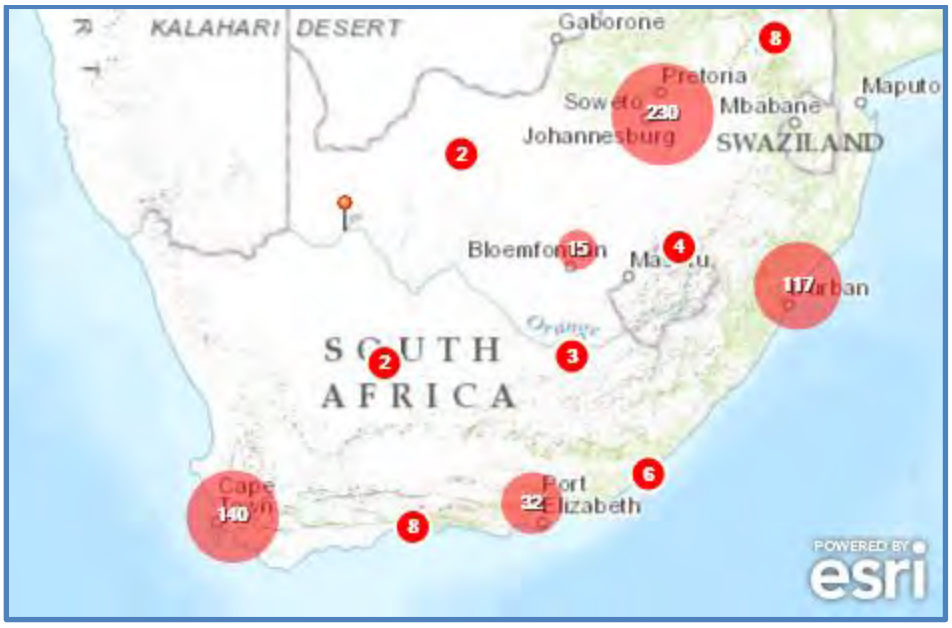


Figure 10: Gazetteer Broker (business address) clustering

4.7.1.3 Bing broker results compared to ArcGIS

For brokers, there were many blank rows returned with the Bing API. It is possible to re-run them through the API, but it does not have the same matching capability as ArcGIS. Subsequent passes returned little or no improvements. ArcGIS matched all but 2 rows with only 24 rows not matching at a score of 100%. The 24 rows that did not match at 100% had a match score between 85.74% and 89.09%.

Bing also returned some obviously incorrect rows. The 'Country' field was designated as 'SOUTH AFRICA', however, in some instances where it did not recognise the suburb and it had 'Town' as 'JOHANNESBURG' it mistakenly geocoded the record in Austria where there is a place named 'Johannesberg'. This was not consistent for all suburbs, as some were correctly recognised. Johannesburg was also found in California and Florida.

The Bing results for broker clustering (of unique Suburb, Town, Postcode combinations) were scattered throughout the world (See Figure 11 below) even though all addresses data had the country designation of 'SOUTH AFRICA'. The first pass can be seen in Figure 11 and the second pass can be seen in Figure 12.

Figure 11: Bing Maps broker (business address) clustering, first pass

Figure 12: Bing Maps broker (business address) clustering, second pass

4.8 Publishing company results

The need for standardisation when attempting a project of this nature can be clearly seen by looking at the results of the geocoding attempts for the publishing company. The number of records processed was significantly less than for the financial company. The aggregated suburb/town/postcode combination for subscriptions, which are equivalent to clients in the financial company, was 1345, compared with 5395 finance aggregates. A summary of the results for both companies can be seen below in Table 18.

Poor data quality was a known problem for the publishing company, and as such the geocoders had a number of issues matching locations. This was most evident with the gazetteer postal code files. They are very basic tables with suburb and town relations to postal codes. The high error rate is directly related to the large number of missing postal codes. When geocoding at our level of detail it was the most important element of the address. Without it, the record has a reduced propensity to be matched correctly. This is due in part to the duplication of suburb names and the fact that the address cleansing software is also unable to correct the address without a postal code. ArcGIS uses a matching score based on matches to the highest level of precision. It is possible for an address to match to different reference features and return a tied score. In this case there is a setting for picking an arbitrary match.

Table 18: Geocoding results for both companies

Publishing Co.	Business	Subs
Number of records	2311	2553
Count of suburb (aggregated prior to cleansing)	1025	1345
Cleansed records aggregated	833	1202
Missing postal code after cleansing	420	344
No match in gazetteer	620	837
No match by Bing	1	51
Bing country only match*	118	102
ArcGIS matched %	95.7%	95.6%
ArcGIS tied %	2.4%	2.3%
ArcGIS unmatched	1.9%	2.1%
ArcGIS < 100% match	14.8%	7.5%
**South Africa centroid*	*-29.0461845*	*25.0628796*

Financial Co.	Brokers	Clients
Number of records	2836	202327
Count of suburb (aggregated prior to cleansing)	1226	29266
Cleansed records aggregated	568	5395
Missing postal code after cleansing	N/A	N/A
No match in gazetteer	44	267
No match by Bing	38	461
Bing country only match*	14	412
ArcGIS matched %	96.3%	98.8%
ArcGIS tied %	3.7%	1.1%
ArcGIS unmatched	0%	0.1%
ArcGIS < 100% match	0.9%	0.4%
**South Africa centroid*	*-29.0461845*	*25.0628796*

The 'Count of Suburb' in Table 18 above indicates the number of unique Suburb/Town/Postcode combinations prior to address cleansing. This can be seen as a cluster for reference in Figure 13. Clustering was also able to assist visually by reducing 833 business address combinations to five groupings, and six groupings for subscription addresses.

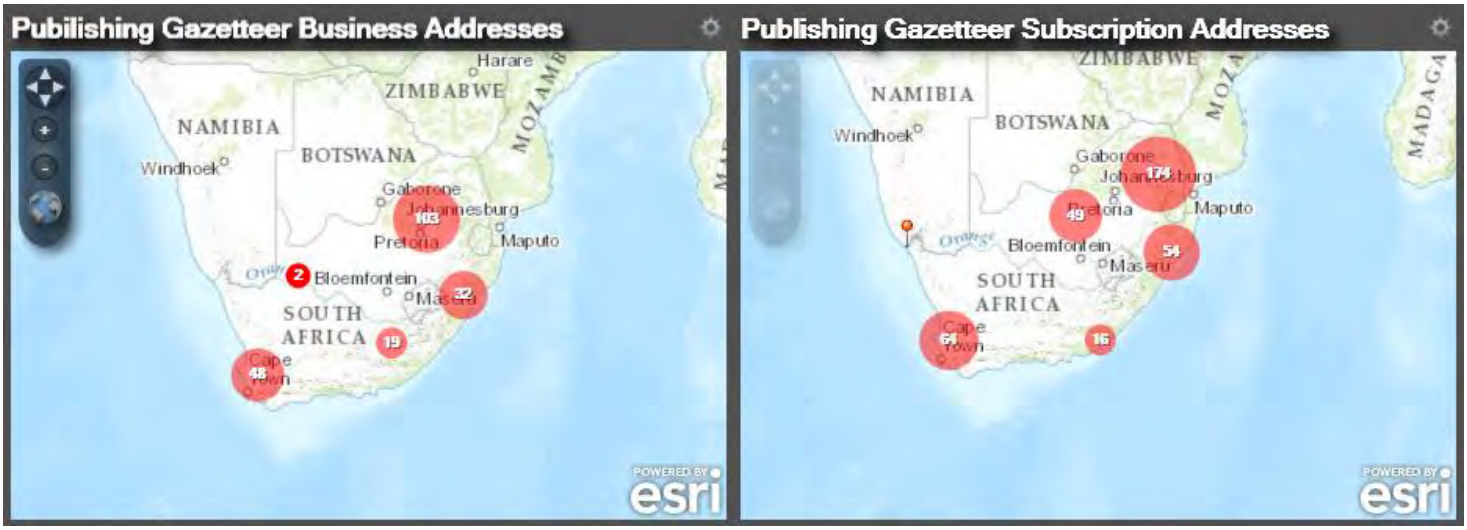


Figure 13: Exact suburb/postal code matches for the publishing co. using gazetteer postal code files

Clustering was useful for spotting obvious errors such as addresses mistakenly geocoded outside of South Africa. Our geocoding input for the publishing company contained only South African addresses, yet because some addresses did not have postal codes, they were matched incorrectly. These can quickly be identified as in Figure 14 below.

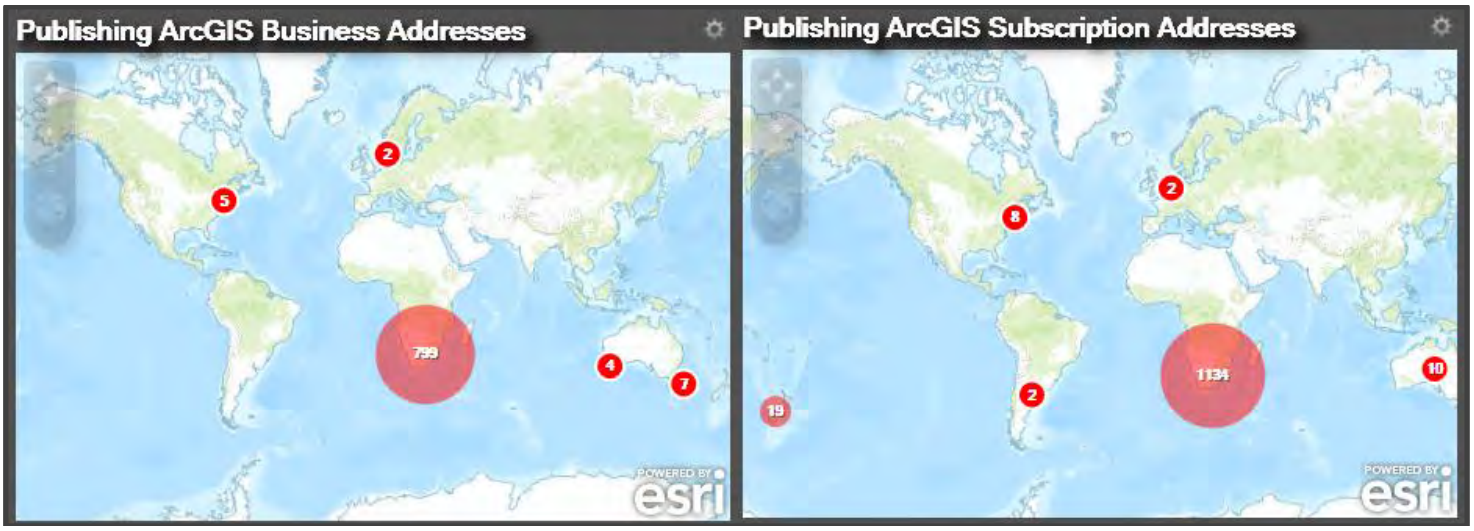


Figure 14: ArcGIS mapped a number of points without postal codes outside of South Africa

The goal of all geocoders is to match every record to a specific point using latitude and longitude; however, sometimes this point will match to a generic centroid that is not useful for analysis. An example where clustering became a useful tool for identifying these incorrectly geocoded addresses can be seen below in Figure 15. In over a hundred cases, Bing Maps matched only the country, placing a number of points in the Karoo just outside Kimberly. This point is the South Africa country centroid.

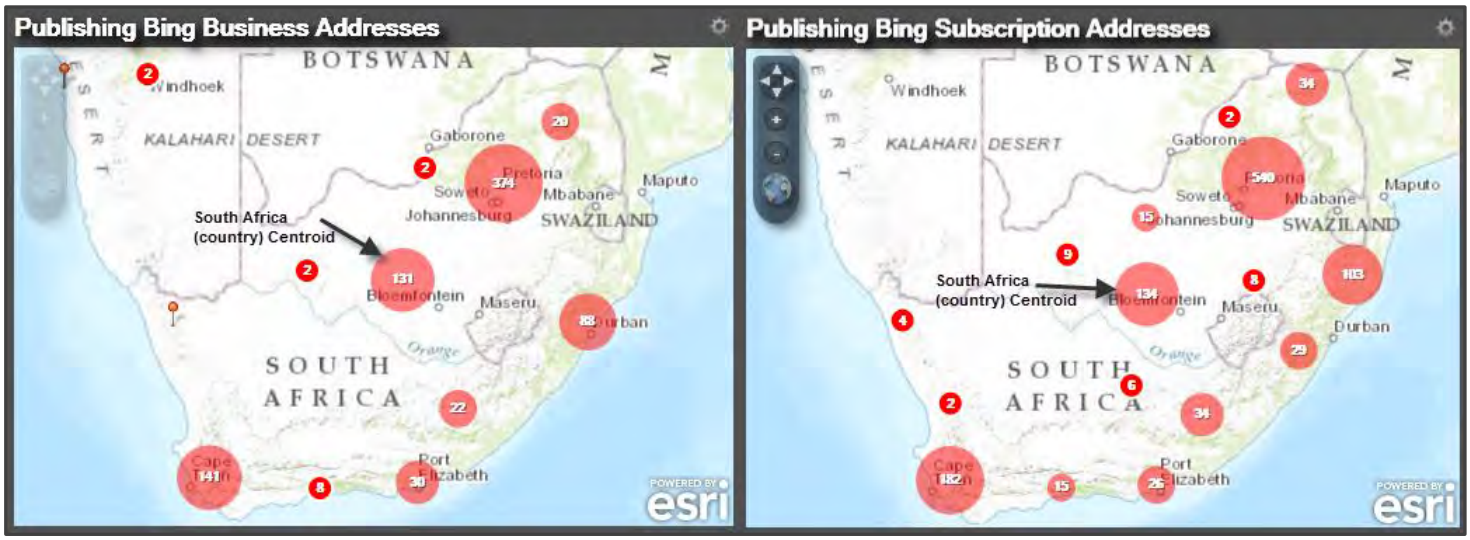


Figure 15: Bing Maps matched many records to the 'country' centroid

4.8.1 Verifying geocoded results

The workflow, being a semi-automated solution, recommends manual checking of results. This is best done by mapping and visually checking this directly against population density maps and by postal code value (since similar postal codes should be geographically clustered together). The accuracy of the gazetteer came into question when verifying the data in this way. The financial company was geocoded initially. The data was first checked for obvious inaccuracies. For example, the Blouberg suburb in the 7441 postcode was defined in 6 different ways, with 3 unique geocodes. Visually, the discrepancy was more readily spotted using the map shown below in Figure 16, as one point ended up in the ocean. This emphasises the importance of knowing the data and the area as it is paramount to successful analysis when using this method.

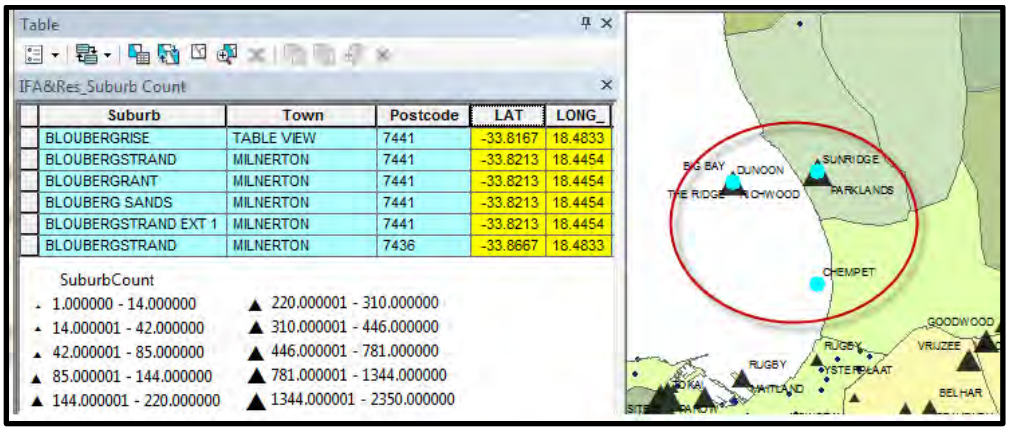


Figure 16: Postal code files, geocoded data in the ocean

The cleaned result set was mapped to compare the count of client vs the count per population normalized. Figure 17 below shows larger blue triangles in areas circled in red. These areas have high numbers of clients, but low population. While this is possible, it warrants investigation. Closer inspection shows that the town names are similar to suburb names in the major cities. These are circled in yellow. Figure 17 also shows how it is common for the same suburb name to be found in multiple postal codes. Just as this complicates the address cleansing effort, it can also complicate the geocoding effort.

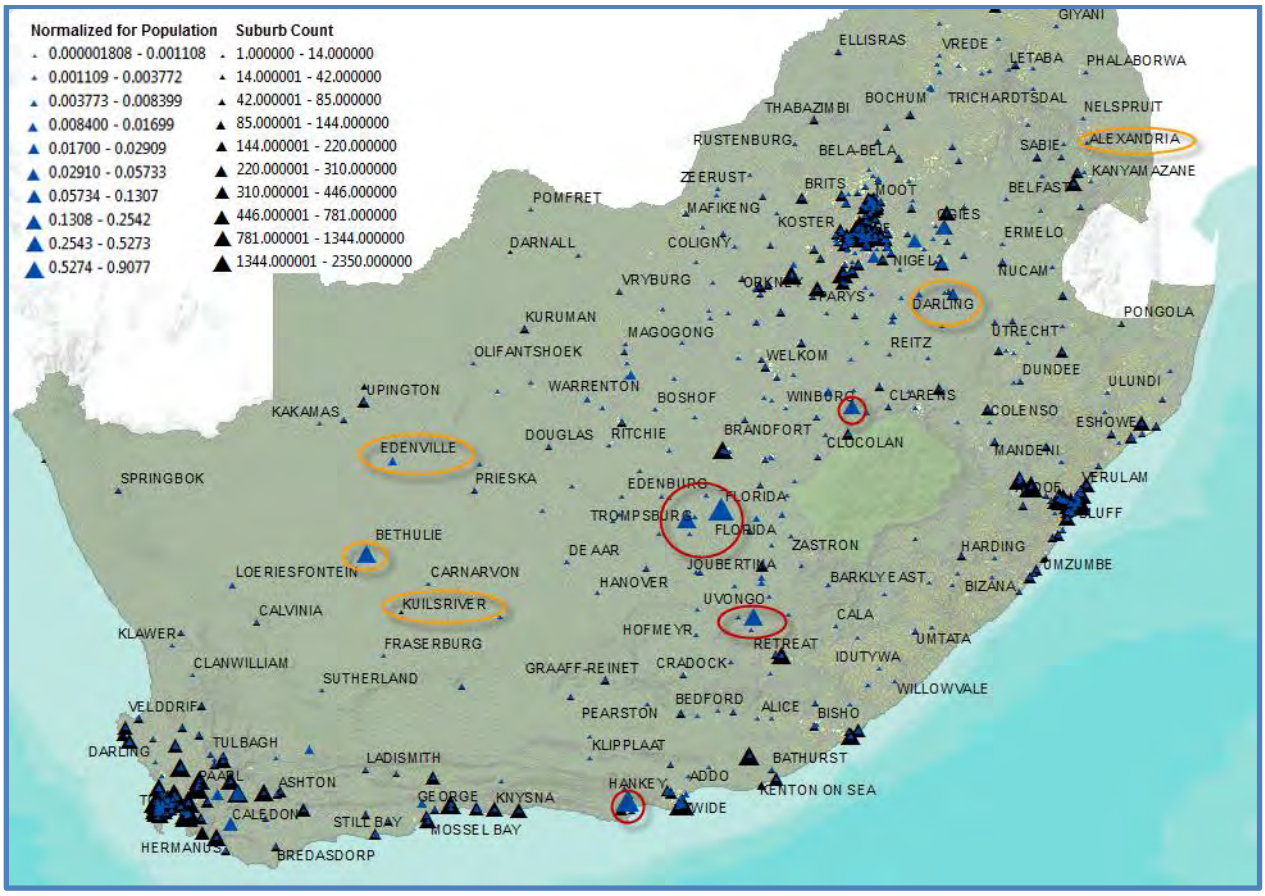


Figure 17: Count of suburb normalized for population shows anomalies in sparsely populated areas

The clients were mapped by postal code as well in order to verify accuracy. These are circled in Figure 18. Postal codes that are not in the correct range by area are clearly evident.

4.9 Comparison

After using the 3 geocoders, the workflow requires a target distance such that geocoder result differences within this distance of each other are considered acceptable and usable, but those where results are more distant than this are unacceptable. We found that in all instances, reducing precision by allowing a greater distance margin between points returned a lower error rate. This was true regardless of geocoder combinations, business type and types of address. All three geocoders worked with the same input data, yet often the point returned could be significantly different. The geocoding tool is only as good as its underlying geographic database which may not be accurate or updated regularly. When comparing the unique combination of suburb, town and postcode and the distance between pairs for all three geocoders, for residential addresses, the average distance between the returned coordinates was 16.89km. Based on this average, results were compared starting at 0.1° (11.1km). For each pair of geocoders, the percentage of results that differed by more than the target distance was computed. For

example, when looking at the publishing company business addresses and comparing ArcGIS and Bing, 17% of the returned points had a difference in distance of over 22.2km. In geographic coordinates an example is the returned points for postal code 2198, Johannesburg. ArcGIS returned (-26.1872, 28.0657) whereas Bing returned (-26.0374, 27.9176). This relates to a difference of 22.27km. Below in Figure 19 is the postal code 2198 with the points of Bing and ArcGIS mapped out. In red on the last map is the boundary of the postal code area. In black is the perimeter distance of the postal area, which if navigated would be in total about 22km.

Looking at the red line in the far right map shows the approximate postal boundary. This shows a postal area that is not uniform in size and can also contain multiple suburbs. A table of the percentage of error based on the target distance for each business and address type can be seen below in Table 19.

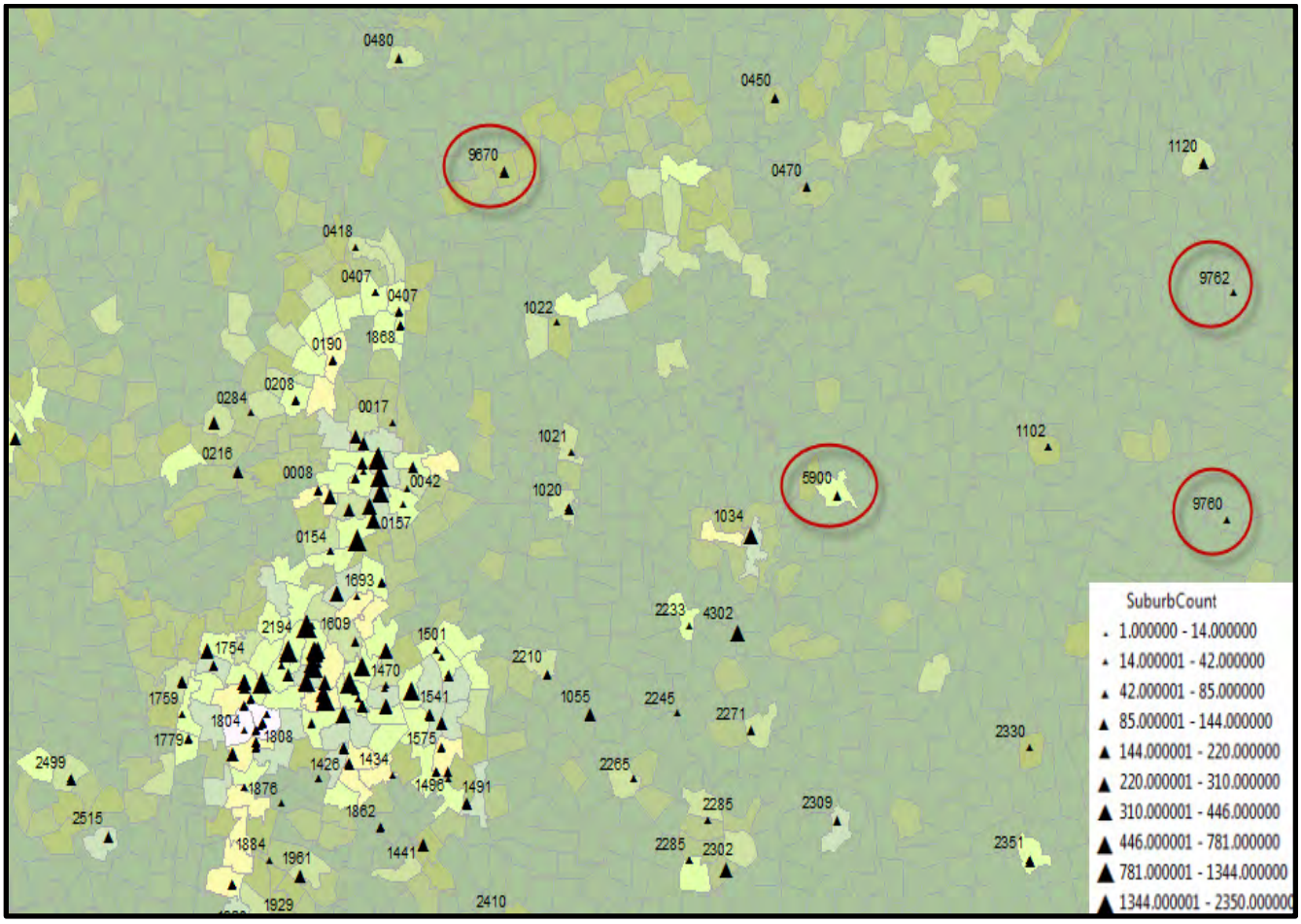


Figure 18: Using postal code geographic segmentation to isolate anomalies

Figure 19: Postal code with postal boundary (red line) and distance guide line in black (Google, 2014)

Table 19: Distance vs accuracy when comparing geocoded results; percent error based on distance target

Publishing Business	ArcGIS and gazetteer	ArcGIS and Bing	Bing and gazetteer	Average
0.1°(11.1km)	41%	38%	33%	**38%**
0.2°(22.2km)	16%	17%	19%	**17%**
0.3°(33.3km)	12%	13%	18%	**15%**
0.4°(44.4km)	12%	13%	18%	**14%**
0.5°(55.5km)	12%	13%	18%	**14%**
Publishing Subscriptions	**ArcGIS and gazetteer**	**ArcGIS and Bing**	**Bing and gazetteer**	**Average**
0.1°(11.1km)	50%	43%	35%	**42%**
0.2°(22.2km)	33%	26%	29%	**29%**
0.3°(33.3km)	29%	23%	27%	**26%**
0.4°(44.4km)	27%	22%	27%	**25%**
0.5°(55.5km)	27%	21%	26%	**25%**
Financial Broker (Business)	**ArcGIS and gazetteer**	**ArcGIS and Bing**	**Bing and gazetteer**	**Average**
0.1°(11.1km)	29%	47%	30%	**35%**
0.2°(22.2km)	8%	30%	28%	**22%**
0.3°(33.3km)	4%	27%	27%	**19%**
0.4°(44.4km)	3%	26%	27%	**19%**
0.5°(55.5km)	1%	26%	27%	**18%**

Financial Client (Residential)	ArcGIS and gazetteer	ArcGIS and Bing	Bing and gazetteer	Average
0.1°(11.1km)	21%	33%	26%	**27%**
0.2°(22.2km)	7%	21%	21%	**17%**
0.3°(33.3km)	3%	20%	20%	**14%**
0.4°(44.4km)	3%	19%	19%	**14%**
0.5°(55.5km)	2%	19%	19%	**13%**

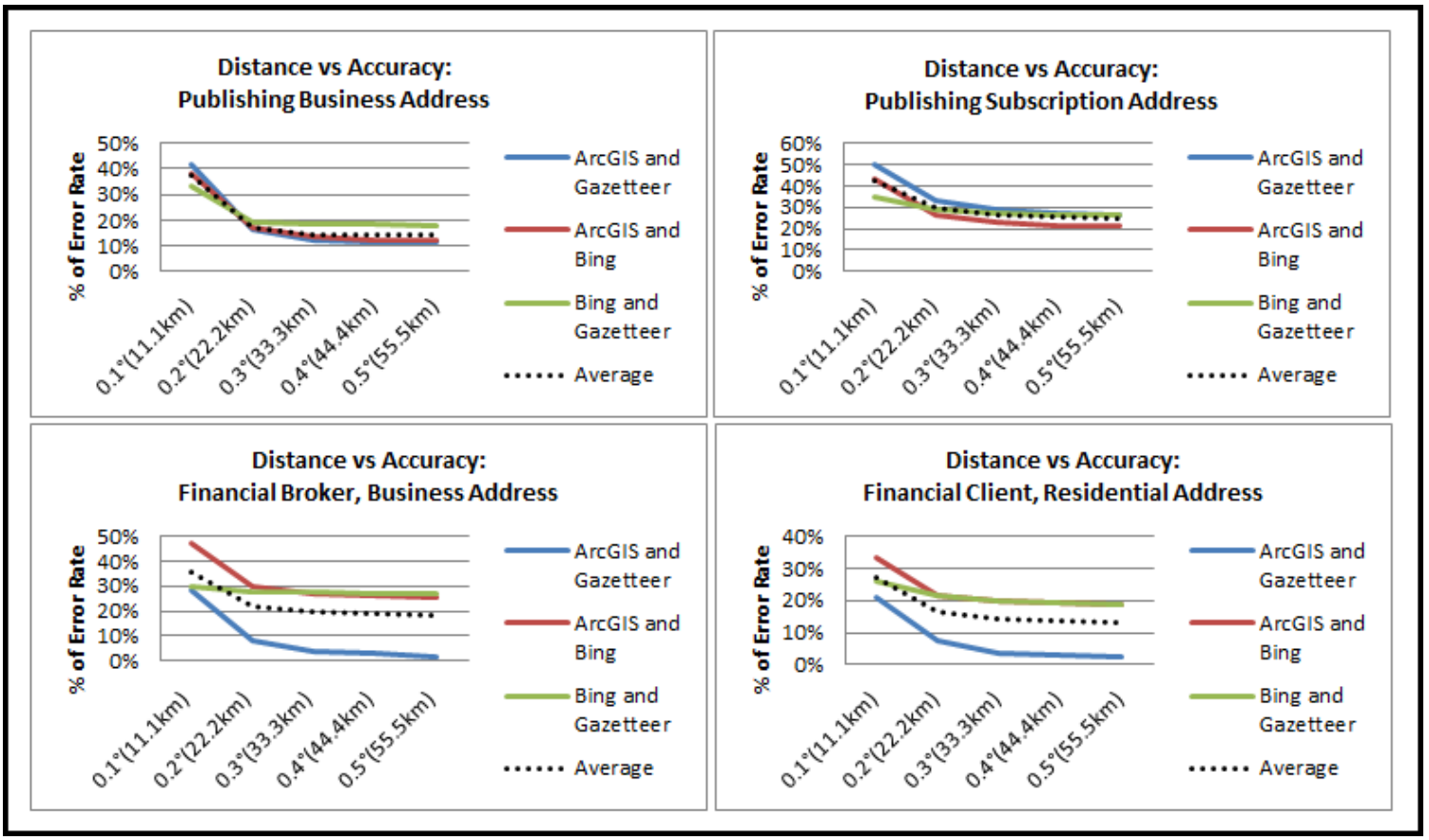


Figure 20: Distance vs Accuracy; Error rates fall as the accepted distance variation increases

The graphs in Figure 20 show the differences in not only the geocoders, but also the business data quality. We knew that the publishing company had inferior data quality, and this is shown in the numbers as the trend levels off at a greater percentage of error. Mapping with distances varying up to 55.5km, calculating what percentage of the population fell outside of the accepted distance for each comparison, indicated that a threshold of 22.2km (0.2°) should be used for our two companies. It is also

worth noting that after 22.2km, the gain from increasing the acceptance distance tapers off in all instances. An acceptance threshold of 22.2km (0.2°) is thus recommended to distinguish usable results from those requiring human intervention.

Figure 21 and Figure 22 show the proportion of data for which all 3 results are within the target distance, as well as the proportion for which any 2 of the 3 results are within that target distance of each other. In what follows, we refer to points that are within the acceptance target distance of each other as matching points.

For the financial company's residential addresses, there was only a 2% difference between two and three matched pairs. This most likely can be attributed to the validation that is required for new clients, including proof of addresses. On the contrary, Bing had problems with the broker file, which can be seen by only 10% having all 3 points match (for 11.1km acceptance target distance). The publishing company had similar numbers in the subscription addresses with the percentage differing between two and three matched pairs at 3% (for a target distance of 22.2km).

Financial Company - Broker @ 22.2km	
TwoClose	Count of TwoClose
N	160
Y	408
Grand Total	568
AllThreeClose	Count of AllThreeClose
N	390
Y	178
Grand Total	568

Financial Company - Broker @ 11.1km	
TwoClose	Count of TwoClose
N	286
Y	282
Grand Total	568
AllThreeClose	Count of AllThreeClose
N	513
Y	55
Grand Total	568

Financial Company - Res @ 22.2km	
TwoClose	Count of TwoClose
N	1253
Y	4142
Grand Total	5395
AllThreeClose	Count of AllThreeClose
N	1355
Y	4040
Grand Total	5395

Financial Company - Res @ 11.1km	
TwoClose	Count of TwoClose
N	1826
Y	3569
Grand Total	5395
AllThreeClose	Count of AllThreeClose
N	2177
Y	3218
Grand Total	5395

Figure 21: Financial company; two vs. three matches in range

Publishing Company - Bus @ 22.2km	
TwoClose	Count of TwoClose
N	267
Y	566
Grand Total	833

AllThreeC	Count of AllThreeClose
N	305
Y	528
Grand Total	833

Publishing Company - Bus @ 11.1km	
TwoClose	Count of TwoClose
N	388
Y	445
Grand Total	833

AllThreeC	Count of AllThreeClose
N	470
Y	363
Grand Total	833

Publishing Company - Subs @ 22.2km	
TwoClose	Count of TwoClose
N	424
Y	778
Grand Total	1202

AllThreeC	Count of AllThreeClose
N	451
Y	751
Grand Total	1202

Publishing Company - Subs @ 11.1km	
TwoClose	Count of TwoClose
N	614
Y	588
Grand Total	1202

AllThreeC	Count of AllThreeClose
N	684
Y	518
Grand Total	1202

Figure 22: Publishing company; two vs three matches in range

5 Spatial Analytics Examples

The motivation behind geocoding of data based on address values is that it enables GIS software to be used to visually display the data geographically. Spatial analytics can be used to better understand the data and identify possible strategies for change. This chapter outlines some of the insights gained from the final product of the application of the workflow to the publishing and financial companies. The data included in the maps below are actual examples from the result set of this study. The acceptance threshold of 22.2km (0.2°) was used in the maps.

5.1 Financial company map analysis

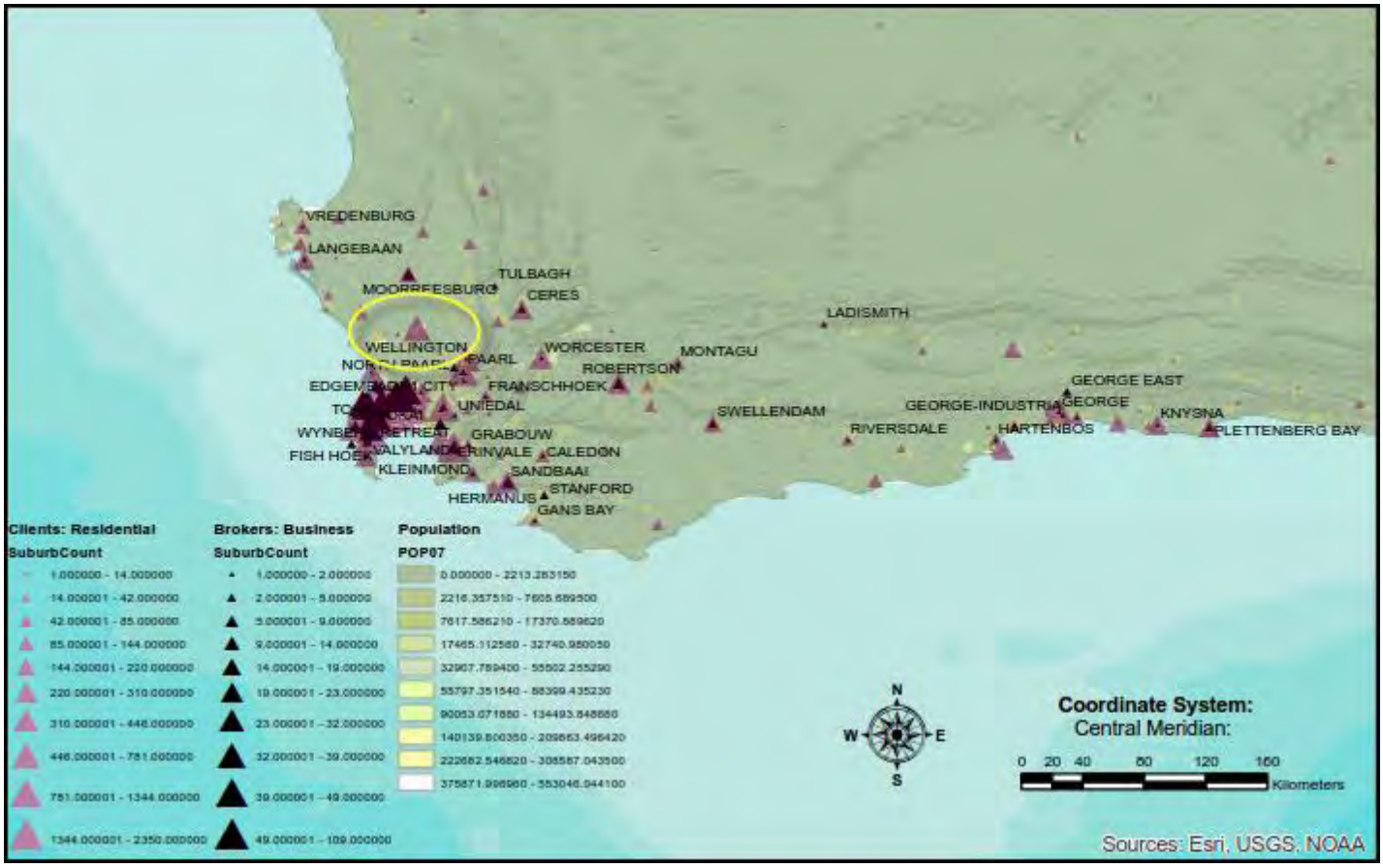


Figure 23: Client count relative to broker count in the Western Cape

Figure 23 shows the Western Cape client numbers relative to broker numbers. What is interesting about this map is that there are areas that are mostly independent from broker influence, yet still have a

high client count. An example is Wellington, in the yellow, highlighted circle. This could indicate an area that has bloomed in spite of lacking broker involvement, or could indicate an area that no longer has a broker and possibly needs to have one replaced. Either scenario gives the business a reason to investigate.

Account balances were joined in and the result was an average balance by suburb as per the map in Figure 24 below.

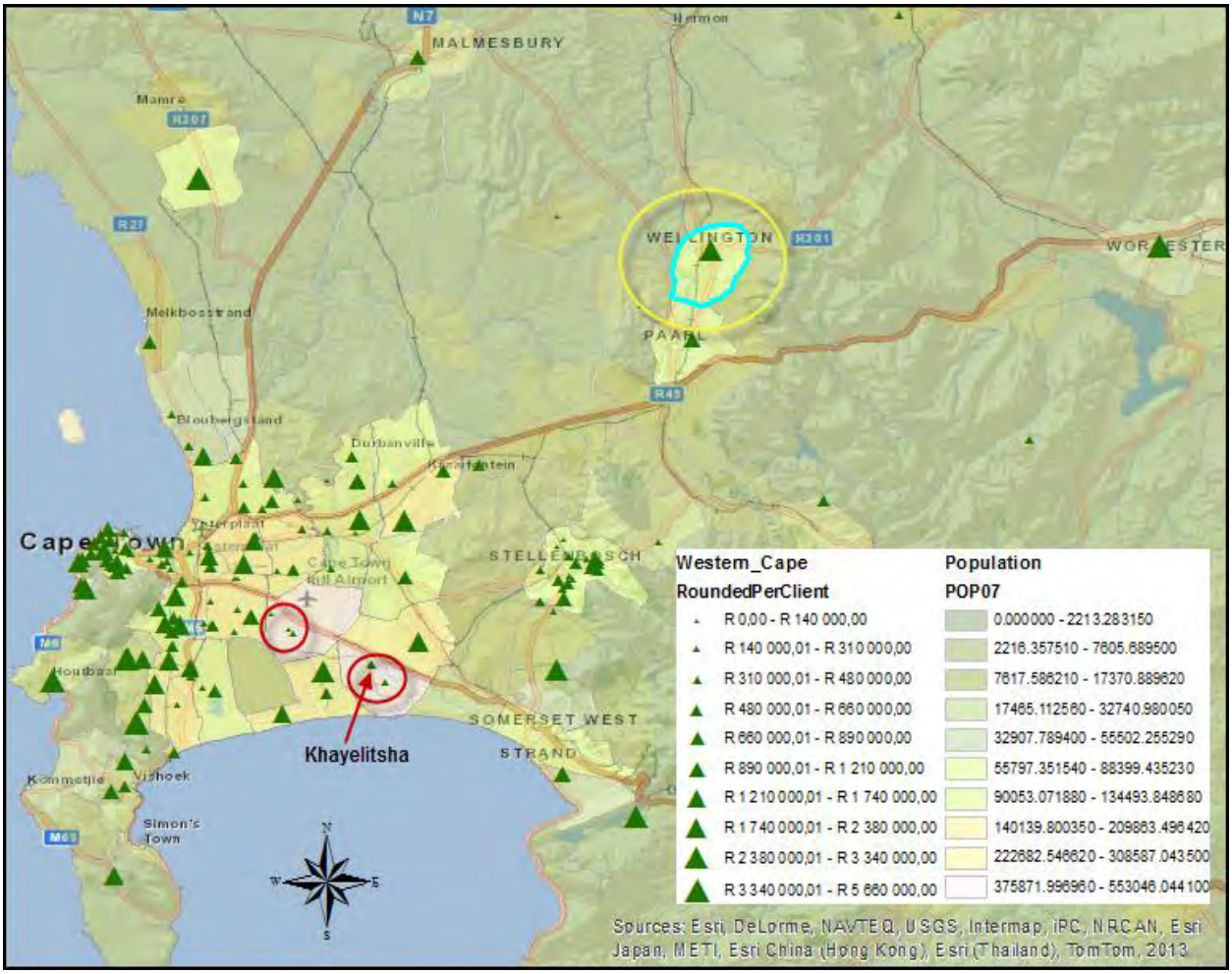


Figure 24: Western Cape discretionary average balance p/client with graduated population by suburb

If we look to the yellow circle, the blue area of Wellington is highlighted. In Wellington we can see that even when we have displayed the data by balance per client, the area is still significant in terms of average balance per client.

Another observation that can also be made is through comparison of balance to population density. The white areas highlighted by red circles are the two densely populated townships of Khayelitsha and

Guguletu. Both are established townships and a large portion of them are still considered to be informal settlements. The larger of the two triangles is Khayelitsha with an average balance between R310,000 and R480,000. This is another scenario that is worth investigating as it has the potential to be a lucrative emerging market.

5.1.1 Heat maps

Using the Getis-Ord Gi* statistic a heat map of spatial clusters was created by mapping the output of the Z scores and P scores. A high Z score and small p-value is illustrated by a cluster of high values or 'hot spots'. A low X score and small p-value is illustrated by a cluster of low values, 'low spots' or 'cold spots'.

The map below in Figure 25 shows a clear hot spot concentration in and around Cape Town, Western Cape and a clear cold spot concentration near Johannesburg, Gauteng. Although the population of Gauteng doubles that of the Western Cape, (Statistics South Africa, 2013) the corporate offices of the company being studied is based in the Western Cape which may have caused this.

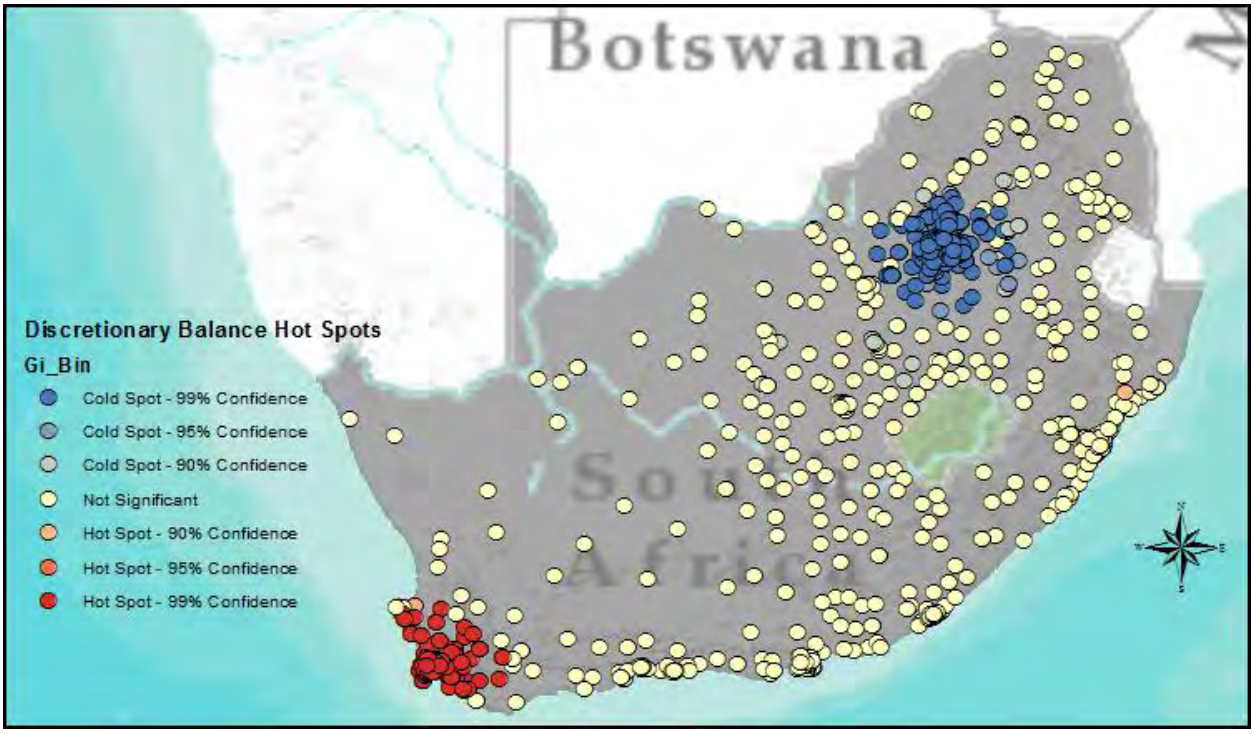


Figure 25: Discretionary balance hotspots; Total sum of area for client data based on residential address (Suburb, Town, Postcode combination)

When distribution is split per year, based on creation date, there is a different picture. In this case we took every 2 years from 2006 to 2012 and cluster mapped them. The Johannesburg, Gauteng area shows clear growth over the years, however, the Cape Town, Western Cape area has been stronger longer. This is shown below in Figure 26.

As the company deals with investments, it makes sense that the contributions over time have been stronger in the Western Cape, where the company started and the Johannesburg area is a growing market. From the numbers in 2008 and 2012 is clear that the Johannesburg market, in client numbers at least, is starting to surpass the Cape Town market.

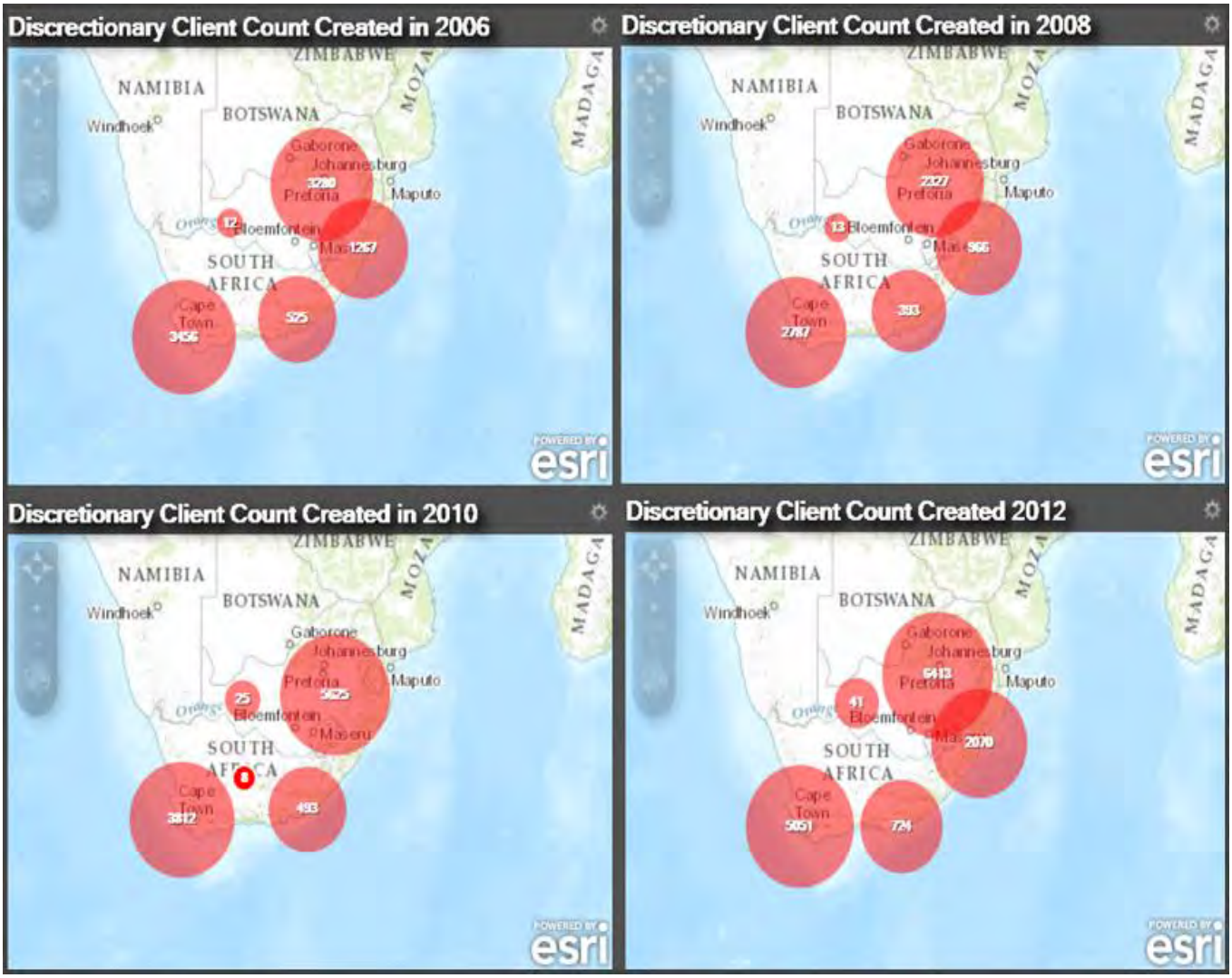


Figure 26: Distribution of new clients per year based on creation date*

***(Note: As addresses are not stored historically, it is possible that the client has moved since the account was created. These are address points based on where the client currently lives and when they were created)**

5.2 Publishing company map analysis

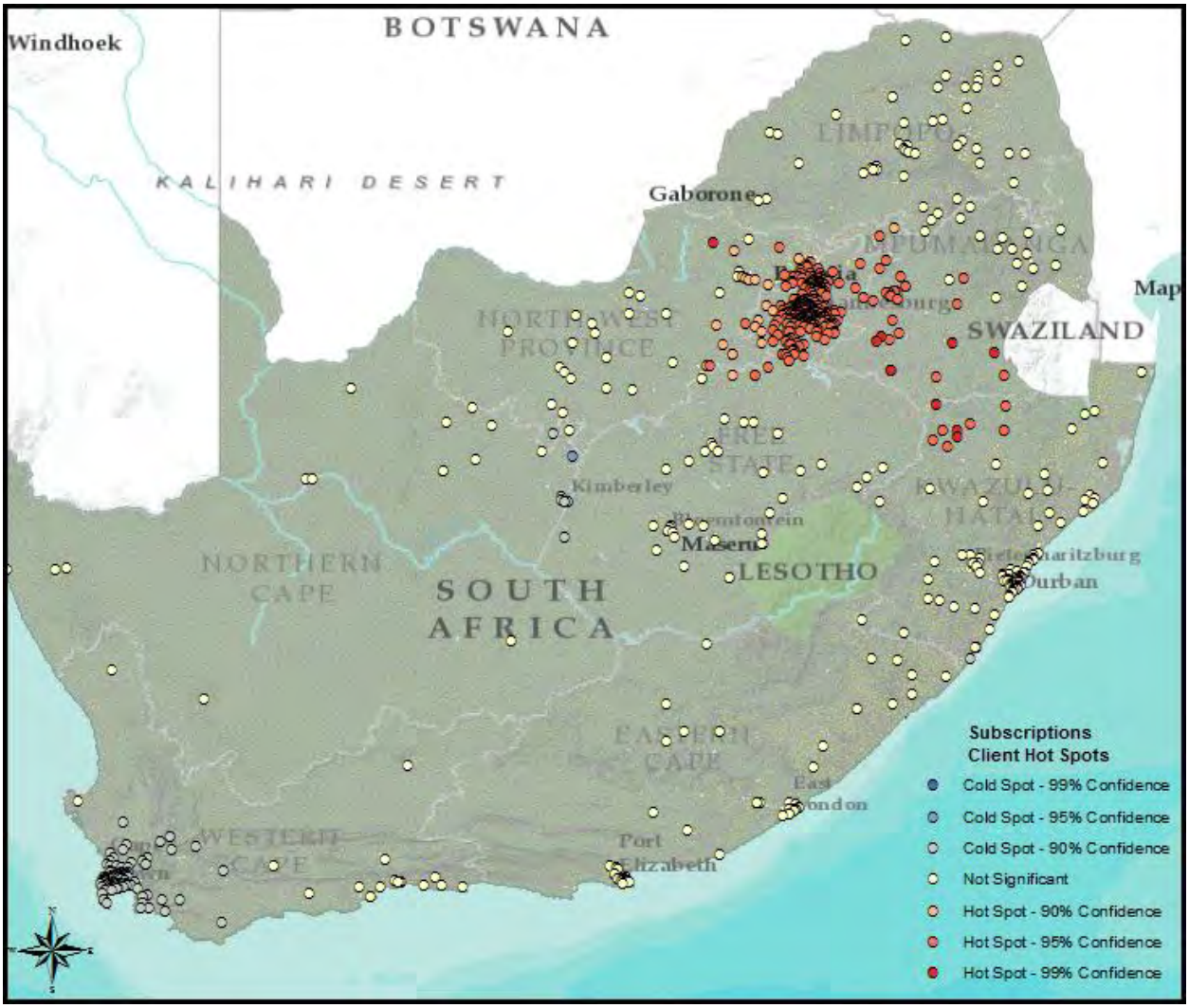


Figure 27: Subscription clients hot spots based on clients per suburb

The first analysis we ran for the publishing company was a hot spot analysis. The same Getis-Ord Gi* statistic was used to generate the heat map in Figure 27. The owner of the company felt that their client base might be concentrated in Johannesburg and the results confirm this. Unlike the financial company, the confidence level generally fell in the 95% range as the population studied is much smaller.

The proceeding analysis was based in Johannesburg as this is became the focus of the analytics.

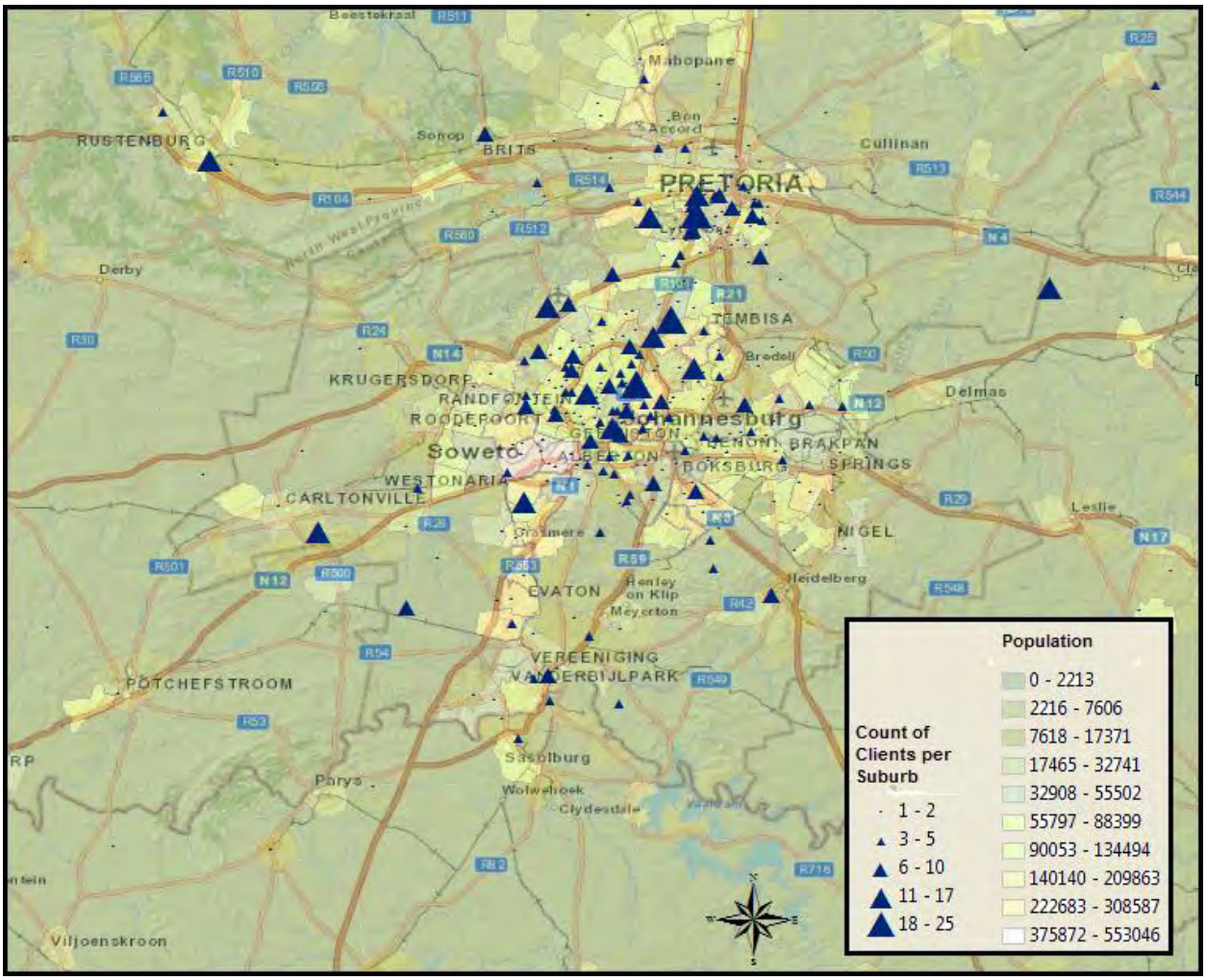


Figure 28: Count of clients by suburb in Johannesburg

Figure 28 shows the number of clients per suburb in Johannesburg. This map gives a good visualization of readership concentration in various suburbs. The owner had given the indication that Pretoria was an up-and-coming market for them. This can be seen by the concentration in various suburbs in Pretoria.

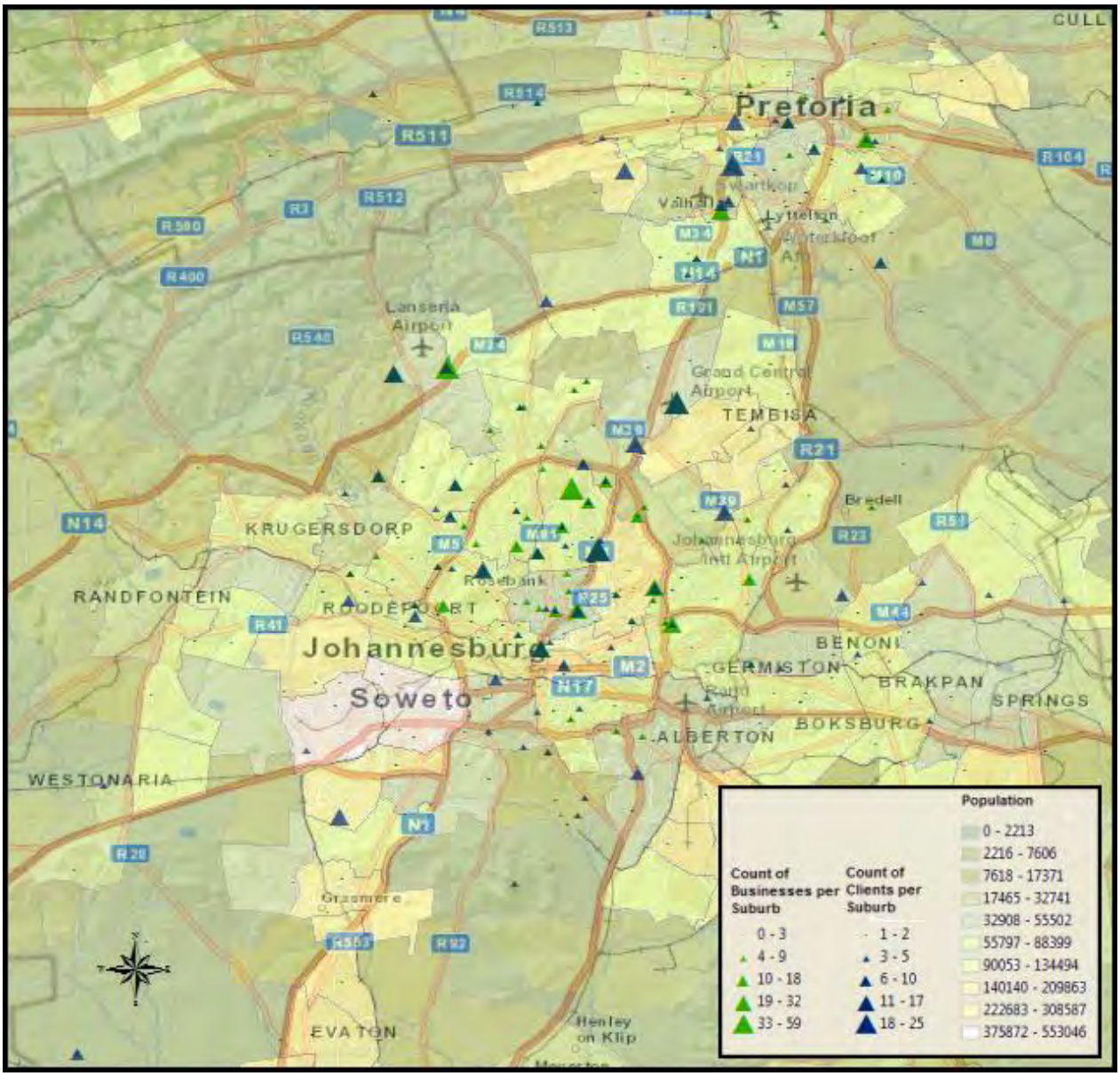


Figure 29: Client count compared to business count

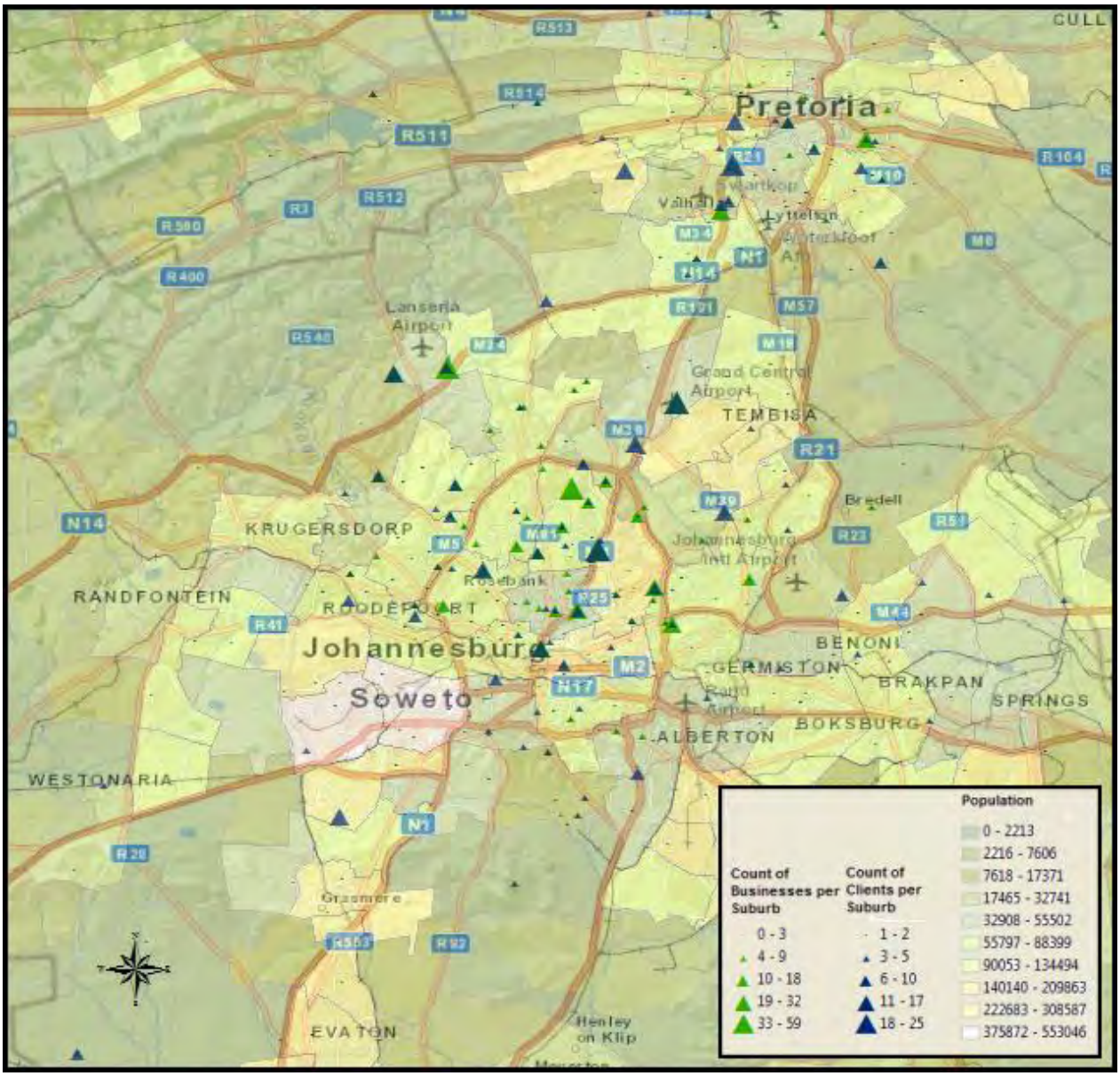


Figure 29 takes both business data and subscription data and maps them comparatively together. As most of the businesses are advertisers, this kind of map is used to show potential for client reach. It can also show areas that aren't being reached and give the company a reason to target these areas for potential advertisers.

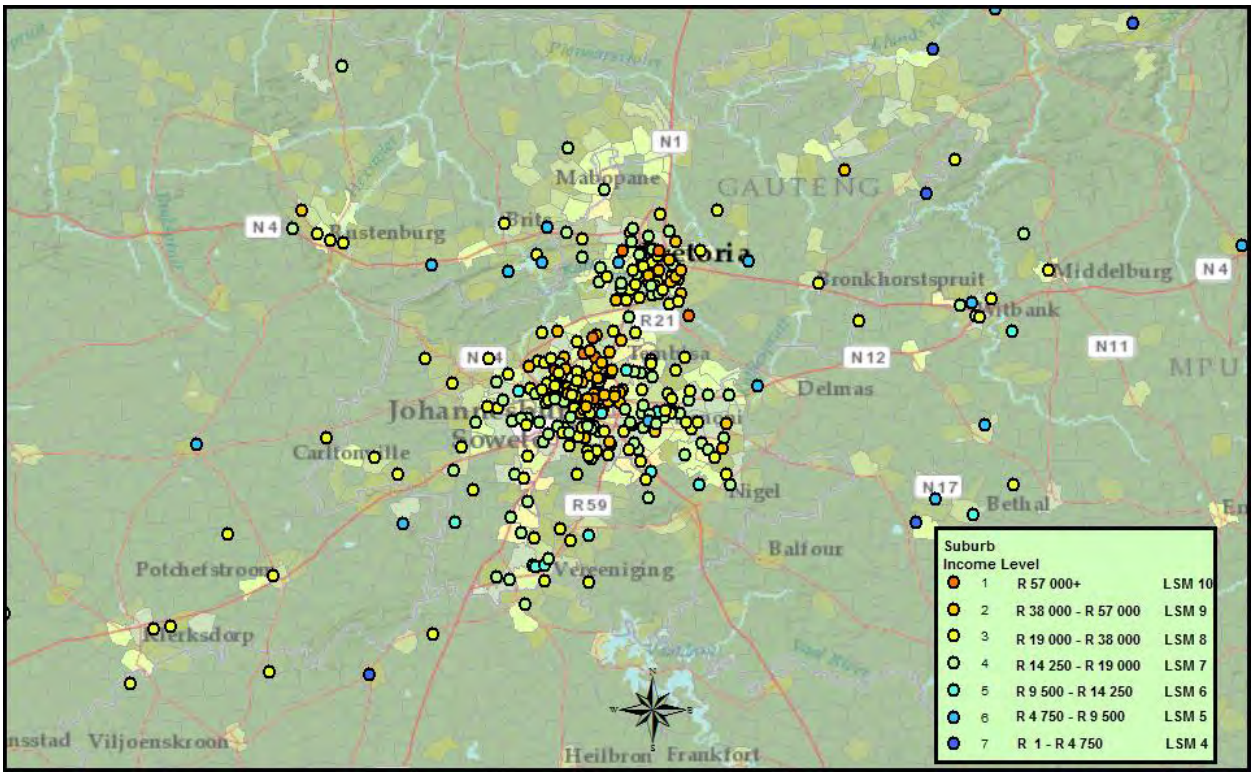


Figure 30: Income level for subscription clients in Johannesburg

Once location data has been confirmed and mapped, there are a multitude of other external and internal databases it can be combined with.

Figure 30 is an example of taking income data per suburb and joining it with the existing client data base. The result is an income level distribution level that can be used for one's clients. These can be displayed all as one, or separated and viewed individually. It gives the business the ability to target certain client groups, or solicit advertising that is in-line with a population group in the area.

The possibilities for analysis become numerous once the issues of poor data quality are overcome. The maps included here are only a small sample of what is possible once a confidence level is achieved in the address data as well as in the geocoding accuracy.

6 Discussion

We were able to address our issues with geocoding in South Africa within the components of the workflow. The first four of these issues can be categorised as address data quality issues. These were as follows:

a. Manually captured address that have incorrect or dirty data
b. Lack of capturing standards
c. Addresses coming from multiple sources
d. No singular source of address truth, i.e., postal delivery vs. utility service

We found South African address data to be inherently dirty regardless of the source of capture. This was compounded by a lack of standardization across organisations that are considered reliable sources for address verification documentation. We were able to solve most of these issues in the elementising and address cleaning steps of the workflow.

The next three issues can be categorised as issues with geocoders. These three are:

e. Geocoding databases that have out of date, incomplete, or incorrect information for South Africa
f. Geocoders can provide conflicting information
g. Commercial geocoding services can be very expensive.

We found that although there are many geocoding sources available for the South African market, there were often varying outputs for the same point. The commercial geocoders available were generally more accurate, but were also more expensive. We were able to address these issues in the geocoding and comparison steps of the workflow. By using multiple geocoders, we were able to identify conflicting information. This also enabled us to use free geocoders for the bulk of geocoding, and only required us to rely on a commercial solution for a smaller subset of data.

The final issue we addressed was with privacy concerns for address data. Described as:

h. Privacy concerns when releasing address data to be cleaned or geocoded

By aggregating our data, we were able keep analyse areas of interest, without disclosing identifying information in the specifics of the address. Aggregation also helped to address the issue of expense as the subset to be geocoded was reduced.

By applying our workflow to two different companies utilizing different types of address data, we were able to take steps to reduce the number of discrepancies in the data without subjecting the businesses to manual clean ups. We were then able to produce meaningful results at different scales by understanding the degree of acceptance when mapping the results.

As the market for spatial analytics grows, the tools and location accuracy will most likely improve. The current maturity and adoption levels in South Africa should not be a deterrent to organisations and individuals who are interested in mapping their data. We have also shown that cost does not have to be an additional barrier as two of the geocoding methods used were free of charge, and aggregation to suburb/town and postal code level makes geocoding by the popular commercial Esri product affordable

6.1 Possible end user solutions

A mock-up of a possible administration screen was created as shown in Figure 31. This could be another alternative to a report to help end users approve and rectify problematic addresses. The colour system could be applied to each address as well as an explanatory message, and users could accept, reject or amend the addresses. We presented the following administration screen as a mock up to the financial company. It was not implemented because manual intervention was not desired by the company.

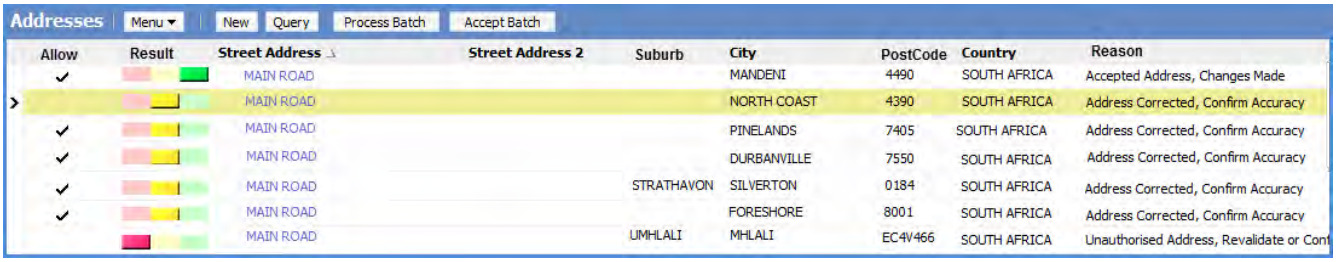


Figure 31: Mock-up of possible administration screen

Adding time to the analysis could also be very valuable. Unfortunately, the way in which both businesses store address data does not make this possible. This is quite common in many businesses. It is possible to see a client's start date; however, the address data generally updates and overwrites previous records when amended. In essence, one can see which clients were created in a certain year, but if the address was amended, there would be no idea how many times they might have moved. In both companies, address history is not kept. There is some form of change logging, but not detailed history on address changes. This plays a big role in the ability to enable time-based spatial analysis. If in the future either company decided they wanted to track client movement over time, it would be recommended to begin keeping history of the old and new addresses. Also, for future study, all new

addresses would need to be subject to standardisation and cleansing. This could be done at capture point or incrementally in batches after capture, but before analysis and geocoding begins.

7 Conclusion

We developed a six step workflow to prepare address data for spatial analytics. This consisted of address elementising, address cleansing, aggregation, geocoding with multiple sources, geocode comparison and finally an extraction of acceptable results. The workflow was successfully applied to two different business types with different data quality standards and different address types. While the software was customised specifically to each business' database, the workflow can be applied other businesses within South Africa who wish to use aggregated address data for analysis.

The address cleansing results showed that geocoding cost savings could be achieved by aggregating cleansed unique combinations of suburb, town, and postal code. The cleansing process greatly reduced the number of items that needed to be geocoded in both instances.

Geocoding with three sources resulted in the ability to confirm points to a determined distance of accuracy. If a client is willing to increase the distance in their degree of acceptance, the error rate would be reduced. Of the three methods used, the commercial product had the highest degree of accuracy. However, the free methods could be used exclusively with higher error rates. The comparison also showed a trend of a reduction of error rate that levels off after 22.2km (0.2^{o}). There was very little gain increasing the degree of acceptance beyond this point.

In conclusion, beneficial spatial analytics can be achieved for companies with less than perfect address data who are interested in protecting client privacy. A high rate of error can be compensated for by adopting the workflow, which pre-processes data and compares results from multiple geocoding methods before assigning geocodes along with accuracy indicators to identify where manual intervention is needed.

8 References

ArcGIS Online Credits Estimator. (2014, October). Retrieved September 9, 2013, from ArcGIS Online: http://www.esri.com/software/arcgis/arcgisonline/credits/estimator

BlueMM. (2007, January 06). BlueMM: Excel formula to calculate distance between 2 latitude, longitude (lat/lon) points (GPS positions). Retrieved Oct 14, 2013, from BlueMM: http://bluemm.blogspot.com/2007/01/excel-formula-to-calculate-distance.html

Breetzke, G. (2006). Geographical information systems (GIS) and policing in South Africa: a review. Policing: An International Journal of Police Strategies & Management, 29(4), 723-740.

Brown, J. R., Ivkovich, Z., Smith, P. A., & Weisbenner, a. S. (2004). The geography of stock market participation: The influence of communities and local firms. Cambridge: National Bureau of Economic Research.

Brownstein, J. S., Freifeld, C. C., Reis, B. Y., & Mandl, K. D. (2008, July 8). Surveillance Sans Frontières: Internet-Based Emerging Infectious Disease Intelligence and the HealthMap Project. PLoS Medicine, 5(7), 1019-1024.

Cairncross, F. (1997). The death of distance, how the communications revolution will change our lives. Cambridge, MA: Harvard Business School Press.

Clarke, K. C. (1996). On Epidemiology and Geographic Information Systems: A Review and Discussion of Future Directions. Emerging Infectious Diseases, 85-92.

Coetzee, S., & Cooper, A. (2007a). What is an address in South Africa? South African Journal of Science , 103(11-12), 449-458.

Coetzee, S., & Cooper, A. (2007b). Value of addresses to the economy, society and governance-a South African perspective. Pretoria: Department of Trade and Industry (dti) and AfriGIS.

Coetzee, S., & Cooper, A. (2008). Can the South African address standard (SANS 1883) work for small local municipalities? Free and Open Source Software for Geospatial Conference, (pp. 71-81). Cape Town.

Esri. (2007/2008). Cape Town's Emphasis on Systems Integration Exemplifies "Smart City" Goals. ArcNews, Winter.

Esri. (2012). ESRI Tapestry segmentation reference guide. Retrieved September 9, 2013, from Esri.com: http://www.esri.com/library/brochures/pdfs/tapestry-segmentation.pdf

Esri. (2013, December 20). ArcGIS REST API - Service Coverage. Retrieved December 23, 2013, from ArcGIS Resources: http://resources.arcgis.com/en/help/arcgis-rest-api/index.html#//02r300000018000000

Esri. (2014, September 15). About Esri | Company Profile - Info & Office Locations. Retrieved September 15, 2014, from About Esri: http://www.esri.com/about-esri

FDGC. (2011). Geospatial Platform Modernization Roadmap v4 Final. Washington D.C.: US Department of the Interior and Federal Geographic Data Committee.

Federal Geographic Data Committee. (2013, Dec 18). The Federal Geographic Data Committee. Retrieved Jan 21, 2014, from Federal Geographic Data Committee: www.fgdc.gov

Ferreira, J., Joao, P., & Martins, J. (2012). GIS for Crime Analysis_Geography for Predictive Models. Electronic Journal Information Systems Evaluation, 15(1), 36-49.

Folger, P. (2009). Geospatial Information and Geographic Information Systems (GIS): Current Issues and Future Challenges. Washington D.C. : Congressional Research Service.

Fotheringham, A. S., Brunsdon, C., & Charlton, M. (2000). Quantitative Geography: Perspectives on Spatial Data Analysis. London: SAGE.

Geiger, J. (2004). Data Quality Management: the most critical initiative you can implement. SUGI 29 (pp. 089-29). Boulder, CO: Intelligent Solutions, Inc.

Geonames.org. (2014, Jan 21). GeoNames. Retrieved Jan 21, 2014, from GeoNames: http://download.geonames.org/export/zip/

Gilpin, L. (2014, August 26). How an algorithm detected the Ebola outbreak a week early, and what it could do next - TechRepublic. Retrieved August 26, 2014, from TechRepublic: http://www.techrepublic.com/article/how-an-algorithm-detected-the-ebola-outbreak-a-week-early-and-what-it-could-do-next/

Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, L., & Brilliant, M. S. (2009, February 19). Detecting influenza epidemics using search engine query data. Nature, 457, 1012-1014.

Goldberg, D., Wilson, J., Knoblock, C., Ritz, B., & Cockburn, M. (2008). An effective and efficient approach for manually improving geocoded data. International Journal of Health Geographics 7 no 1, 7(1), 60.

Goodchild, M. F., & Janelle, D. G. (2010). Toward Critical Spatial Thinking in the Social Sciences and Humanities. Santa Barbara, California: University of California, Santa Barbara. Department of Geography and the Center for Spatial Studies.

Google. (2013, November 25). Google Maps/Google Earth APIs Terms of Service - Google Maps API — Google Developers. Retrieved Dec 16, 2013, from Google Developers: https://developers.google.com/maps/terms#section_10_12

Google. (2014, November 13). Johannesburg - Google Maps. Retrieved November 13, 2014, from Google Maps: https://www.google.co.za/maps/place/Johannesburg,+2198/@-26.1577651,28.0701644,13z/data=!4m2!3m1!1s0x1e950db89cb0585d:0x6a566425b8dfe622

Graham, J. (2015, January 15). Uber steering data to Boston. Retrieved January 14, 2015, from Boston Herald: http://www.bostonherald.com/business/business_markets/2015/01/uber_steering_data_to_boston

Gregory, I. N. (2005). A place in history: A guide to using GIS in historical research. 2nd Edition. Belfast: Ian Gregory.

Grimshaw, D. J. (2001). Harnessing the power of geographical knowledge: the potential for data integration in an SME. International Journal of Information Managment, 21(3), 183-191.

HealthMap. (2014, March 14). Contagious Disease Surveillance | Virus Awareness | Ebola Map | HealthMap. Retrieved August 26, 2014, from HealthMap: healthmap.org/ebola/#timeline

Hill, L. L., & Zheng, Q. (1998). Indirect geospatial referencing through place names in the digital library: Alexandria digital library experience with the developing and implementing gazetteers: Analysis and preliminary evaluation of the classical digital library model. Proceedings of the Annual Meeting-American Society for Information Science. Vol. 36, pp. 57-69. Information Today.

Hongjian, Z., Fei, X., & Chunxiu, W. (2011). Research on the Application of GIS Technology in Spatial Mode Customer Relationship Management System. Journal on Innovation and Sustainability, 2(2), 34-37.

Johnson, R. (2000). GIS Technology for Disasters and Emergency Management . Redlands, CA: Esri.

Johnson, S., & Yespolov, T. (2014). The Role of Extension Systems and GIS Technology in Formation and Predicting Global Agricultural Policy: Precision Agriculture is Coming Fast. International Scientific Electronic Journal Earth Bioresources and Life Quality Founded by National University of Life and Environmental Sciences of Ukraine (NUBiP of Ukraine) and Global Consortium of Higher Education and Research for Agriculture (GCHERA)(4).

Jonckie. (2013, January 22). Digital mapping: hidden secret of SA business. Retrieved Nov 27, 2013, from Insurance Chat: http://www.insurancechat.co.za/2013-01/digital-mapping-hidden-secret-of-sa-business/

Jones, K., & Simmons, J. (1987). Location, location, location. Toronto: Methuen Publications.

Khatri, V., & Brown, C. V. (2010). Designing data governance. Communications of the ACM 53.1, 148-152.

Kim, K. W., Choi, B.-J., Hong, E.-K., Kim, S.-K., & Lee, D. (2003). A taxonomy of dirty data. Data mining and knowledge discovery 7, no. 1 , 81-99.

Lombaard, M. (2010). South African Postcode Geography. Retrieved Nov 27, 2013, from Statistics South Africa: http://mapserver2.statssa.gov.za/geographywebsite/africaGIS.html

Maguire, D. J. (1991). An Overview and Defiinition of GIS. Geographical Information Systems: Principals and Applications, 9-20.

Maletic, J. I., & Marcus, a. A. (2000). Data Cleansing: Beyond Integrity Analysis. IQ, 200-209.

Marshall, P. (2013, Oct 18). GIS becomes indispensable for managing agriculture -- GCN. Retrieved Jan 21, 2014, from GCN.com: gcn.com/Articles/2013/10/18/USDA-GIS.aspx?p=1

Meankaew, P., Kaewkungwal, J., Khamsiriwatchara, A., Khunthong, P., Singhasivanon, P., & Satimai, W. (2010). Application of mobile-technology for disease and treatment monitoring of malaria in the "Better Border Healthcare Programme. Malaria Journal, 9(1), 1-14.

Microsoft. (2015, January 23). Location Data. Retrieved January 23, 2015, from Microsoft Developer Network - Bing Maps REST Services: https://msdn.microsoft.com/en-us/library/ff701725.aspx

Nassar, A., Blackburn, A., & Whyatt, D. (2012). Quantifying Urban Growth in Dubai Emirate: A Geoinformatics Approach. Proceedings of the GIS Research UK 20th Annual Conference (pp. 375-382). Lancaster: Lancaster University.

Neves, J. N., & Camara, A. (1999). Virtual Environments and GIS. In Geographical Information Systems (pp. 557-65).

Newsom, S. W. (2006). Pioneers in infection control: John Snow, Henry Whitehead, the Broad Street pump, and the beginnings of geographical epidemiology. Journal of Hospital Infection, 64(3), 210-216.

Ortmann, J., Limbu, M., Wang, D., & Kauppinen, T. (2011). Crowdsourcing Linked Open Data for Disaster Management. Proceedings of the Terra Cognita Workshop on Foundations, Technologies and Applications of the Geospatial Web in conjunction with the ISWC (pp. 11-22). Muenster, Germany: Institute for Geoinformatics, University of Muenster, Germany.

Parish, J. S. (2009). Geographic Information Systems and Spatial Analysis of Market Segmentation for Community Banks. Greensboro: University of North Carolina.

Prasad, K., & Ramakrishna, S. (2011). Role of Geographical Information Systems on Intersecting Functionalities of Banking and Insurance Sectors. International Journal of Computer Science and Emerging Technologies, 2(3), 383-388.

Rahm, E., & Do, H. H. (2000). Data Cleaning: Problems and Current Approaches. IEEE Data Eng. Bull. 23.4, 3-13.

Republic of South Africa. (2013, Nov 26). Act No. 4 of 2013: Protection of Personal Information Act, 2013. Vol. 581(No. 37067).

Rossouw, P. (2009, January 26). International address standard compliance for SA Post Office - EE Publishers. Retrieved Jan 22, 2015, from EE Publishers: www.ee.co.za/article/international-address-standard-compliance-for-sa-post-office.html

Rossow, P. S. (2008). Addressing and the new postcode system. PositionIT, 59-65.

SAARF. (2012). Do-It-Yourself LSM® Classification. Retrieved Feb 06, 2014, from SAARF: http://www.saarf.co.za/LSM/lsm-diy.asp

SAARF. (2012). SAARF Living Standards Measure. Retrieved Jan 20, 2014, from SAARF: http://www.saarf.co.za/LSM/lsms.asp

SAARF. (2012, Feb 28). SAARF Segmentation Tools. Retrieved Feb 06, 2014, from SAARF: http://www.saarf.co.za/lsm-presentations/2012/LSM Presentation - February 2012.pdf

Sampath, T. (2012, July 20). Cluster Heat Maps Could Help Insurers Make Sense of Big Data. Retrieved January 28, 2015, from Insurance Networking: http://www.insurancenetworking.com/blogs/cluster-data-heat-map-30735-1.html

Shacklett, M. (2014, Feb 11). Data quality: The ugly duckling of big data? Retrieved Feb 12, 2014, from TechRepublic: http://www.techrepublic.com/article/data-quality-the-ugly-duckling-of-big-data/#ftag=RSS56d97e7

Skupkin, A., & Fabrikant, S. I. (2003). Spatialization methods: A cartographic research agenda for non-geographic information visualization. Cartography and Geographic Information Science, 99-119.

Statistics South Africa. (2013). Mid-year population estimates 2013. Pretoria: Stats SA.

Swift, J., Goldberg, D., & Wilson, J. (2008). Geocoding best practices: review of eight commonly used geocoding systems. Los Angeles, CA: University of Southern California GIS Research Laboratory.

Thompson, S. (2011). The Geography of Geocoding: Impecations on Precision and Accuracy . (p. 54). Redlands, CA: Esri.

Trillium Software. (2009). An Introduction to Geocoding: What every enterprise needs to know about precision location data. Billerica, MA: Harte-Hanks Trillium Software.

Tuan, Y.-F. (1979). Space and Place: Humanistic Perspective. University of Minnesota.

U.S. Department of Agriculture. (2014, Jan 21). CropScape General Information. Retrieved Jan 21, 2014, from National Agricultural Statistics Service: http://www.nass.usda.gov/research/Cropland/sarsfaqs2.html#Section1_1.0

US Office of Management & Budget. (2010). Budget of the US Government, Fiscal Year 2011. Washington D.C.: US Governement Printing Office.

Appendix LSM® Classification Scoring

Table 20: LSM® Classification Scoring (SAARF, 2012)

Question	Answer	Weight if True
TV set	True/False	0.120814
Swimming pool	True/False	0.166031
DVD player/ Blu Ray Player	True/False	0.09607
Pay TV (M-Net/DStv/Top TV) subscription	True/False	0.12736
Air conditioner (exl. fans)	True/False	0.178044
Computer /Desktop/ Laptop	True/False	0.311118
Vacuum cleaner/floor polisher	True/False	0.164736
Dishwashing machine	True/False	0.212562
Washing machine	True/False	0.149009
Tumble dryer	True/False	0.166056
Home telephone (excluding a cell)	True/False	0.104531
Deep freezer –free standing	True/False	0.116673
Refrigerator or combined fridge/freezer	True/False	0.134133
Electric stove	True/False	0.16322
Microwave oven	True/False	0.126409
Built-in kitchen sink	True/False	0.132822
Home security service	True/False	0.151623
3 or more cell phones in household	True/False	0.184676
2 cell phones in household	True/False	0.124007
Home theatre system	True/False	0.096072
Tap water in house/on plot	True/False	0.123015
Hot running water from a geyser	True/False	0.185224
Flush toilet in/outside house	True/False	0.113306
There is a motor vehicle in our household	True/False	0.16731
I am a metropolitan dweller	True/False	0.079321
I live in a house, cluster or town house	True/False	0.113907
I live in a rural area outside Gauteng and the Western Cape	True/False	-0.129361
There are no radios, or only one radio (excluding car radios) in my household	True/False	-0.245001
There are no domestic workers or household helpers in household (incl. both live-in & part time domestics and gardeners)	True/False	-0.30132

The totals are added up and the constant of 0.810519 is subtracted from the total score giving the LSM® group. The scores can be seen in Table 21.

Table 21: LSM® Scoring (SAARF, 2012)

Score	LSM®
Less than -1.390140	1
Between -1.390140 to -1.242001	2
Between -1.242000 to -1.011801	3
Between -1.011800 to -0.691001	4
Between -0.691000 to -0.278001	5
Between -0.278000 to 0.381999	6
Between 0.382000 to 0.800999	7
Between 0.801000 to 1.168999	8
Between 1.169000 to 1.744999	9
More than 1.744999	10