UNIVERSITY OF THE WITWATERSRAND, JOHANNESBURG

**Faculty of Science**

**SCHOOL OF GEOGRAPHY, ARCHAEOLOGY AND ENVIRONMENTAL STUDIES**

**MSC GEOGRAPHICAL INFORMATION SYSTEMS (GIS) AND REMOTE SENSING**

2021

NAME : BETH PULANE CRANKSHAW

PERSON NUMBER : 819865

# RESEARCH PROJECT

# DASYMETRIC ESTIMATION OF TUBERCULOSIS INCIDENCE RATES AND FACILITY-LEVEL HOT SPOT ANALYSIS IN JOHANNESBURG

***SUPERVISORS:***

***Prof Elhadi Adam:*** *University of the Witwatersrand, School of Geography, Archaeology and Environmental Studies, Johannesburg.*

***Dr Harry Moultrie:*** *Centre for Tuberculosis, National Institute for Communicable Diseases (NICD), Johannesburg.*

# Declaration

I declare that this research report is my own, unaided work. It is being submitted to the Faculty of Science, University of the Witwatersrand, Johannesburg, in partial fulfilment of the requirements for the degree of Master of Science in Geographic Information Systems and Remote Sensing. It has not been submitted before for any degree or examination at any other University.

(Beth Pulane Crankshaw)

Signed on the 16th day of September 2021 in Johannesburg.

# Abstract

South Africa is listed by the World Health Organisation (WHO) (2020) as one of the top 20 global contributors to absolute case numbers for TB burden and is among the 14 countries with the highest burden of HIV associated TB, and multi-drug resistant TB (MDR-TB). Understanding the geographical disparities in TB distribution provides a foundation upon which informed decisions around targeted interventions can be made. Determining TB disease incidence at a fine scale is therefore critical in the field of public health where strategic interventions, e.g. targeting disease hotspots, can result in exponential reductions in transmission and incidence.

However, routine geographic analysis of the distribution of TB are not at an appropriate scale for such knowledge generation due to the format in which population and health data are available. This project therefore explores an alternative dasymetric areal interpolation technique for determining incidence of microbiologically-confirmed pulmonary TB (mPTB) from facility-level data where catchment populations are unknown.

A series of hot spot detection techniques are then applied to the dasymetric mPTB output to identify areas of high incidence. Standard approaches are explored including K-Nearest Neighbour (KNN), inverse distance and fixed distance weighting techniques, as well as the cluster morphology AMOEBA method.

# Contents

Declaration ..... i

Abstract ..... ii

List of Figures ..... vi

List of Tables ..... vii

List of Acronyms ..... viii

Chapter 1: Introduction ..... 1

1.1 Background ..... 1

1.2 Research Problem ..... 2

1.3 Aims and Objectives ..... 3

1.4 Study significance ..... 3

1.5 Ethical considerations ..... 5

1.6 Research report outlines ..... 5

Chapter 2: Literature review ..... 7

2.1 Introduction ..... 7

2.2 Choropleth mapping ..... 7

2.3 Spatial interpolation ..... 9

2.3.1 Dasymetric mapping ..... 12

2.4 Comparison of choropleth and dasymetric methods ..... 14

2.5 Spatial Autocorrelation ..... 15

2.5.1 Hot spot analysis ..... 16

Chapter 3: Methodology ..... 18

3.1 Introduction ..... 18

3.2 Study Area ..... 19

3.3 Data acquisition and preparation ..... 24

3.3.1 Microbiologically-confirmed Pulmonary TB (mPTB) data ..... 25

3.3.2 Population data ..... 25

3.3.3 Land-use dataset ..... 26

3.4 Research assumptions ..... 28

3.4.1 Study population ..... 28

3.4.2 Utilisation of nearest facility ..... 29

3.5 Choropleth method ..... 31

3.6 Dasymetric method ..... 32

3.7 Comparing the two methods ..... 36

3.8 Spatial autocorrelation ..... 37

3.8.1 The AMOEBA method ..... 39

Chapter 4: Results ..... 43

4.1 mPTB incidence rate outputs ..... 43

4.1.1 Choropleth output ..... 43

4.1.2 Dasymetric Output ..... 44

4.1.3 Variance and standard deviation ..... 48

4.2 Spatial Autocorrelation ..... 49

4.2.1 Global results ..... 49

4.2.2 Local results ..... 50

4.2.3 AMOEBA ..... 51

Chapter 5: Discussion ..... 53

Chapter 6: Conclusion and Recommendations ..... 56

4.3 Limitations of the study ..... 57

4.4 Recommendations and future research ..... 58

References ..... 59

Addenda ..... 65

1. SANLC Classification Map ..... 65

2. AMOEBA overlayed with residential pixels ..... 66

3. Ethical Clearance Certificate ..... 67

# List of Figures

Figure 1: Johannesburg's health sub-districts ..... 19

Figure 2: Median individual income by EA for 2011 ..... 21

Figure 3: Number of individuals living below the poverty line per EA for Johannesburg ..... 22

Figure 4: Population density (per square kilometre) estimates for 2017 ..... 23

Figure 5: Example of discrepancies between SBC and SANLC populated area pixels ..... 27

Figure 6: Creating Thiessen polygons from points (ESRI, 2016) ..... 32

Figure 7: Gauteng facility catchment areas (1), filtered to those intersecting with the Johannesburg boundary (2) ..... 33

Figure 8: Binary pixelated map of populated areas, derived using SANLC and SBC LULC data ..... 36

Figure 9: The AMOEBA procedure ..... 40

Figure 10: Health sub-district TB rates per 100,000 population for the two-year period 2018-2019 ..... 43

Figure 11: Dasymetrically-derived mPTB rate (per 100,000 population) for Johannesburg 2018-2019 ..... 45

Figure 12: Dasymetric map of mPTB rate (per 100,000 population) for Johannesburg 2018-2019 ..... 47

Figure 13: Hot spot analysis results for (1) KNN, (2) Distance threshold, and (3) Inverse distance & Inverse Distance Squared ..... 50

Figure 14: mPTB incidence clusters determined using the (1) AMOEBA and (2) AMOEBA(+) algorithms .....51

# List of Tables

Table 3.1: Summary of datasets utilised .....24

Table 4.1: Mean values from a choropleth mapping process, and standard deviation of dasymetric values from this mean.....48

Table 4.2: Result of Global Moran's I test for spatial autocorrelation .....49

# List of Acronyms

| | |
|---|---|
| CoJ | City of Johannesburg |
| HIV | Human Immunodeficiency Virus |
| LULC | Land use/land cover |
| MDR-TB | Multiple drug-resistant tuberculosis |
| NDoH | National Department of Health |
| NICD | National Institute for Communicable Diseases |
| NHLS | National Health Laboratory Service |
| SANLC | South African National Land Cover |
| SBC | Satellite Building Count/ SPOT Building Count |
| TB | Tuberculosis |
| WHO | World Health Organisation |
| XDR-TB | Extensively drug-resistant tuberculosis |

# List of Abbreviations

| | |
|---|---|
| CI | Confidence Interval |

# Chapter 1: Introduction

## 1.1 Background

Tuberculosis (TB) was declared a notifiable disease in South Africa's Cape and Natal regions in 1904, and nationally in 1919, obligating registration of all cases (Churchyard et al., 2014). Moving forward to the early 1990s, activation and spread of the disease was fuelled by the HIV epidemic, and from 2006 was further exacerbated by a growing burden of multiple drug-resistant (MDR) TB and extensively drug-resistant (XDR) TB (ibid.). Cases have, however, declined since 2008 due to upscaling antiretroviral treatment (The First National TB Prevalence Survey, 2021) and interventions championed by the National TB Programme (NTP), established in 1994 (Department of Health, 2021).

Despite these efforts, South Africa is still listed by the WHO (2020) as one of the top 20 global contributors to absolute case numbers for TB burden and is among the 14 countries with the highest burden of HIV associated TB, and multi-drug resistant TB (MDR-TB). The recently released 'The First National TB Prevalence Survey' estimated that TB prevalence (for all ages and all forms of TB) in South Africa in 2018 was 737 (95% CI 580-890) per 100,000 population; substantially more than the reported WHO estimate of 301 (CI 215-400) for the same year (WHO, 2019). South Africa has aligned itself with the WHO End TB Strategy, adopted at the World Health Assembly in 2014, which targets a 90% reduction in patients suffering from TB, and a 95% reduction in deaths from TB by 2035 (Loveday, Mzobe, Pillay, & Barron, 2019).

The above-mentioned prevalence survey reported that the highest proportions of TB were in males (with a prevalence almost 1.6 times that of females), and in populations that were HIV-negative (71.2%). This latter finding contrasts with corresponding National TB programme (NTP) data where 58% of individuals already on treatment prior to enrolment in the TB survey

were HIV-positive, meaning only 42% were HIV-negative. HIV-positive participants in the study's sample exhibited more symptoms than those who were HIV-negative (a high proportion of which reported no symptoms, yet had bacteriological confirmation of TB) and therefore were more likely to be diagnosed. This introduces the rising concern of sub-clinical, asymptomatic TB believed to be an underestimated contributor to TB burden (The First National TB Prevalence Survey, 2021). Detection requires more active, sensitive screening solutions in areas of high prevalence.

## 1.2 Research Problem

Detailed studies of TB distribution have made use of clinical case data acquired from the ETR.Net (pre-2019), later migrated to TIER.Net (2019 onwards), data repositories (Obaromi, Ndege, & Yongsong, 2019; Wekesa et al., 2020). However, this data is not routinely available and where local access to a spatial and temporal subset of the data is permitted, the individual address data is often not geocodable due to ambiguous residential addresses often located in townships, informal settlements and rural areas (Coetzee, Cooper & Katumba, 2020).

In the public health sector, automated, real-time laboratory test data is provided at the facility level, and similarly to literature, this is most commonly monitored and analysed at the sub-district level (McIntosh et al., 2018). Despite the known limitations of this aggregated scale of analysis, it is necessitated – even where raw, patient-level data is available – due to difficulties in coding unformatted address data in south Africa and missing patient residential address data in these national datasets (Mokgalaka, 2015). Sub-district level data may be adequate for monitoring TB on a national level, but does not provide the necessary detail for meaningful data analysis and planning of strategic interventions, especially in cities. Metropolitan areas are characterised by a diversity of healthcare needs with higher population densities and migration patterns contributing to the spread of airborne diseases such as TB (Lurie & Williams, 2014).

The availability of facility-level data combined with advances in computing power and geospatial software development presents an opportunity to determine disease distribution at a fine scale, i.e. by facility catchment area. However, this scale also introduces challenges for spatial analysis and interpretation given the intricacies of determining catchment extents. Catchment area and accessibility analyses are widely researched topics within the field of Geographic Information Systems (GIS), with numerous proposed approaches offered with varying levels of complexity.

Determining disease incidence also requires robust population estimates, representing the total count of people potentially affected by a disease. The finest scale official census data available in South Africa is that of enumeration area (EA) polygons, but these are administrative (not functional) boundaries that do not resemble facility catchment population estimates.

### 1.3 Aims and Objectives

The overall aim of this research project is to determine whether spatial interpolation techniques offer a more representative depiction of TB incidence in a hot spot detection application.

Given the above aim, the following objectives were to:

1. Determine TB incidence using a conventional choropleth mapping approach
2. Determine TB incidence using a dasymetric mapping technique
3. Compare the TB incidence outputs of methods in objectives (1) and (2)
4. Identify TB incidence hot spots using the preferred, most representative method identified in objective (3)

### 1.4 Study significance

Recently, dCXR assisted by computer-aided detection (CAD) technologies has been recommended as a front-line screening tool for detecting asymptomatic TB which is estimated to account for >154,000 undiagnosed cases in South Africa each year (Department of

Health, 2021). The prevalence survey estimated that 390 000 people had TB in 2018, with 235 652 notified cases, while 154 348 were "missed TB patients" or undetected. The latest versions of the CAD component – a machine learning approach to detection – offer detection accuracies equal to expert human readers, and these results can be processed in under two minutes at a lower average cost per subject screened (Murphy et al., 2020).

The WHO has endorsed dCXR usage in its latest *Rapid communication on systematic screening for tuberculosis* (2020) and *Consolidated Guidelines on Tuberculosis: Systematic Screening for the Tuberculosis Disease* (2021), necessitating South Africa's consideration of a planned roll out. However, there are not enough resources available to implement dCXR in every clinic (there are over 4300 health facilities reporting TB cases throughout the country, with around 130 in Johannesburg alone), necessitating strategic placements of this intervention (H.Moultrie, personal communication, February 24, 2021). There are three potential approaches to determining optimal locations for maximal impact/ detection (ibid.), targeting:

1) Facilities with the highest burden
2) Areas with the highest positivity
3) Areas with the highest incidence

Option (1) considers positive counts per facility, (2) considers the number of positive cases per population that gets tested, and (3) considers the number of new positive cases over the total population. The third option provides the most contextual representation of burden as it takes into account the entire underlying population, regardless of the extent of testing. Spatially, this approach offers an intuitive statistical surface from which to determine hot spots, as seen in existing literature (Dowdy, Golub, Chaisson, & Saraceni, 2012; Ibrahim, Hamisu, & Lawal, 2015).

## 1.5 Ethical considerations

Ethical clearance for this research was attained from the University of the Witwatersrand School of Geography, Archaeology and Environmental Studies (GAES) (see Addendum 3).

Tuberculosis laboratory data was acquired in an aggregated format at facility level, ensuring no personal details were disclosed to the researcher; avoiding all potential anonymity and confidentiality risks.

## 1.6 Research report outlines

**Chapter 1**: Introduction

This chapter has provided a background into the dissertation topic, has set out the research aims and objectives, and explained the significance of the study and its findings.

**Chapter 2**: Literature review

Chapter 2 will discuss the literature on key concepts and available techniques for deriving accurate population surfaces and detecting areas of spatial autocorrelation.

**Chapter 3**: Methodology

Chapter 3 introduces the study area and presents the data to be utilised for the various methods to be performed. This includes an explanation of the necessary data preparation processes and lists the limitations of the data. Two approaches to presenting mPTB are then presented as well as a method for comparison thereof. Methods for detecting spatial autocorrelation and there discussed, including a cluster morphology technique, AMOEBA.

**Chapter 4**: Results

In this chapter, the main findings of the analysis are presented, with limited discussion thereof.

**Chapter 5**: Discussion

A more in-depth discussion of the findings is presented here, with specific reference to the aims and objectives of the research, and correlations with existing literature.

**Chapter 6**: Conclusions and recommendations

This chapter draws conclusions about the study based on the identified aims and objectives. Recommendations are provided based on limitations of the study and the potential for further research given the findings and procedural lessons learnt.

# Chapter 2: Literature review

## 2.1 Introduction

Human population data often form the denominator for various spatial analyses in the field of epidemiology and public health (Hay, Noor, Nelson, & Tatem, 2005), one such popular application being that of defining populations at risk of disease. The accuracy of any consequent analytical surfaces is therefore in part determined by the quality of the input population data. Inaccuracies in this underlying data introduce inherent errors – those which are carried in the data before the data are entered into analyses or models (Bian & Butler, 1999) – often induced during aggregation.

Frequently this aggregation is carried out using arbitrary, administrative areal units e.g. census EA's. Where population data boundaries are assumed to coincide with the phenomenon to be explored, analyses are straightforward. This scenario is presented below through a discussion of choropleth mapping (section 2.2). However, where these boundaries do not align, this poses a challenge for integrating datasets. Spatial interpolation techniques provide one potential solutions for transferring data from source to target units (discussed in section 2.3), one such approach being that of dasymetric mapping. This technique utilises ancillary land use/land cover (LULC) data to transfer data to populated areas of the study site (section 2.3.1).

A useful point of departure in analysing any spatial distribution of a disease is to assess the degree of spatial autocorrelation of the incidence rates (Rushton, 2003). Therefore, application of global and local test statistics for assessing spatial autocorrelation of the TB rate outputs are discussed (section 2.5).

## 2.2 Choropleth mapping

The development of rate maps has only been possible in cartography since statistics provided methods to treat quantitative data as linear arithmetic (Barrozo, Pérez-Machado, Small, &

Cabral-Miranda, 2016). The most common graphic portrayal of statistical surfaces, choropleth mapping, also known as statistical mapping, has been widely applied to represent rates due to its consistency in depicting relative data (Richards et al., 2010). Choropleth maps "assign a shading or color to geographic areas (defined by their boundaries) as a means of representing some measure of centrality with respect to a given characteristic" (Gale & Halperin, 1982, p 330). This measure of centrality is a 'representative value' used to characterise data in the same class (Mu & Tong, 2018); a concept to be utilised later in this study. As a thematic mapping technique, the aim is to portray spatial variations and relationships between various kinds of phenomena (Gale & Halperin, 1982).

In the fields of human geography and epidemiology, this usually requires some consideration of population density as part of calculating disease incidence, and this census enumeration data is traditionally provided at some administrative level(s) (Steiner & Paulus, 2007). Unfortunately, administrative boundaries are often political, and do not adequately depict population distribution or the phenomenon being studied (Openshaw, 1984). This may be an adequate representation of information when applied to small areas where an assumption of homogeneity over the area's surface risks limited loss of information (ibid.). However, Talbot et al. (2000) posit that under conditions of known large spatial variability in the distribution of the variable of interest, spatial filters (i.e. geographic units) should be adjusted in geographical size in relationto the distribution and density of observed events.

In application to disease mapping, all maps that portray disease rates for discrete areas are likely to have their extreme rates (high and low) in areas with small populations at risk; especially where disease is rare and when measured over a short period of time (Rushton, 2003). These issues feed into what is known as the Modifiable Areal Unit Problem (MAUP) to be discussed below.

## 2.3 Spatial interpolation

Spatial interpolation describes the process of creating a surface model from spatial data, either in the form of discrete points or subareas, with the intention of predicting values at other points or for other subareas (Lam, 1980). These two applications address distinctly different problems; point interpolation (through isarithmic or contour mapping) attempts to designate values from sample points to locations where there are no values, commonly utilising exact methods or approximation methods (ibid.). Areal interpolation (through isopleth maps) instead transfers attributes from a spatially comprehensive set of boundaries with known values to other, often smaller but not always, geographical units with unknown values (Goodchild & Lam, 1980). This addresses the common problem of needing to obtain comparable estimates, such as population, for a set of *target* regions that do not respect the boundaries of the first, *source* dataset. This latter technique of areal interpolation is applicable to this research paper which utilises aggregated population data and will therefore be discussed further.

Nearly all geographic phenomena are scale-sensitive, therefore the ability to transfer data to the correct scale and units is paramount for accurate representation thereof. Openshaw (1984) theorizes this sensitivity as the modifiable areal unit problem (MAUP). MAUP is potentially experienced in two forms: the scale effect and the zoning/aggregation problem. The former refers to unavoidable analytical differences dependent on the size of units, with correlations generally being more pronounced for larger units (ibid.). Avoiding this is therefore reliant on decisions around how many zones should be generated. The zonation effect refers to the major differences dependant on how the zones are demarcated, even at the same scale (Openshaw, 1984; Chen et al., 2019).

There are several established areal interpolation methods including simple areal weighting, Pycnophylactic interpolation, area-to-point interpolation without ancillary information,

dasymetric mapping, street-weighting, statistical or geostatistical approaches, and point-based intelligence approaches with ancillary information (Comber & Zeng, 2019).

Simple areal weighting presents a cartographic overlay method that preserves volume, meaning that subdivided populations sum to the original population (Hallisey et al., 2017). Weights are determined based on the size of the overlap between source and target regions, calculated using a regular grid overlay (raster surface) (Hay et al., 2005). Populations are then assigned to each grid cell and used to reallocate to target zones. The key limitation associated with this is approach is the assumption of an even distribution of population within each source zone (Langford, 2006).

Building on from simple areal weighting, Pycnophylactic interpolation presents a mass-preserving method which proceeds to smooth the grid raster values iteratively, considering the weighted average of nearest neighbours (Hay et al., 2005). These weightings are however subjectively determined, and the method unrealistically assumes that no discernible boundaries exist in the distribution of the phenomenon, e.g. human population (Tobler, 1979).

Continuous surfaces are also created in the area-to-point interpolation without ancillary information, which is more commonly applied to estimating rates than underlying population counts (Comber & Zeng, 2019). In one such application, Goovaerts (2008) applies a Kriging formulation to create continuous maps of mortality risk. Benefits of this technique include reducing the visual bias associated with the interpretation of choropleth maps, however, there are a number of issues surrounding choice of location for the control point (Bentley et al., 2013). Point-based interpolators are also known to not be volume-preserving, therefore requiring rescaling (ibid.).

Several geostatistical approaches that were originally intended for point interpolation have subsequently been applied to areal interpolation (Comber, 2019), including areal-point interpolation via cokriging (Kyriakidis, 2004). Together with statistical areal interpolation

techniques, these methods are most useful where limited data is available, and often incorporate covariates to refine model estimates (Comber, 2019). Notably, regression analyses are often used to model the variable of interest with other ancillary data (ibid.). Estimates for unknown values are generated through a smooth prediction surface or probability surface, after which populations can be estimated for target zones (Hallisey et al., 2017). However, these models introduce complexity, require transformation of covariate geographies, and assume smooth distribution changes across the landscape (ibid., 2017).

Methods that incorporate ancillary data to improve their predictive accuracy can be grouped into four groups: those which apply areal masks to inform interpolation (dasymetric mapping), those using road networks to allocate populations along target zone road segments (street-weighting method), those which establish a statistical relationship between the ancillary data and the source zone data to guide allocation (statistical approaches), and those which use point data as ancillary information (point-based approaches) (Comber & Zeng, 2019).

In the field of public health, dasymetric approaches are often cited as the most useful method to better represent population, socioeconomic and health data (Barrozo et al., 2016; Wilson & Mansfield, 2017). Accurate, detailed population data can assist in visualising risk patterns and socioeconomic risk factors associated with the health issue of interest (Barrozo et al., 2016). The social and environmental factors most frequently correlated with TB are (in order of significance) (Shaweno at al., 2018): poverty, overcrowding and non-standard housing, ethnic minority populations, population density, low education status, health care access and immigrant populations.

Given the aim of this study to provide detailed population estimations, as well as the availability of land cover auxiliary information via SANLC and Eskom Satellite Building

Count (SBC) data, dasymetric interpolation methods have been selected as the most appropriate approach to compare against a conventional choropleth method. This chosen technique aims to increase the spatial resolution of data by incorporating related high resolution ancillary data layers, allowing the primary data to be represented at finer resolutions (Ngidi, Mans, McKelly, & Sogoni, 2017).

### 2.3.1 Dasymetric mapping

Dasymetric mapping, is an area-based cartographic technique popularised in the early 20th century by Wright (1936); exemplified via mapping Cape Cod into inhabitable and unhabitable regions. A subsequent body of literature illustrates a variety of methods to produce these thematic maps (Eicher & Brewer, 2001; Sleeter, 2004; Steiner, 2007; Mennis, 2009). The most common variations include the grid and polygon binary methods, the polygon and grid three-class methods, and the limiting variable method (Steiner, 2007). The use of ancillary data at the heart of these methods enables more precise population density locations and estimates. The most common form of ancillary data in dasymetric mapping is land-use/land-cover (LULC) information derived from satellite imagery (Mennis, 2009).

The method variations listed above are distinguishable by their approach to re-allocation of the disaggregated data (Mennis, 2009). Two underlying approaches exist: the binary allocation method and weighted method. In the binary allocation method, the quantitative attribute from the source zone (e.g. population) is allocated solely to dasymetric (ancillary) zones that have a functional relationship with the attribute. Population data, for example, may be exclusively allocated to residential zones that are considered to be inhabited (de Greer, 1926; and Wright, 1936 as cited by Mennis, 2009).

This allocation may be taken a step further by weighting the target zones prior to allocation, e.g. weighted land classifications. This is illustrated by the ‘three-class’ method as reviewed and presented in Eicher and Brewer (2001), Mennis (2003) and Sleeter (2004). This may be a

tempting method where hierarchical urban density classifications are available in the data. However, the population weightings of each urban class are often subjectively determined and assume homogeneity within each urban class across the whole study area. This can therefore create a false sense of accuracy (Eicher & Brewer, 2001).

The grid binary method, known for its simplicity, requires the ancillary data to be assigned into two classes. The frequently referenced Fisher and Langford (1995) were the first to publish a dasymetric areal interpolation method applied using 2-D land use data as the ancillary information. In this case, the binary dasymetric method is based on the area of populated and unpopulated land units. Similar to the urban classifications mentioned above, determining these parcel allocations is a subjective decision that is heavily dependent on the quality of the LULC classifications and knowledge of the study area. The classified land use data act as control units to disaggregate the source unit population and then again for reaggregate to obtain the target unit population.

A resurgence of the dasymetric methods, often attributed to Mennis (2003), is now almost solely associated with application to population density mapping. The catalyst for this revitalisation is accredited to advances in geospatial technology and the significant increase in availability and spatial resolution of spatial data more generally (Widener, Crago & Aldstadt, 2012). Satellite remote sensing is not a perfect indicator of population density but can describe the urban morphology or built-up and non-developed areas with increasing precision (Mennis, 2003). The dasymetric technique is now considered to be a special case of areal interpolation where the target units (raster grid cells) are much smaller than the source zones, and the intention is to approximate a continuous surface (Mennis, 2003).

Population mapping generally has two purposes: to cartographically portray the extent and density of population across an area of interest, and to derive quantitative estimates of

population density for use in subsequent spatial analytical modelling tasks (Langford 2003). The main benefit of the above proposed surface-base representation of population is that it allows for data aggregation to nearly any desired areal unit required for said tasks. Hence it is not subject to the MAUP and other areal unit-derived problems (Bracken 1993; Eicher & Brewer, 2001).

In application, this approach may help to identify fine-scale risk patterns of infectious and chronic diseases. In Nigeria, dasymetric interpolation of population data have been used to support health facility location modelling (Jega, Comber & Tate, 2017). The outcome was deemed incredibly useful for the site of application due to the presence of large water body which could be filtered out using populated LULC data. A key limitation identified, however, was the lack of appropriate approaches for validating the dasymetrically-derived populations. The focus of the study, as seen in similar international papers (Álvarez-Hernández, Lara-Valencia, Reyes-Castro, & Rascón-Pacheco, 2010; Ibrahim et al., 2015; Queiroga et al., 2012; Tadesse, Enqueselassie, & Gebreyesus, 2018), is on determining the relationship between TB and environmental/socio-economic conditions. There remains a gap in the literature for application of this method in the field of health in South Africa, and a focus on the methods themselves.

### 2.4 Comparison of choropleth and dasymetric methods

Geographical data accuracy is a statement of how closely data represent the real world. Factors to consider may include; omitted features, non-existent features present, correct classifications, up-to-dateness, and feature location accuracy.

Ensuring that features are discernible requires use of an appropriate spatial resolution. Consideration of data density – a measure of how many features per area are stored – may indicate a minimum feature size and therefore an appropriate scale of analysis. In core remote

sensing literature, Woodcock and Strahler (1987) propose the local variance method, which considers the relationship between the size of objects in a scene and the spatial resolution of sensors. As the size of spatial resolution increases, the size of objects can be approximate and the likelihood of neighbours being similar decreases as the values tend to be different from each other, thus causing the local variance to increase (Wu & Li, 2009). In the case of LULC for population mapping, a distinct scale of high variance could be related to the size of individual houses, or to groups thereof indicating suburbs, towns or cities (ibid.).

Local variance relates closely to spatial autocorrelation, but may also be used to compare values generated at different scales; thus becoming a measure of information lost due to inappropriate scale selection (Wu & Li, 2009). Choropleth maps, by definition, provide an aggregated representation of features with an inherent, acknowledged loss of information. The extent of this loss and its effects on analysis, however, is often unknown.

The dasymetric mapping method aims to provide more detailed population data in order to optimise any subsequent analyses reliant thereon. Given the proposed use of remote sensing LULC data, this accuracy is dependent on a use of an appropriate scale that can differentiate between the spatial objects being investigated.

## 2.5 Spatial Autocorrelation

Spatial autocorrelation provides a measure for statistically significant patterns in data (Mitchell, 1999). The concept was first proposed through global statistics which consider landscape-wide patterns (Moran, 1948; Krishna-Iyer, 1949; Geary, 1954), however the 1990s provided consequent local versions that explore local deviations from global trends (Getis & Ord, 1992; Anselin, 1994). More recently, other approaches to spatial autocorrelation have been proposed including cluster morphology (Kracaliket al., 2012).

Spatial data are unique in that the normal principles of independence assumed for transactional data in traditional statistics cannot be applied (Getis & Ord, 1992). Tobler's first law of geography instead declares that *'All things are related, but nearby things are more related than distant things'* (Tobler, 1979), motivating the consideration of spatial autocorrelation in datasets with geographic characteristics (Zhang et al, 2003). Consequently, spatial autocorrelation can be defined as "the correlation among georeferenced observations arising from their relative locations in geographic space" (Griffith, 2017) and is a fundamental quantitative and qualitative geospatial data feature, manifesting itself in assorted patterns on maps (Griffith, 2021).

There are several interpretations of 'relative' locations, dependent on how space is conceptualised. Common approaches utilise distance, contiguity/adjacency, neighborhood, proximity polygons and space–time window concepts (Cromley & McLafferty, 2012). This is an important consideration as "if the applied conceptualization method fails to reflect the inherent structure of the spatial relationships of the spatial features, the outcomes of the analysis will be misleading" (Grekousis, 2020, p 16).

### 2.5.1 Hot spot analysis

Once the desired conceptualisation method is selected, subsequent analyses such as cluster detection, better-known as a hot spot analysis, can be performed. This method identifies localised groups of similarity; clustering proximate spatial phenomena with low values result into cold spots, and clusters of high values form hot spots. A hot spot analysis is essentially an optimization problem, and several algorithms have been developed to determine optimal clustering of spatial units, according to their relationships in geographic space and similarity in attributes (Cromley & McLafferty, 2012).

In application to disease mapping, hot (or cold) spots are interpreted as any spatially-bounded area of significantly elevated (or reduced) risk (Lawson, 2018). Dowdy et al. (2012) illustrate

the important role of high-incidence hot spots in propagating TB epidemics. They estimate the proportion of TB transmissions originating from a high-incidence hot spot within Rio de Janeiro in order to determine the relative impact of targeted interventions in the hot spot area versus a blanket intervention throughout the wider community. They consider two scenarios; one where the hot spot population is self-contained and does not transmit to the surrounding community, and one where these cross-transmissions are allowed. In the latter scenario – which is more realistic given global migration trends in cities – the hot spot generated 35.3% of all transmission events despite only containing 6% of the city's population and 16.5% of its TB incidence. Although not an exact science, the paper's findings strongly affirm the vital role that strategic interventions that target hot spots can play in the wider battle against TB.

## 2.6 Research assumptions

### 2.6.1 Study population

This study considers the total population of Johannesburg, as available in the census data, when calculating mPTB incidence. This population therefore innately includes those who are insured and uninsured. It was decided not to treat these groups as separate populations due to the complex cross-utilisation of private and public facilities often associated with diagnosis and treatment of the disease. Although there is a gap in the literature for providing substantial quantitative evidence of true utilisation of medical facilities, it has been suggested that in South African physicians do not traditionally treat TB in the private sector, and are more likely to refer these patients to the public health care system (Sinai et al., 2018; Boffa et al., 2019). This may seem counterintuitive given that Gauteng accounts for 40% of the country's medical aid beneficiaries, with approximately 24% of the region's population covered (Council for Medical Schemes, 2019). TB diagnosis, treatment and care costs should in theory be covered in full by Prescribed Minimum Benefits (PMBs), regardless of the plan that any member and their

dependents are on (Council for Medical Schemes, 2017). It has, however, been suggested that many private doctors simply do not know how to treat TB, leading to an overburdened public health sector (Boffa et al., 2019).

Other indicators of the small role played by the private sector in TB diagnosis and treatment include figures from 2012 that assign 93% of the overall volume of TB Xpert tests to the public sector (TB Diagnostics Market Analysis Consortium, 2015), meaning only 7% are conducted in private facilities. Similar proportions were confirmed in the TB Prevalence Survey (2021) where 8.1% of surveyed patients sought care in the private sector. Private TB medication sales

also only accounted for 15% of total sales in South Africa in 2019 according to the WHO Global TB Report 2019 (as referenced in Stallworthy, Dias & Pai, 2020).

### 2.6.2 Utilisation of nearest facility

A significant assumption made in this study is that people utilise their nearest medical facility; presented in the data as the facility of diagnosis. This assumption is supported by the General Household Survey (2016) where the vast majority of the population (92.6% given a representative sample of 19 657) indicated that they attend their nearest health facility when falling ill or requiring emergency assistance.

There is limited application of GIS present in South African literature on PHC accessibility for general or TB-specific purposes (Mokgalaka, 2014; Kapwata et al, 2017; Tanser & Wilkinson, 1999; Tanser, 2006; Tanser & Gijsbertsen, 2006; Nteta, Mokgatle-Nthabu & Oguntibehu, 2010) with the majority of existing research orientated around assessing whether populations are within the recommended distance and travel-time guidelines proposed by *National Core Standards for Health Establishments in South Africa* (NDoH, 2011). Proximity to care is an important factor for diseases such as TB which have demanding treatment regimens (Wilkinson & Tanser, 1999). A number of these studies also present incidental findings related to nearest facility usage, but those with a general focus on PHC actively exclude specialised TB facilities, therefore omitting a portion of the TB patient cohort.

Mokgalaka (2014) considers accessibility (including capacity) and utilisation (i.e. 'revealed' accessibility) of health facilities in Johannesburg. This study applies the CSIR's calculation of capacity ('the potential to accommodate visits') as a function of facility size, staffing levels and range of services (Mokgalaka et al., 2013). Patient register ETR.Net TB data is used as a proxy for primary healthcare (PHC) demand, and reveals that almost 45.26% of the patients from the register did not use their nearest facility as a first point of contact, despite only 1% of

the patients residing outside the City's boundary (thus avoiding what could have been a potential source of skewness in the data). However, the study also only considers the uninsured population which is problematic given the complex cross-utilisation of private and public facilities, discussed above. A deeper understanding and more accurate linkage of public and private patient data is needed before these insured and uninsured populations are treated separately. Kapwata et al. (2017) perform a network analysis from a sample of adults and children diagnosed with XDR-TB in KwaZulu-Natal. A comparison between distance to nearest facility and actual distance travelled to facility of diagnosis reveals that the majority of the XDR-TB positive population do not attend their nearest facility. Findings also acknowledged the disparities between rural and urban communities with regards to accessing PHC. Residents of eThekwini, classified as 84.8% urban and 15.2% rural, had an average distance to nearest hospital or clinic of 0.49 and 6 kilometres respectively, but claimed an average actual distance travelled of 19.57 kilometres. Results also showed that 43% of the XDR-TB patients in the study enrolled at specialised provincial TB hospitals. These findings are critical to spatial understandings of XDR-TB diagnosis and treatment, but may not be applicable to more common drug susceptible (DS) TB which is less severe and may be treatable at non-specialised facilities located closer to patients.

The last notable studies to be discussed are those of Tanser et al. (2001) and Tanser, Gijsbertsen and Herbst (2006). In both cases, the rural homesteads of Hlabisa in northern KwaZulu Natal are investigated. In 2001, the authors compared actual PHC clinic usage to predicted usage determined by simple Thiessen polygons. It was found that the Thiessen-based model had an accuracy of 87%, indicating that this percentage of homesteads made use of their nearest clinic. In 2006, the authors decided to build on this research, conducting a catchment area analysis that factored in auxiliary data on physical access. Their model takes into account the proportion of people likely to be using public transport (as a function of estimated walking time to clinic),

and the quality and distribution of the road network and natural barriers. Overall, the model was able to predict reported clinic usage with an accuracy of 91%, only 4% better than the previous study. The Thiessen, distance-based polygons therefore provide a relatively competitive indication of clinic usage with significantly less resources, time and computations required.

It should be noted here that geographic accessibility is just one indicator of utilisation of and access to healthcare services, alongside quality of care, suitability of available services, and financial constraints of patients and healthcare systems. Other factors related to preference may also influence where patients choose to seek care, including: seeking care in areas of employment rather than place of residence; seeking care in urban areas perceived to be offering better quality treatment, facilities, medication and staff; shorter waiting times in facilities with larger capacities elsewhere; and seeking care away from place of residence due to stigma associated with TB (Kapwata et al., 2017). There is significant international literature on health care facility preference, with a gap in literature focusing on South Africa. We therefore assume patient utilisation of the nearest medical facility whilst acknowledging the need for deeper investigation into this topic.

# Chapter 3: Methodology

## 3.1 Introduction

This chapter introduces the study area in detail, describes the datasets acquired, and outlines the steps taken to process said datasets into workable formats for the comparison of choropleth mapping processes and the dasymetric mapping method. Processing was required for non-spatial attribute data, geographic data in vector format, and geographic data in raster format.

The two mapping techniques were considered in the context of spatial epidemiology, understood as being "concerned with describing, quantifying, and explaining geographical variations in disease, especially with respect to variations in environmental exposures at the small-area scale" (Elliott et al., 2000, p V). These methodologies aim to address the first half of this definition; seeking to present an optimised approach to describing variations in mPTB incidence. This contributes to the literature of using spatial analysis methods to provide decision support (Rushton, 2003).

The methodology can be classified as a comparative case study; contrasting currently utilised choropleth mapping with an advanced dasymetric method.

## 3.2 Study Area

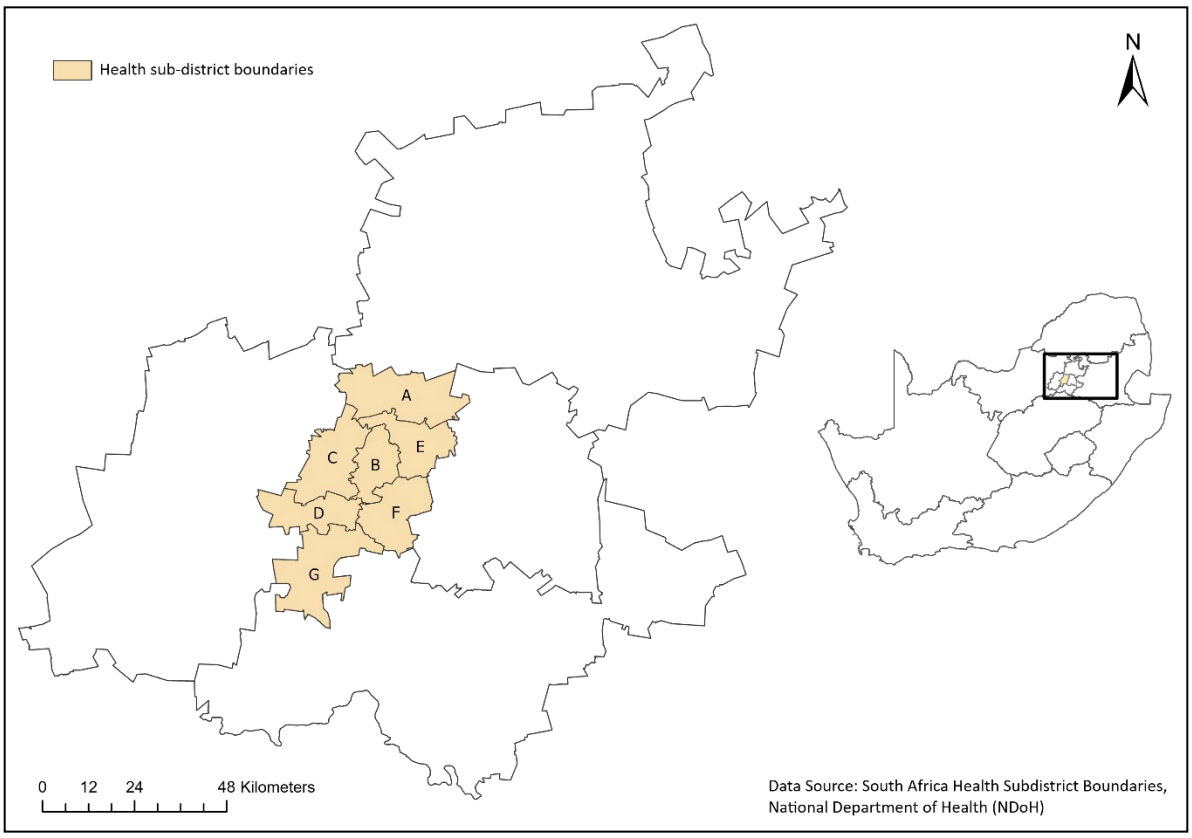



*Figure 1: Johannesburg's health sub-districts*

The City of Johannesburg (CoJ) was chosen due to its metropolitan nature, the researcher's existing knowledge of the study area, and availability of various sources of topical data for the region. Over time, the urban metro has consistently been faced with the paradox of being a world-class, economic powerhouse whilst harbouring high levels of inequality (Crankshaw & Parnell, 2004; Everatt, 2014; Pieterse & Owens, 2018). It can be expected that there is greater heterogeneity in mPTB incidence than one would, for instance, expect in more socio-economically homogeneous rural areas. Existing studies have, however, indicated that access to clinics within the CoJ is very good as clinics are well-distributed throughout the area, reflecting population distribution (Green et al., 2012). However, environmental risk levels vary with many households inhabiting high density informal settlements; a hotbed for TB transmissions (ibid.). Johannesburg is also known as an economic destination for seasonal workers,

illustrated by circular migration patterns between rural and urban areas (Lurie & Williams, 2014). This population movement may too contribute to transmission of mPTB into and out of the city.

Information available in the ESRI Demographic dataset acquired for this paper allowed us to map poverty and population density in the CoJ as two initial exploratory indicators of potential sites of increased disease burden. Figure 2 illustrates 2017 estimates for median individual income with much of the metro's wealth concentrated in regions B and C (predominantly characterised by suburbia, including some of the country's wealthiest areas such as Sandton), with the lowest income groups located in the southern areas of G and D (containing Soweto, South Africa's most well-known township). These low-income regions correspond with areas containing the highest counts of people living below the poverty line (Figure 3). Given that data is displayed at an Enumeration Area (EA) scale, with between 100 and 250 households, the smaller polygons with darker purple shades represent high density areas with high proportions of residents living below the poverty line. The extraordinary population densities (Figure 4) present in the data, with certain areas exceeding 80 000 persons/km$^2$, correspond to existing literature on the study region (Mokhele, 2011; GCRO, 2021).

In terms of existing knowledge of mPTB burden in the region, the province of Gauteng falls into the lowest burden stratum, as categorised in The First TB Prevalence Survey 2021, and only accounts for 14.6% of total cases for 2018 and 2019 combined. The CoJ is the greatest contributor in the province, accounting for approximately 37.7% of cases.

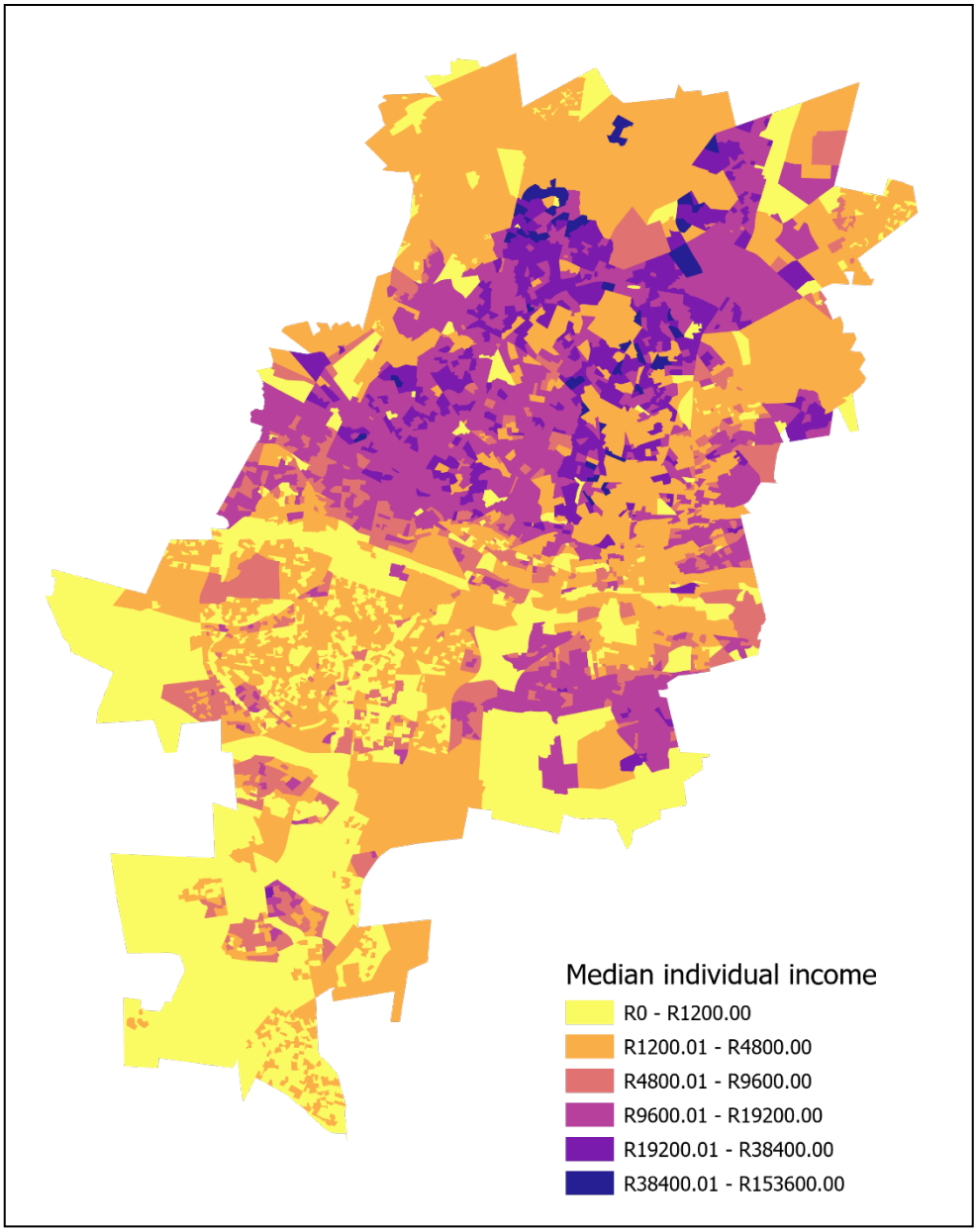



*Figure 2: Median individual income by EA for 2011*

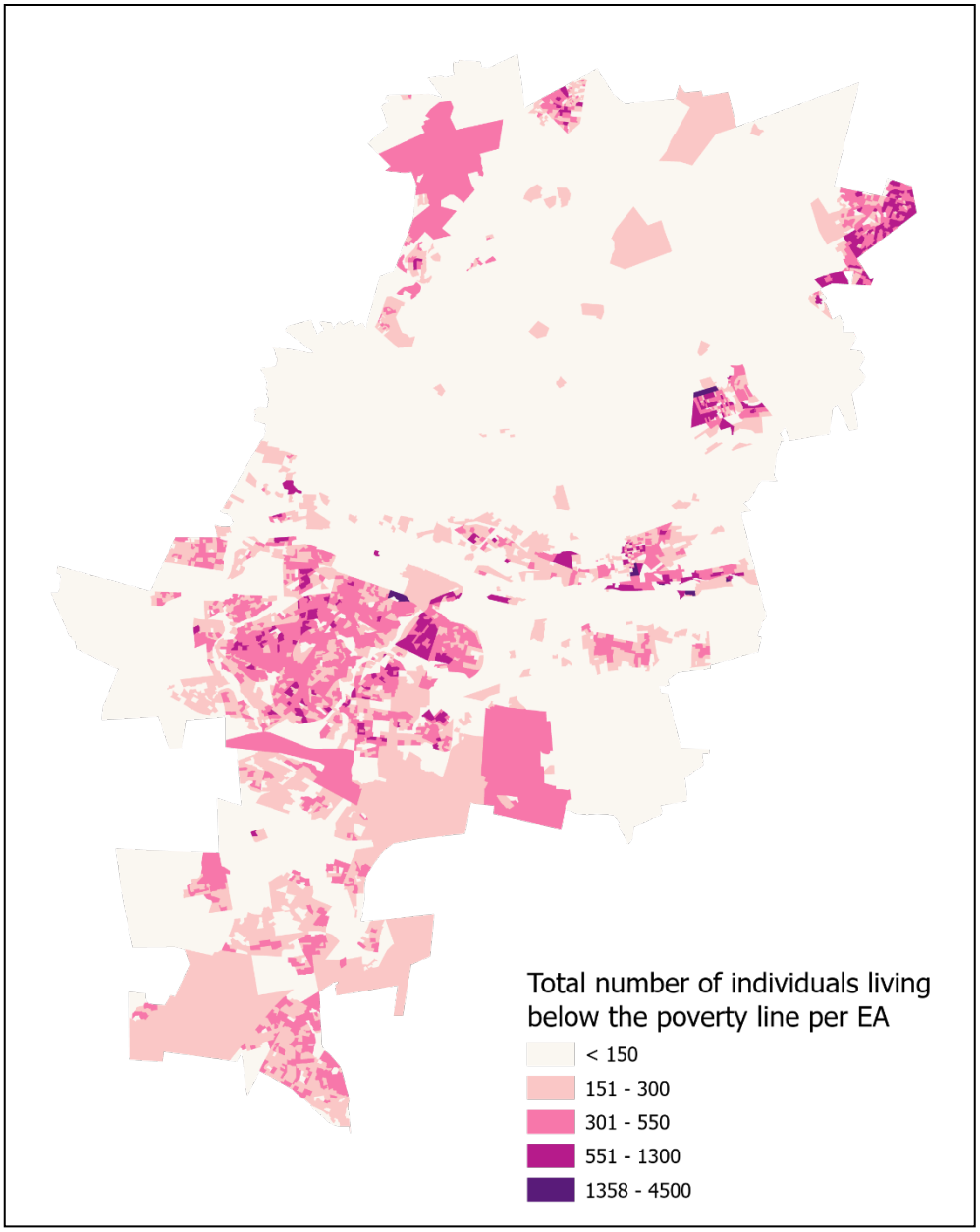



*Figure 3: Number of individuals living below the poverty line per EA for Johannesburg*

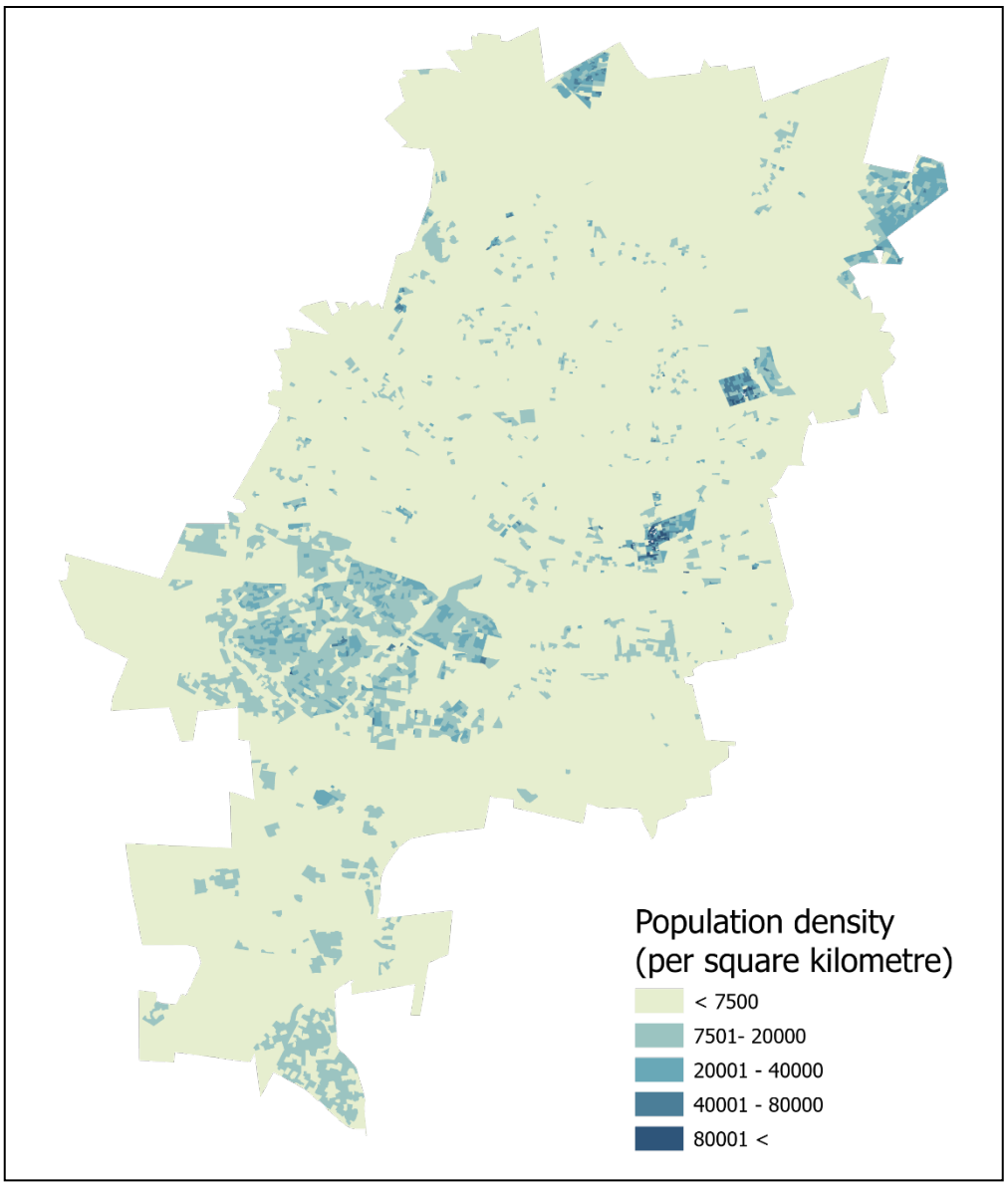



*Figure 4: Population density (per square kilometre) estimates for 2017*

### 3.3 Data acquisition and preparation

The datasets listed in Table 3.1 were utilised for determining mPTB incidence using choropleth mapping and dasymetric mapping methods, and will be discussed in detail in this chapter.

*Table 3.1: Summary of datasets utilised*

| **Dataset** | **Spatial resolution** | **Temporal resolution** | **Data format** |
|---|---|---|---|
| TB laboratory data (NICD/NHLS) | Facility-level | Jan 2018 – Dec 2019 | Positive laboratory counts by facility for the years 2018 and 2019 (csv) |
| ESRI Ideal data | Enumeration area (EA) | 2017 estimates | Population estimates derived from 2014 Stats SA data (csv) |
| Health sub-districts for South Africa (SAEON) | Health sub-district | 2016 | Polygon shapefile (shp) |
| South African National Land Cover (SANLC) | 20 metre (raster pixels) | 2018 | GEOTIFF format raster with 72 land cover classifications generated from 20-meter multi-seasonal Sentinel 2 satellite imagery |
| Eskom Spot Building Count data (SBC) | Derived from 8-metre satellite imagery | 2018 | Vector points and polygons derived from 2017 satellite imagery from the SPOT 6 sensor system (with a multi-spectral image resolution of 8m), classified into 10 categories |

### 3.3.1 Microbiologically-confirmed Pulmonary TB (mPTB) data

The use of health records for research is considered to be secondary use of healthcare data; that is, the data were first collected as part of routine patient care and will be secondarily used for research (Zozus et al., 2015). mPTB laboratory data was acquired from the NICD, via the NHLS central data warehouse (CDW), providing positive pulmonary TB laboratory counts aggregated to the facility level for 2018 and 2019.

These facilities can be grouped into a number of location types: care centres, community health clinics, clinics, gateway clinics, hospitals, mobile clinics, prison clinics and SANDF/SAPS clinics. These can further be grouped into five umbrella categories; hospital, clinic, prison, military and private.

For the purposes of this study, public clinics with permanent locations were included. Hospitals were excluded as their catchment areas are more complex requiring unique considerations, and prison and military facilities are excluded as these populations are not spatially organic.

### 3.3.2 Population data

Enumeration area (EA) level census data (ESRI demographic data) was also provided by the NICD, as previously acquired from ESRI. EA's are the smallest geographical unit used for enumeration purposes in South Africa, containing between 100 and 250 households (Stats SA, 2011). The acquired data provides estimated night-time population counts for 2014 by EA, calculated using Stats SA's 2011 modelled mid-year estimates.

When needing to determine TB incidence, the numerators will be annual mPTB figures (3.3.1) and denominators will be these population estimates.

### 3.3.3 Land-use dataset

The most commonly used ancillary data for determining populated areas is classified land use/ land cover (LULC) raster data generated from satellite imagery (Shaweno et al., 2018). Two applicable datasets were identified from South African literature.

The South African National Land Cover (SANLC) data for 2018 is publicly available and was obtained in GEOTIFF format from the National Department of the Environment, Forestry and Fisheries website. This classification system is generated from 20-metre multi-seasonal Sentinel 2 satellite imagery and presented in 72 land cover classifications. There is evidence of use of previous versions of the SANLC dataset for dasymetric mapping of population (Hodgson, 2017) and land cover change (Schoeman, Newby, Thompson, & Van den Berg, 2013). The overall accuracy for the dataset has been estimated to be 90.14% (Thompson, 2019). Consideration was given to residential land use classifications in the dataset.

Other case studies (Mokgalaka, Smit, & Mans, 2015; Ngidi et al., 2017) utilise Eskom's Satellite Building Count (SBC) (also known as Eskom dwelling layer) point and polygon datasets as vectorised forms of LULC. This data is available to all government departments and academic institutions on request. The latest publication is derived from 2017 imagery from the SPOT 6 satellite system (with a multi-spectral image resolution of 8 metres). Building point locations are provided and classified into 10 categories.

Since we are only interested in potentially populated residential buildings, the SBC point data were filtered to only include features classified as class name = 'Dwelling', defined in the metadata as a 'building structure where people live'. A separate polygon feature shapefile is provided for dense informal settlements, defined as areas where settlements are too dense to determine the number of structures (HDA, 2012).

The following rules were followed by Eskom in defining a *point* for the output point shapefile (Eskom Settlement Growth Dataset Standard, 2017):

a) One point per rural housing structure in farms and villages.
b) One point per grouping of housing structures as found in KwaZulu-Natal and Eastern Cape.
c) One point per grouping of urban structures, e.g. townhouses, and schools.
d) One point per any other individual identifiable building structure.
e) One point per formal cadastre polygon; few exceptions exist based on the accuracy of the cadastral.

It was presumed that the SANLC would provide a more thorough representation of populated areas, given that some Eskom SBC points represent groupings of housing and urban structures. However, on inspection of the data, the latter was able to detect settlements unbeknownst to the former (Figure 5).

*Figure 5: Example of discrepancies between SBC and SANLC populated area pixels*

It was therefore decided that a combination of the raster and vector datasets would be used to provide a comprehensive representation of populated areas.

### 2.7 Choropleth method

The basic choropleth mapping method was carried out to represent how standard microbiologically-confirmed Pulmonary TB (mPTB) incidence rates are most frequently represented given routinely available data. mPTB notification data was aggregated by health sub-district, with no consideration of true facility catchment areas. The population EA boundaries correspond with the health sub-district boundaries, therefore incidence was calculated using a cumulative population total for each polygon, at a rate of mPTB count (accumulated over two years) per 100 000.

## 2.8 Dasymetric method

One of the key benefits of the dasymetric process is the ability to transfer data from one source spatial unit to an entirely different set of target units, regardless of coincidence of the layer boundaries (Mennis, 2003). This study took advantage of this to aggregated population by facility catchment area, rather than health sub-districts used in the choropleth method above.

Facility catchment areas were generated using Thiessen, or Voronoi, polygons; a method use to proportionally divide and distribute a point coverage into regions (ESRI, 2016). The process generates a triangulated irregular network (TIN) from all the points, after which perpendicular bisectors for each triangle edge are generated, forming the edges of the desired polygons (Figure 6). Utilising Euclidean distance, these distance polygons divide space such that the underlying population data could be allocated to the nearest facility.

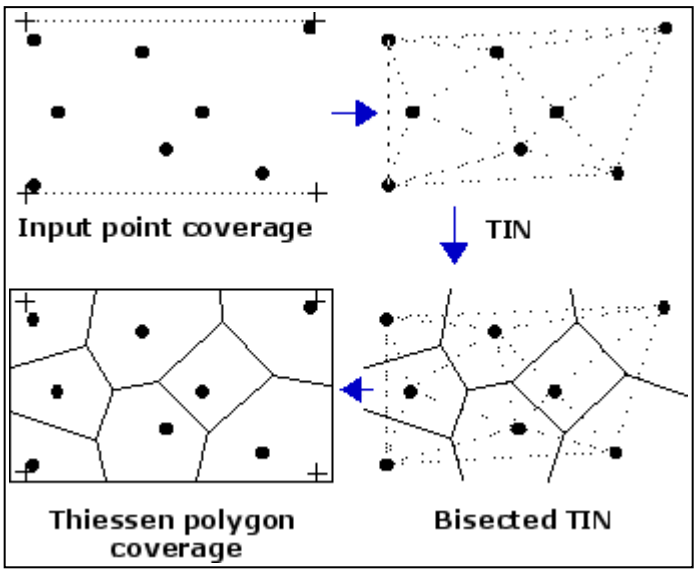



*Figure 6: Creating Thiessen polygons from points (ESRI, 2016)*

Given that these polygons do not conform to administrative boundaries, it was decided to generate catchment areas for all of the facilities in Gauteng, and any polygons intersecting with Johannesburg would be considered in all analyses. Figure 7 shows the location of each clinic with a catchment area that intercepts with the boundary of the CoJ. The dark grey polygons

represent the study area for later dasymetric mapping procedures (Figure 7(1)). These are then clipped to the boundary of Johannesburg and compared to the EA boundaries containing population data (Figure 7(2)). It should be noted that since the derived facility catchment areas do not respect the boundaries of Johannesburg, a larger total population was captured by the new boundaries, especially on the peripheries of the study area. This may not have a substantial impact on areas of lower density – which is often associated with urban peripheries – however, Johannesburg, along with several other southern African cities (Melo, 2016), are unique in that the outskirts of the metropolitan is experiencing increased formal housing developments as well as being scattered with low-income, high density informal settlements as a result of high demand for housing nearer to jobs and services (Harrison et al., 2014).

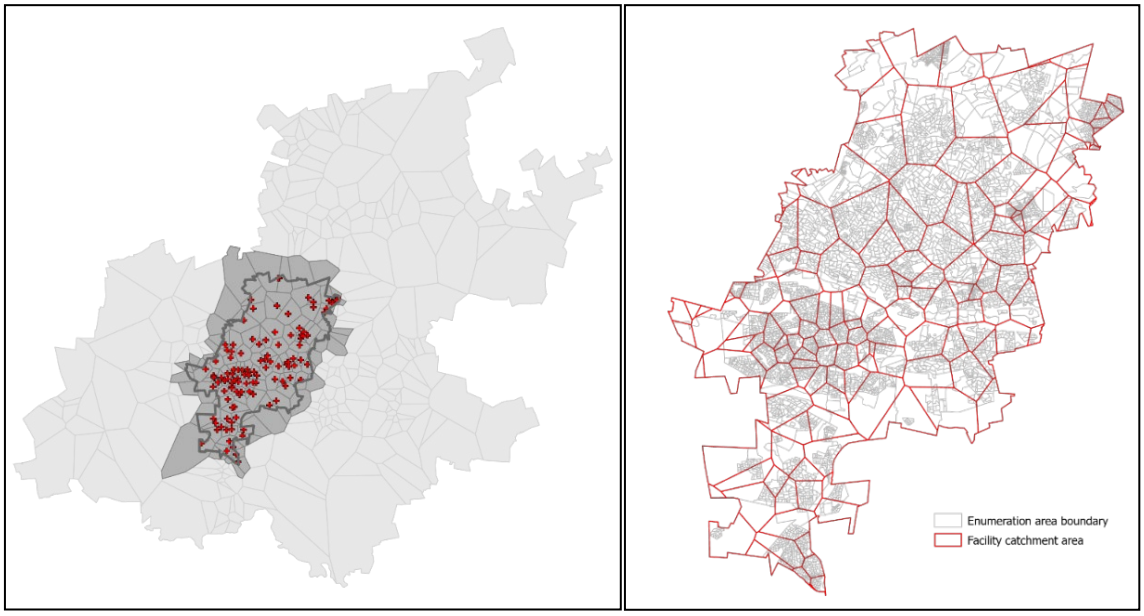



*Figure 7: Gauteng facility catchment areas (1), filtered to those intersecting with the Johannesburg boundary (2)*

These facility catchment areas define the target boundaries for the dasymetric methodology presented next.

The following grid binary methodological steps were taken (following the logic of Eicher & Brewer (2001)) to achieve a dasymetrically-derive mPTB incidence rate for each facility catchment area:

1. Create a binary (residential versus non-residential) dataset from LULC inputs
2. Convert EA census areas to raster
3. Zonal statistics tool was used to calculate the number of residential pixels per EA census tract
4. Simple arithmetic was used to calculate population per residential pixel in each EA polygon
5. Zonal statistics was then used to sum the pixel values by facility service area; achieving population count per Thiessen polygon
6. TB incidence (per 100,000 persons) could then be calculated per facility service area: (total sum of mPTB cases * population / 100 000)

The grid cell resolution of the SANLC raster (20 metre) serves as the resolution for the final population surface. Higher resolutions are naturally preferential but need to be balanced with considerations of acceptable processing times of subsequent GIS operations. This resolution is acceptable given the researcher's available computational capacity, and is believed to be fine enough to capture the desired spatial variation of population and mPTB incidence within the areas of interest.

In order to combine the raster and vector LULC layers, the SBC point layer as well as the informal settlement polygon layer, were transformed into a raster grid with equal sizes to that of the SANLC (for consistency and merging purposes). Any pixel overlapping with an SBC point or polygon was allocated a value of 1 (indicating that it is populated) using the reclass tool.

The two SBC outputs could then be merged with the binary SANLC layer, using maximum pixel values to resolve overlapping pixels; i.e. to ensure the value of populated pixels with value = 1 are transferred to the final layer, with the remaining pixels taking the value of 0.

Figure 8 displays the final binary LULC map to be utilised as target zones for the population and mPTB incidence data. When calculating disease incidence, this method assumes an even spread of rates amongst the inhabited pixels in each facility catchment area.

Figure 8 also shows that the majority of Johannesburg's surface area is potentially inhabited, especially in the central and northern areas. The empty, white band running along the east-west axis illustrates the Witwatersrand ridge, or previously proclaimed gold fields. Historically this created a topological barrier between the white northern suburbs of Johannesburg, and the government provided housing to the south of the ridge where black residents were legally restricted to (Mubiwa & Annegarn, 2013). Areas south of the ridge are characterised by colder micro-climates and intrusive dust from the mine dumps; also often associated with respiratory health issues (SAHO, 2011; Kootbodien et al., 2019). These socio-economic fractures in the Johannesburg landscape were reinforced through various apartheid spatial planning strategies, the legacies of which are still present, as seen previously in Figure 3 where poverty is largely concentrated in the south.

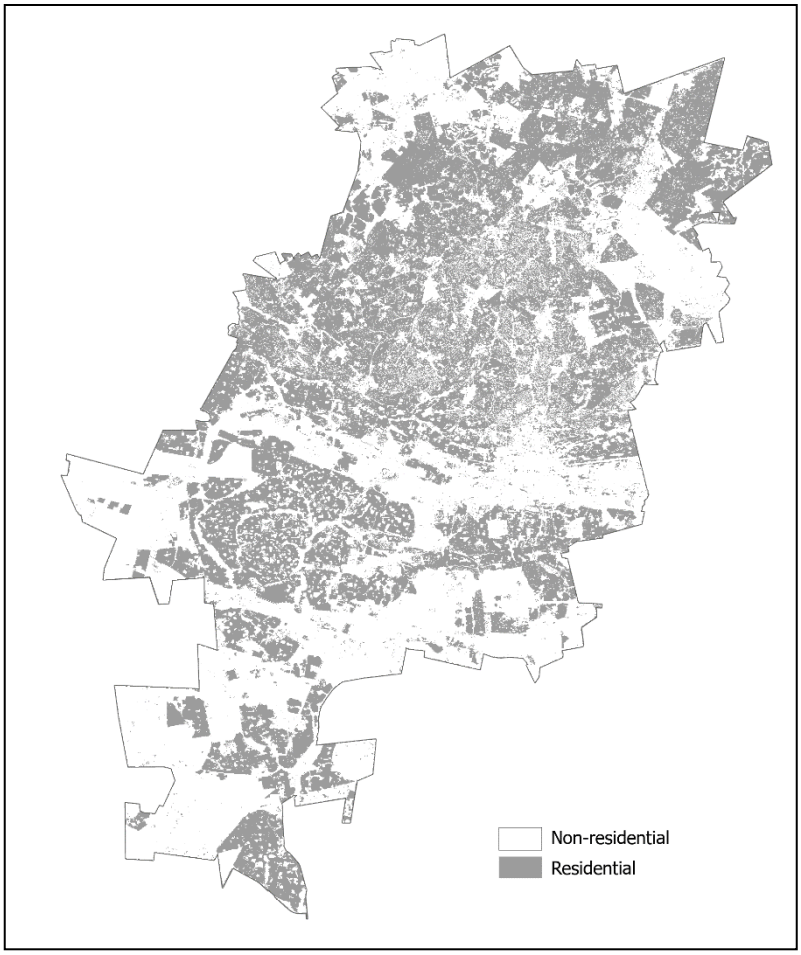



*Figure 8: Binary pixelated map of populated areas, derived using SANLC and SBC LULC data*

### 2.9 Comparing the two methods

The results from 3.5 and 3.6 each provide univariate maps of mPTB incidence rates that we could then compare. Again, it is noted that the underlying population for the choropleth and dasymetric mapping outputs slightly dissimilar and the overlaying mPTB data varied between the models due to different interpretations of health facility catchment areas. The comparison between rates therefore includes the effects of both mPTB count and population differences.

In traditional statistics, standard deviation provides an indication of the spread of the data around the mean. This basic concept was utilised and adjusted for the purposes of this project, where we were interested in how the dasymetric output mPTB rates differ with those produced using a choropleth mapping method. Therefore, the choropleth rate values (which assume homogeneity across each health sub-district) were treated as the representative values, or

pivotal means, around which the variation of the dasymetric rates were calculated.

Given standard measures of variation for the population, the following equations will be used:

$$Population\ Variance\ = \frac{\sum(x-\mu)^2}{n}$$

$$Population\ Standard\ Deviation\ = \sqrt{\frac{\sum(x-\mu)^2}{n}}$$

Where the choropleth value is substituted in for $\mu\mu$ (traditionally the population mean).

These variations were measured for each health sub-district. Values for each pixel of the dasymetric mPTB rates map were extracted by health sub-district, and standard deviation (substituting mean values for the choropleth outputs) was then calculated in RStudio.

### 2.10 Spatial autocorrelation

The Getis Ord Gi* inferential statistic is one of the most commonly utilised local statistics and works around the null hypothesis that states that there is no spatial clustering of feature values (Kiani et al., 2021). The statistic returns a z-score, indicating the intensity of positive and negative clusters (hot and cold spots). For statistically significant positive clusters, higher z-score values indicate more intense hot spots, whilst the inverse is true for statistically significant negative clusters where smaller values indicate more intense cold spots. This statistic can be applied to almost any localmethod of cluster analysis.

In order to determine mPTB hot spots from the dasymetric mPTB incidence output derived from section 3.5, different conceptualisations of spatial relationships were considered, each utilising the Getis Ord Gi* statistic. Firstly, consideration was given to global trend using the Global Moran's I statistic. Thereafter, three standard relationships were illustrated; KNN, distance threshold and inverse distance. Lastly, an increasingly recognised more complex,

cluster morphology algorithm was demonstrated; A Multidirectional Optimum Ecotope-Based Algorithm (AMOEBA).

A variety of standard methodologies can be carried out using the in-built Getis Ord Hot Spot Analysis (Getis Ord Gi*) spatial statistics tool in ArcGIS Pro 2.7.2. The recommended default conceptualisation of spatial relationships for hot spot analyses tool in ArcGIS is Fixed Distance Band was used as this works well for point data, or for polygons which vary significantly in size. This approach constrains the algorithms search for feature neighbours according to a user-specified distance limit. Popular alternative methods include the two inverse distance weightings; inverse distance and inverse distance squared. These are most appropriate for continuous data or to model processes where the closer two features are in space, the more likely they are to interact with or influence each other (ESRI, n.d.). Here a threshold distance can be specified to reduce the number of computations required for the function, or, for smaller datasets, it is possible to treat every feature as a potential neighbour of every other feature. In this latter scenario the algorithm is exhaustive.

Another consideration within the function was that of whether to adopt a correctional parameter to account for multiple testing and spatial dependency. This is of increasing importance with larger datasets. Classical statistical correction methods are available, e.g. Bonferroni correction, however they are often incredibly conservative and overcompensate in their effort to avoid false positives thus missing statistically significant results. False Discovery Rate (FDR) corrections offer a less conservative, ranked approach that only exclude those results with the weakest p-values. Given that there are only 141 facility catchment areas in this paper's study area, no corrections were used, but if this method should be extended to a national scale, it would be advised.

### 3.8.1 The AMOEBA method

The Multidirectional Optimum Ecotope-Based Algorithm (AMOEBA) process was carried out in RStudio, and the results were mapped back into ArcGIS Pro. Given that this is a lesser-known method, it is described in more detail below.

In theory (although out of the scope of this study) the output and any clusters discovered would be discussed and assessed a posteriori (Rushton, 2003), most commonly to further investigate relationships between disease rates and other covariates.

This method provides an algorithmic approach for generating spatial weights from univariate areal data with a spatial component (e.g. census enumeration units) (Aldstadt & Getis, 2006; Grubesic et al., 2014). It is a spatial cluster detection method as well as a spatial weight matrix construction tool (Tao & Thill, 2019). Therefore, used in conjunction with the Gi* algorithm, the procedure offers an exhaustive measure of local spatial autocorrelation while detecting irregularly shaped geographic clusters.

The conceptualisation of space used here is similar to the neighbourhood approach illustrated by KNN in that AMOEBA is a contiguity-based algorithm. However, it does not assume that clusters only exist in circular or compact regions, providing a novel approach whereby irregularly shaped clusters may be determined. This is achieved by adopting a region growing technique, similar to that found in image segmentation literature where segmentation is achieved through an iterative process of examining the neighbours of initial seed locations and deciding whether to add them to the region or not. This process is illustrated in Figure 9 below.

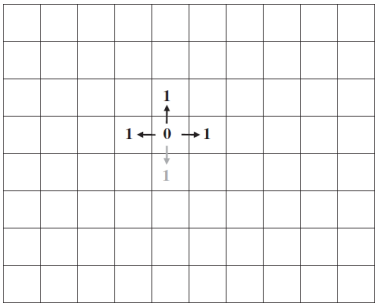
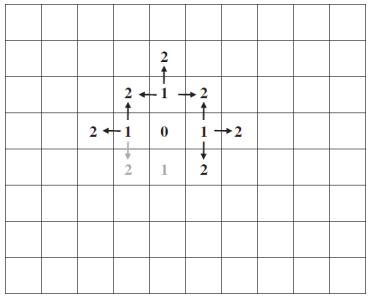
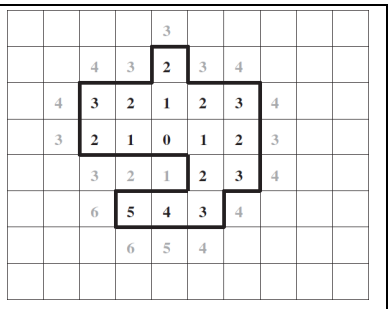
| | | |
|---|---|---|
|  |  |  |
| 1 a) Calculate Gi* value for each spatial unit, denoted $G_i^*(0)$<br><br>b) Calculate Gi* for the combination of the target and surrounding pixel. Only include those where the new combined value is greater than the individual i[th] Gi* value (greyed arrow indicates links that are excluded). | 2) Compute $G_i^*(1)$ for each region $i$ and all combinations of its contiguous neighbours. | 3) The final ecotope ($k_{max}$) for k number of links is identified when the addition of any set of contiguous units fails to increase the absolute value of the $G_i^*$ statistic. |

*Figure 9: The AMOEBA procedure* (Jankowska, Aldstadt, Getis, Weeks, & Fraley, 2008)

In the first step, a value of $G_i^*(0)$ greater than 0 indicates that the value at location $i$ is greater than the mean of all units. A value less than 0 indicates a value smaller than the mean (Aldstadt & Getis, 2006). A Gi* statistic is then calculated for this initial seed location and its contiguous neighbours. Neighbours are included to form a cluster (or ecotope) if the combined Gi* statistic is greater than that of the seed value. This process is iterated over all of the features ($i$) in the study area. Once the ecotopes with the highest Gi* values are identified (each with a value for $G_{max}^*$), a Monte Carlo-type permutation test is performed to establish the statistical significance of each ecotope.

The AMOEBA package in R (as part of the CRAN repository) presents the Amoeba function as follows:

```
AMOEBA(outc, neig, power, cpu)
```

Where:

`outc` = Numeric vector with the study outcome

`neig` = Neighbours object from "spdep" package

`power` = Number of times to apply AMOEBA

`cpu` = Total number of cpu to run faster the algorithm

The power value tells the algorithm how many times the AMOEBA must be applied. When applied once, the algorithm is called AMOEBA, and when applied more than one time, the algorithm is called AMOEBA+ (Aldstadt & Getis, 2006; Duque et al., 2011). It is stressed that the power selected should satisfy that $3^{power} < length\ (outc)$, otherwise the algorithm will fail. The AMOEBA and AMOEBA+ algorithms were both explored, with the latter utilising a power of 4 given that there were 139 facilities (i.e. *outc* = 139) to consider.

The available parameters also allow for parallel computational implementation of AMOEBA, as developed by Widener, Crago and Aldstadt (2012), through the use of multiple central processing units (CPUs). This spreads the processing load between CPUs in order to improve processing time, especially for large datasets. For the purposes of this project with relatively small datasets, however, one CPU was used.

The output of the above function returns a vector with a 3-way classification of the outcome. The classification is ordered from lowest (clustering of low values) to high (clustering of high values), and identifies which areas are lower than the mean, those areas near to the mean and the areas which are over the mean. Where AMOEBA(+) is applied, this outcome has as many

as $3^{power}$ classification levels; thus increasing the resolution of so-called 'risk' levels (Castellano et al., 2013).

# Chapter 4: Results

## 4.1 mPTB incidence rate outputs

### 4.1.1 Choropleth output

Incidence was calculated for a two-year period (2018-2019) per 100,000 population, and the relative rates between regions are illustrated using a graduated colour-scheme of varying increments with darker colours for areas with greater rates (Figure 10). These values range from just under 180 to over 461.

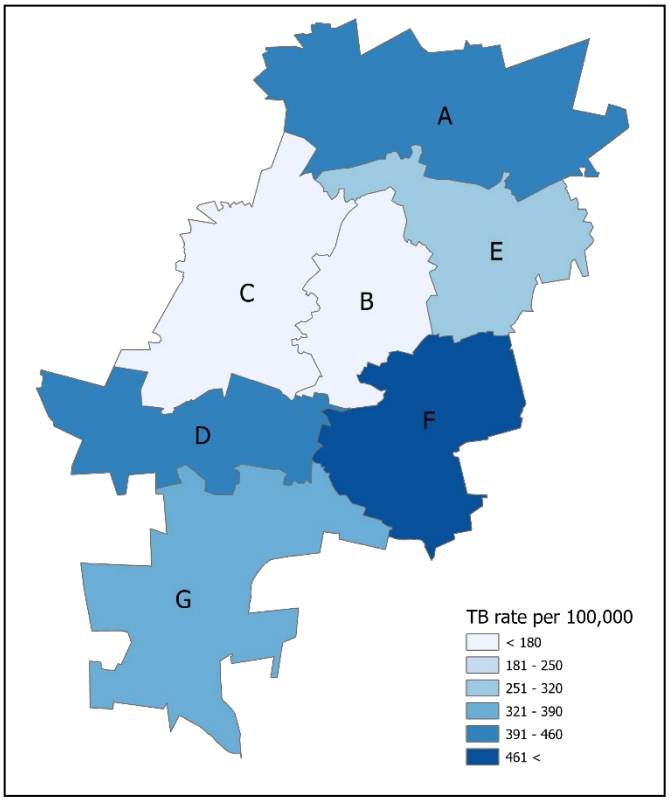



*Figure 10: Health sub-district TB rates per 100,000 population for the two-year period 2018-2019*

This choropleth representation of mPTB incidence rates generally reflects a distribution that could be predicted from previous illustrations of the distribution of population income and areas of poverty (Figures 2, 3, and 4). The lowest incidence rates are located in regions B and C which are high income, low density regions. It is however, surprising to see that Region F

lights up as the metropolitan's worst affected area with a rate exceeding 461 (averaged at 230.5/100 000 per annum). Nevertheless, these mPTB rates are in general reasonably low.

This map would suggest all efforts should be targeted towards region F (containing 15 facilities), and potentially D (31 facilities) and A (15 facilities). As previously mentioned, rates are influenced by underlying population with smaller populations being more vulnerable to extremely high or low values. This could be at play here as region D has a significantly larger number of cases and population than region F (5001:1 164 021 and 3638:635 233, respectively), but results in a lower overall rate.

### 4.1.2 Dasymetric Output

The dasymetrically-derived mPTB incidence rates for 2018/19 are first presented in a polygon format (Figure 11) for ease of visual comparison with the above choropleth output. The final grid binary dasymetric output (Figure 12) presents these same rates as constrained to LULC-identified populated areas.

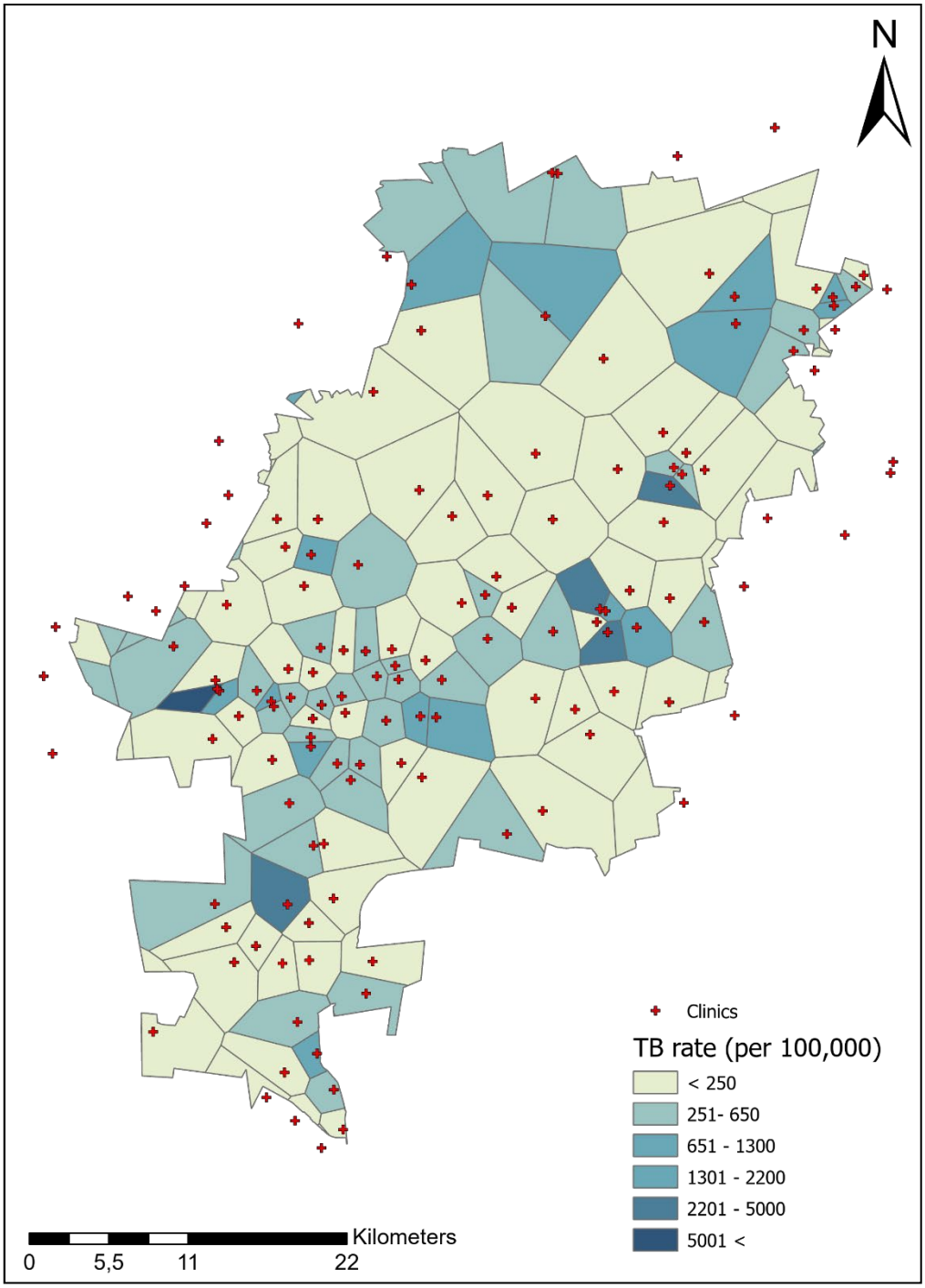



*Figure 11: Dasymetrically-derived mPTB rate (per 100,000 population) for Johannesburg 2018-2019*

The dasymetrically-derive mPTB incidence output reveals a much more nuanced distribution of rates than that of the previously presented choropleth map (Figure 10). Of interest here is the fact that areas of high incidence are more isolated and spread out across the region. There is noticeable clustering of medium and high rates in the north-west, north-east and south-west.

Figure 11 is useful in displaying facility distribution in the region as well as the associated mPTB rates. As previously mentioned, the facilities are well-distributed according to population densities. Thus, the smaller polygons in the south-western region correspond with our previous illustrations of high population densities (Figure 4).

In low-income areas in the north-east and south-west we see that the facilities are closer together, resulting in smaller catchment areas. This gives the visual perception that these areas cater to smaller populations and are less significant, however, the underlying population densities should be kept in mind.

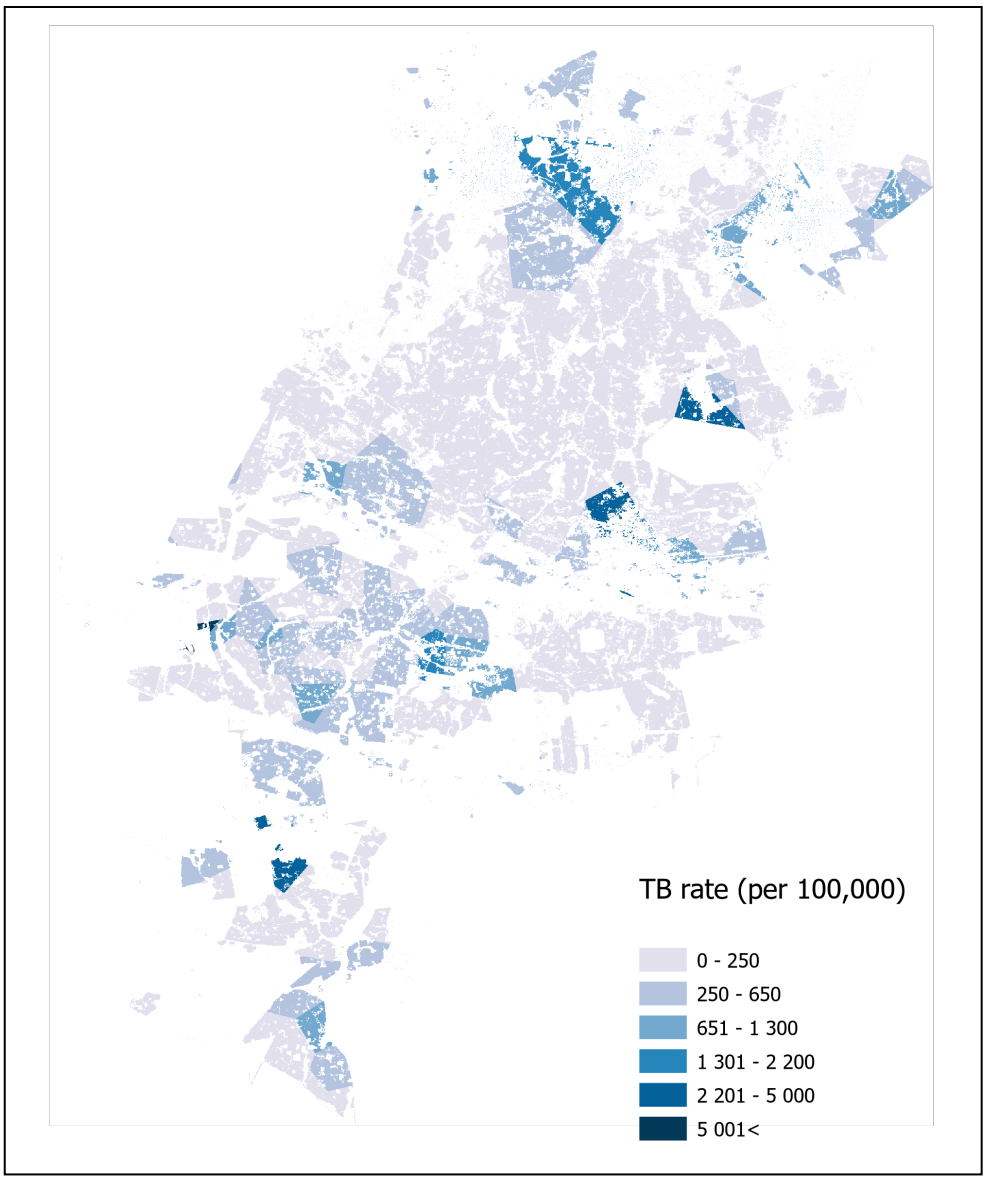



*Figure 12: Dasymetric map of mPTB rate (per 100,000 population) for Johannesburg 2018-2019*

Figure 12 provides an added layer of context to the rates presented in Figure 11, as the rates are limited to the populated areas within each catchment area. High value polygons may seem less daunting when it is realised that interventions should be allocated to these restricted surface areas.

However, this method also reveals areas of incredibly high incidence, with some areas exceeding 5000 cases (per 100,000). These figures may be partially inflated due to inadequate facility catchment area derivation, however, this provides a strong indication that areas of high

incidence are being missed when utilising simple choropleth methods. A quantification of these differences is provided in using measure of variation below.

### 4.1.3 Variance and standard deviation

Through calculation of variance and standard deviation (Table 4.1), it is shown that the dasymetric mPTB rate data for each health sub-district deviates significantly from the average rates suggested by the choropleth representative values. These differences may be interpreted as the information that is lost when using the latter technique which does not consider underlying population distributions and facility catchment areas.

*Table 4.1: Mean values from a choropleth mapping process, and standard deviation of dasymetric values from this mean.*

| | Choropleth value (proxy mean) | Variance | Standard deviation |
|---|---|---|---|
| Johannesburg Health Sub-district A | 431.7 | 73728.7 | 271.5 |
| Johannesburg Health Sub-district B | 162.5 | 34390.3 | 185.4 |
| Johannesburg Health Sub-district C | 152.9 | 33582.0 | 183.3 |
| Johannesburg Health Sub-district D | 429.6 | 42799.7 | 206.9 |
| Johannesburg Health Sub-district E | 281.1 | 51989.9 | 228.0 |
| Johannesburg Health Sub-district F | 572.7 | 66686.0 | 258.2 |
| Johannesburg Health Sub-district G | 348.7 | 22895.2 | 151.3 |

Region F illustrated the highest incidence rates in the choropleth Figure 10, yet here the dasymetric data demonstrates the second highest variance and standard deviation from the choropleth mean. This indicates how non-representative this single measure is for the area. Similarly, the highest standard deviation is seen in Region A, located in the north of Johannesburg. The region has predominantly low population densities but also hosts two

incredibly dense, low-income settlements (Diepsloot and Ivory Park); resulting in significant deviation from the choropleth attempt at a representative value for the region.

Remote sensing imagery has long been considered a revolutionary tool for monitoring and forecasting human diseases (Rush & Vernon, 1975; Hay, 2000). The ability of the grid binary dasymetric method to detect greater variation in the data, as illustrated here, therefore endorses this technique for presenting disease incidence rates.

## 4.2 Spatial Autocorrelation

### 4.2.1 Global results

Firstly, we confirmed what patterns are present in the general data; i.e. whether the pattern was clustered, dispersed or random. This was achieved by measuring global spatial autocorrelation., the results of which are shown below in Table 4.2.

*Table 2.2: Result of Global Moran's I test for spatial autocorrelation*

| Global Moran's I Summary | |
|---|---|
| Moran's Index: | -0,003629 |
| Expected Index: | -0,007194 |
| Variance: | 0,000032 |
| z-score: | 0,634214 |
| p-value: | 0,525941 |

The resultant p-value is not statistically significant, therefore we could not reject the null hypothesis of CSR. This indicates that the overall pattern and trend of the data is not significantly clustered or dispersed. However, this does not necessarily indicate that there are

no isolated clusters, or hot spots. These could instead be detected using local spatial autocorrelation statistics which assess each feature within the context of neighbouring features.

### 4.2.2 Local results

Significant local spatial autocorrelation is presented here by type of association (Figure 13); (1) KNN, (2) Distance threshold, and (3) Inverse distance. The number of statistically significant hot spots, illustrated in red, varied for each ($n$ = 18, 11, and 6, respectively). All three methods revealed no significant cold spots.

In all three cases, there are clear clusters in regions F and D. The inverse distance method detected the fewest number of clusters, but detected two additional hot spots.

The default distance threshold for neighbourhood search in the two distance-related methods (fixed distance band and inverse distance/inverse distance squared) was 5309 metres. This was deemed adequate as it ensured that each neighbour had at least one neighbour and did not result in excessively large neighbourhoods (Grekousis, 2020).

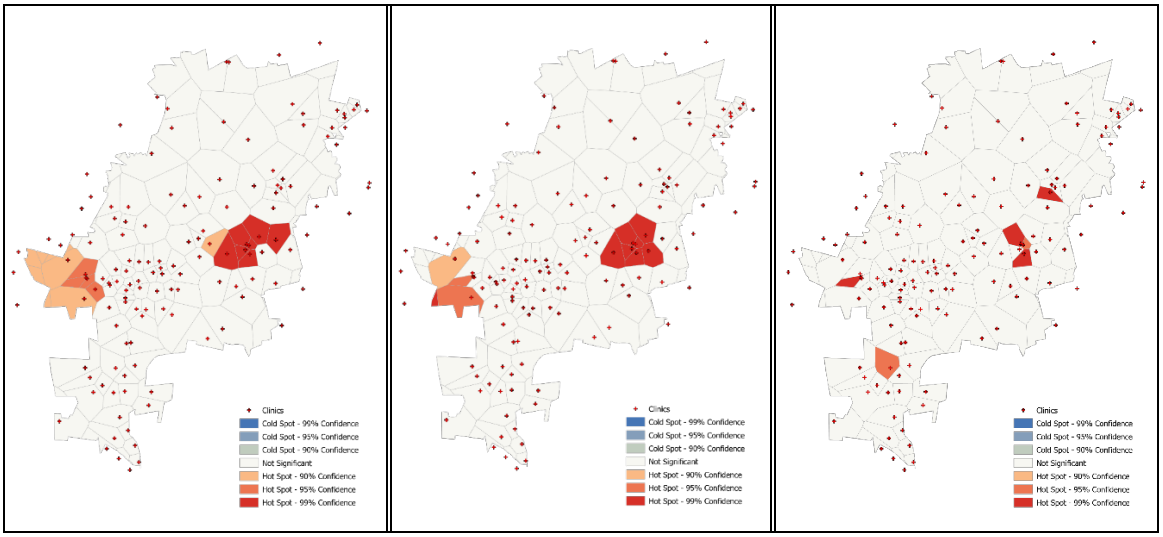



*Figure 13: Hot spot analysis results for (1) KNN, (2) Distance threshold, and (3) Inverse distance & Inverse Distance Squared*

These hot spots refine the dasymetric findings in 4.1.2 and identify catchment area clusters which may account for higher transmission rates within their boundaries and to neighbouring

areas. Targeting these facilities as part of new detection and treatment interventions could have exponential effects.

### 4.2.3 AMOEBA

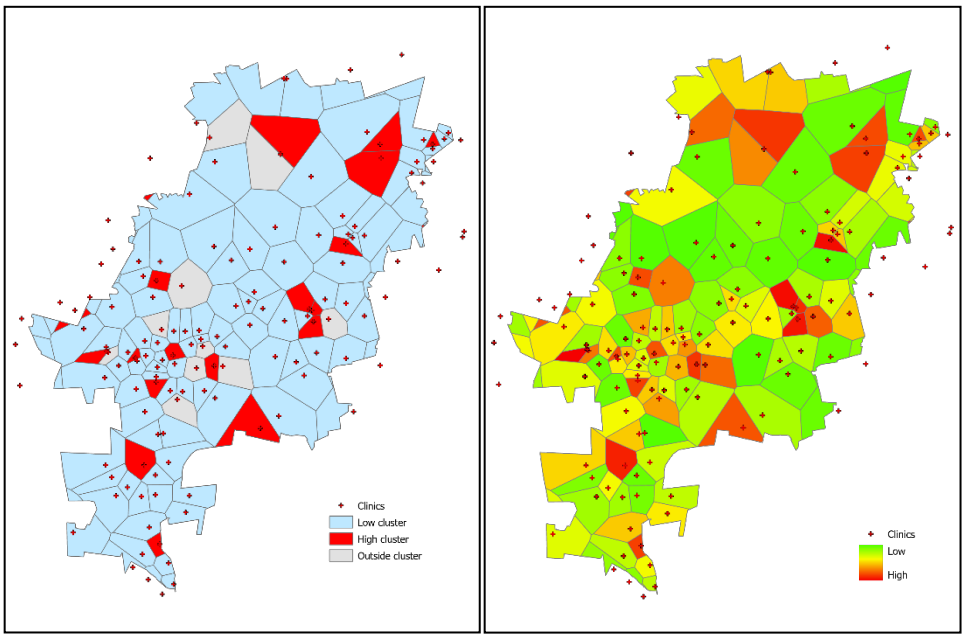



*Figure 14: mPTB incidence clusters determined using the (1) AMOEBA and (2) AMOEBA(+) algorithms*

Figure 14(1) shows the results of a 3-class AMOEBA. 18 hot spots were identified, with the majority of the study area falling into low value clusters. Those values which are closest to the mean of the overall study area do not fall into the low or high cluster classifications and are therefore designated as 'outside' of a cluster (Barro et al., 2015).

Visualisations of clusters is highly controlled by the divisions of the resultant $G^*_{max}$ values as derived for each ecotope. Therefore the AMOEBA+ algorithm was also executed, resulting in a wider spectrum of classifications. The output in Figure 14(2) provides a range of clustering classes from low to high. This visualisation, better illustrated through a stretched colour symbology, divides the $G^*_{max}$ output into 81 value ranges. We can therefore visually identify areas with less extreme highs and lows than required for Figure 14(1). This may be useful in future, for example if one were to create a hierarchy of facilities in need of increased intervention. It is also interesting to note the visualised multi-directional groupings of high and

medium values offer by the AMOEBA(+) algorithm. This is clear in a congregate cluster in the north of the city, as well as in a more serpentine manner in the mid-south region.

# Chapter 5: Discussion

The choropleth and dasymetric results have illustrated the heterogenous nature of mPTB incidence in our study region of Johannesburg to varying degrees. The differences between the outputs are attributed to the scale at which the data has been calculated and presented, as well as the consideration of discrete facility catchment boundaries as explored in the latter.

The choropleth output considered incidence as a health sub-district level with no consideration of true facility service reach. This epitomises the MAUP issue where arbitrary boundaries misrepresent the organic boundaries of phenomena. Whilst it would have been reasonable to have anticipated the highest incidence rates to be located in Region D which contains the greatest number of clinics, some of the highest population densities and the largest concentration of people living below the poverty line, instead, the output suggested that Region F is of greatest concern. Reasons for the choropleth's inherent weaker performance are well-known, with assumptions of homogeneity resulting in true variation being lost. Smaller units of analysis are therefore preferential, however, in this case the approach offered little information on the distribution of mPTB incidence given that there are just 8 health sub-districts in the City of Johannesburg local municipality.

A dasymetric representation of mPTB was then accomplished using a combination of population census data, facility mPTB count data, and LULC data in vector and raster formats. The final output (Figure 12) provides incidence at a health facility service area scale. The incorporation of a binary (residential vs non-residential) LULC classification system, allowed results to be filtered to populated areas. This output still assumes homogeneity, but does so at a much smaller scale. Given the varying densities and presence of large unpopulated areas in Johannesburg, this had a significant impact on the mPTB incidence rates, as similarly seen in

previous literature (Jega, Comber & Tate, 2017). High incidence service areas were identified in every sub-district, with the majority falling within Region D.

In addition to providing more detailed visualisations of incidence, these data are presented in a format that can easily be integrated with putative socioeconomic and environmental factors for further analyses.

The differences between the choropleth and dasymetric methods were quantified using variance and standard deviation to measure the spread of the dasymetric outputs around the choropleth representative values for each health sub-district. Standard deviations were significantly high, varying between 151 and 182 for each sub-district, indicating substantial underlying fluctuations in mPTB incidence within the areal units. Our third objective, to determine optimal locations for dCXR interventions, requires analysis on a meaningful scale, which therefore cannot be addressed through the standard choropleth approach. The dasymetric output was therefore considered.

The resultant, better dasymetric visualisation of geospatial 'risk' of mPTB, calculated from a more realistic, detailed interpretation of implicated population, made it possible to study the spatial clusters of incidence. A global spatial autocorrelation analysis did not reveal any significant pattern of clustering or dispersion when considering the spread of the data in the whole study area. This is understandable given succeeding local spatial autocorrelation outputs which identified a maximum of 18 positively significant polygons (with varying degrees of significance) out of a total of 139. These local indices did, however, identify clusters of significantly positive incidence – supporting a global body of literature on mPTB that consistently enforce that spatial distributions of the disease are not uniform in cities (Queiroga et al., 2012; Shaweno et al., 2018).

These are incredibly useful in answering our study aim of identifying hot spots where dCXR interventions may be targeted. Instead of developing strategies aimed at all facilities in the seemingly most affected health sub-district, these interventions can be more strategically placed throughout the city.

Three standard local hot spot techniques were compared to the cluster morphology AMOEBA (and AMOEBA+) method which has been endorsed in literature as a superior technique for detecting clusters or a non-compact nature. There is spatial agreement between the highest values illustrated in the AMOEBA output (Figure 14) and that of the 6 hot spots identified in the inverse distance output (Figure 13). This indicates that the assumption made by the IDW – that local influence that diminished with distance – may hold true for disease incidence. This makes intuitive sense given the transmissible nature of TB.

The hot spot outputs should be considered alongside the LULC-constrained distribution of mPTB (Figure 9); thus an additional, exploratory overlay was generated using the AMOEBA output and the binary populated raster data (Addendum 2). From here we can identify certain service areas with seemingly small population distributions and resultant high mPTB incidences. Other high incidence areas contain higher densities of residential pixels, representing facilities that, if targeted, could serve a larger audience and have a greater impact.

In theory, it would be possible to carry out AMOEBA on raster data (grids) using the `spdep` package in R to create neighbours via the `cell2nb` command. However, this is only appropriate for smaller datasets (Crago & Aldstadt, 2012; Duque et al., 2010) as the resulting neighbour list is a product of the number of rows and columns (in this case 2294*2859 = 6 558 546). Therefore available computational power presented a limitation to this study, indicating that the full potential of the AMOEBA method requires substantial resources not commonly available in public health settings.

# Chapter 6: Conclusion and Recommendations

The overall aim of this research project was to determine whether spatial interpolation techniques offer a more accurate, detailed representation of TB incidence in a hot spot detection application.

The following objectives were completed as a means to achieve this aim:

- Determine TB incidence using a standard choropleth mapping approach
- Determine TB incidence using an advanced dasymetric technique
- Compare the TB incidence outputs of the choropleth and dasymetric methods
- Identify TB incidence hot spots

After a review of existing international and local literature, a dasymetric form of areal interpolation was identified as the most appropriate technique given the available data and the research problem that required transferring population data to different spatial units. This method was presented as an alternative to conventional TB disease choropleth mapping.

Consequently, the study illustrated two methods for presenting disease incidence data: a standard choropleth method, and a dasymetric mapping method. In the latter case, a focus was placed on creating an accurate and detailed depiction of underlying population distributions in order to calculate TB incidence at a finer scale. This required a diversity of datasets, some of which may not usually be associated with disease mapping. LULC raster and vector datasets, derived from satellite imagery, provided auxiliary information identifying areas that are most likely populated.

The results from the dasymetric technique were then compared to values derived from the choropleth process, considering location variance in the form of standard deviations. From the statistical output, it could be concluded that the choropleth mapping technique vastly under-

captures the distribution and variation in TB incidence rates across Johannesburg. In directly answering the research aim, it can be said that the dasymetric spatial interpolation technique provides an alternative, more detailed representation of TB. The accuracy of the output is implied by literature that indicates that the majority of populations use their nearest facility (Stats SA, 2016; Tanser et al., 2001; Tanser, Gijsbertsen & Herbst, 2006). The true accuracy of the output would require additional ground-truthing data in the form of *actual* clinic usage.

Consideration was then given to spatial autocorrelation in the data. The choropleth sub-district depiction of incidence was not of a fine enough scale to warrant a hot spot analysis. Instead, it was possible to visually identify which sub-districts were of concern directly from the mapping output. The dasymetric method, however, provided a detailed account of TB incidence across the study area. Although a significant spatial pattern was not present at a global scale, hot spots were detected using a local autocorrelation index, applied through various models. It was not possible to perform the analysis on the final raster output due to computational limitations, therefore significant clusters were identified using the facility catchment area polygons.

### 4.3 Limitations of the study

1. Thiessen-polygon-generated proxy facility catchment areas were used to define the area from which patients are drawn. This assumed that people utilise their nearest facility based on simple Euclidean distance from facilities. A better understanding of utilisation and preference, using qualitative and quantitative information, could greatly improve the accuracy of population allocations to clinics. More accurate catchment areas could also be explored, e.g. utilising routable transport network data, however this may not improve results as substantially as one may think (as seen in Tanser et al. 2006, previously discussed).
2. For the purposes of this research, only clinics were considered. Further research required for integrating data from other types of facilities.

3. Further investigation is required regarding quality of the census data known to be less accurate in metropolitan areas, given their complex nature, calling its quality into question (Weir-Smith, 2016).
4. This study did not consider any underlying variables, apart from population, when determining TB distribution. This therefore makes it an incomplete study if wanting to identify optimal locations for implementing dCXR initiatives. Dasymetric mapping methods do cater for additional auxiliary data, e.g. through combining the LULC data with more detailed information relating to socio-economic characteristics of the populated pixels (e.g. informal vs formal, rural vs urban, high density vs low density).

### 4.4 Recommendations and future research

The findings of this study provide provisional recommended facilities for targeted dCXR interventions. Final locations should be determined using methods that solve the limitations, as listed above.

On a more general note, it is recommended that standard TB spatial monitoring practices be reviewed to ensure that the information provided to policy makers is at the scale required for making informed, strategic decisions.

The dasymetrically determined population raster, one of the outputs of this project, provides a foundation upon which analysis of a phenomenon may be carried out almost any scale. This means that further investigations into human and environmental factors that influence or are correlated with mPTB, such as income, household size, population density may be included in models. Utilization of multivariate auxiliary datasets in this way may be explored through intelligent dasymetric mapping.

# References


Aldstadt, J., & Getis, A. (2006) *Using AMOEBA to create a spatial weights matrix and identify spatial clusters*, Geographical Analysis, 38(4), 327–343, https://doi.org/10.1111/j.1538-4632.2006.00689

Álvarez-Hernández, G., Lara-Valencia, F., Reyes-Castro, P. A., & Rascón-Pacheco, R. A. (2010) *An analysis of spatial and socio-economic determinants of tuberculosis in Hermosillo, Mexico, 2000-2006*, International Journal of Tuberculosis and Lung Disease, 14(6), 708–713.

Barro, A. S., Kracalik, I. T., Malania, L., Tsertsvadze, N., Manvelyan, J., Imnadze, P., & Blackburn, J. K. (2015) *Identifying hotspots of human anthrax transmission using three local clustering techniques*, Applied Geography, 60, 29–36. https://doi.org/10.1016/j.apgeog.2015.02.014

Barrozo, L. V., Pérez-Machado, R. P., Small, C., & Cabral-Miranda, W. (2016) *Changing spatial perception: dasymetric mapping to improve analysis of health outcomes in a megacity*, Journal of Maps, 12(5), 1242–1247. https://doi.org/10.1080/17445647.2015.1101403

Bentley, G. C., Cromley, R. G., & Atkinson-Palombo, C. (2013) *The network interpolation of population for flow modelling using dasymetric mapping*, Geographical Analysis, 45(3), 307–323. https://doi.org/10.1111/gean.12013

Bian, L., & Butler, R. (1999) *Comparing Effects of Aggregation Methods on Statistical and Spatial Properties of Simulated Spatial Data*, Photogrammetric Engineering & Remote Sensing, 65(1), 73–84.

Boffa, J., Moyo, S., Chikovore, J., Salomon, A., Jooste, S., Mkhombo, T., Pai, M., & Daftary, S. (2019) *Why we need to evaluate the quality of tuberculosis care in South Africa's private health sector*, SAMJ Correspondence, SAMJ, November 2019, Volume 109, No 11.

Castellano, G. V., Lopez-Quilez, A., Perez-Vicente, R., Merlo, J. (2013) *AMOEBA+ with R*, The R User Conference, useR! 2013, University of Castilla-La Mancha, Albacete, Spain.

Child, K. (2016) *Private TB patients have to be treated in State*, Sunday Times [newspaper], 22 February 2016, South Africa.

Churchyard, G. J., Mametja, L. D., Mvusi, L., Ndjeka, N., Hesseling, A. C., Reid, A., & Pillay, Y. (2014) *Tuberculosis control in South Africa: Successes, challenges and recommendations*, South African Medical Journal, 104(3), 244–248. https://doi.org/10.7196/SAMJ.7689

Coetzee, S., Cooper, A. K., & Katumba, S. (2020) *Strengthening Governance in the Gauteng City-Region through Spatial Data Infrastructure: the case of address data*, GCRO, Governing the GCR Provocation Series #5, May 2020.

Comber, A., & Zeng, W. (2019). *Spatial interpolation using areal features : A review of methods and opportunities using new forms of data with coded illustrations*. (June), 1–23. https://doi.org/10.1111/gec3.12465

Council for Medical Schemes (2017) *Tuberculosis*, CMScript, issue 2, 2017, available from: https://www.medicalschemes.com/files/CMScript/CMScript2Of2017.pdf

Council for Medical Schemes (2019) *Annual Report 2018/19*, Council for Medical Schemes, Pretoria, South Africa, Available from: https://www.medicalschemes.com/files/Annual%20Reports/CMSAR2018_19.pdf

Crankshaw, O. & Parnell, S. (2004) *Johannesburg: race, inequality and urbanisation*. In J. Gugler (Ed.), World cities beyond the west: Globalisation, development and inequality, Cambridge: Cambridge University Press.

Cromley, E.K. , & McLafferty, S. (2002). *GIS and public health*. New York: Guilford.

Department of Health (DoH) (2020). *The First National TB Prevalence Survey; South Africa 2018*.

Dowdy, D. W., Golub, J. E., Chaisson, R. E., & Saraceni, V. (2012) *Heterogeneity in tuberculosis transmission and the role of geographic hotspots in propagating epidemics*, Proceedings of the National Academy of Sciences of the United States of America, 109(24), 9557–9562. https://doi.org/10.1073/pnas.1203517109

Duque, J. C., Aldstadt, J., Velasquez, E., Franco, J. L., & Betancourt, A. (2011) *A computationally efficient method for delineating irregularly shaped spatial clusters*, Journal of Geographical Systems, 13, 355–372. https://doi.org/10.1007/s10109-010-0137-1

Elliott P., Wakefield J.C., Best N.G., Briggs D.J., eds. (2000) *Spatial Epidemiology: Methods and Applications*, Oxford: Oxford University Press

Eicher, C. L., and Brewer, C. A. (2001) *Dasymetric mapping and areal interpolation: Implementation and evaluation*, Cartography and Geographic Information Science 28 (2):125–38.

Eskom (2017) *Eskom Settlement Growth Dataset Standards*.

ESRI (2016) *Creating Thiessen polygons from points*, Environmental Systems Research Institute Inc., https://desktop.arcgis.com/en/arcmap/10.3/tools/coverage-toolbox/how-thiessen-works.htm

ESRI (n.d.) *Modeling spatial relationships*, https://pro.arcgis.com/en/pro-app/latest/tool-reference/spatial-statistics/modeling-spatial-relationships.htm

Everatt, D. (2014) *Poverty and inequality in the Gauteng city-region*. In Harrison P., Gotz G., Todes A., & Wray C. (Eds.), Changing Space, Changing City: Johannesburg after apartheid - Open Access selection (pp. 63-82). Johannesburg: Wits University Press. doi:10.18772/22014107656.7

Goodchild, M. F., & Lam, N. S. (1980) *Areal interpolation: A variant of the traditional spatial problem*, Geo-Processing, 1, 297–312.

Green, C.A., Mans, G., Le Roux, A., Schmitz, P., Mokgalaka, H.M., McKelly, D., Ngidi, M., Badenhorst, W. and Zietsman, H.L. (2012) *Geographic Accessibility Study of Social Facility and Government Service Points for the Metropolitan Cities of Johannesburg and eThekwini 2011/12*, CSIR/BE/SPS/ER/2012/0061/B, CSIR: Built Environment.

Grekousis, G. (2020*) Think Spatially: Basic Concepts of Spatial Analysis and Space Conceptualization*, in Spatial Analysis Methods and Practice: Describe – Explore – Explain through GIS (pp. 1-58). Cambridge: Cambridge University Press. doi:10.1017/9781108614528.002

Grubesic, T. H., Wei, R., Murray, A. T., Grubesic, T. H., Wei, R., Murray, A. T., & Murray, A. T. (2014) *Spatial Clustering Overview and Comparison : Accuracy , Sensitivity , and Computational Expense Spatial Clustering Overview and Comparison : Accuracy , Sensitivity , and Computational Expense*. *5608*. https://doi.org/10.1080/00045608.2014.958389

Hallisey, E., Tai, E., Berens, A., Wilt, G., Peipins, L., Lewis, B., … Lunsford, N. B. (2017) *Transforming geographic scale: A comparison of combined population and areal weighting to other interpolation methods*, International Journal of Health Geographics, 16(1), 1–16. https://doi.org/10.1186/s12942-017-0102-z

Harris, B., Goudge, J., Ataguba, J. E., McIntyre, D., Nxumalo, N., Jikwana, S., & Chersich, M. (2011) *Inequities in access to health care in South Africa*, Journal of Public Health Policy, Vol 32, S1, 102-123

Harrison P., Gotz G., Todes A., & Wray C. (2014) *Materialities, subjectivities and spatial transformation in Johannesburg*. In Harrison P., Gotz G., Todes A., & Wray C. (Eds.), Changing Space, Changing City: Johannesburg after apartheid - Open Access selection (pp. 63-82). Johannesburg: Wits University Press. doi:10.18772/22014107656.7

Hay, S. I. (2000) *An overview of remote sensing and geodesy for epidemiology and public health application*, Adv Parasitol. 2000;47:1-35. doi: 10.1016/s0065-308x(00)47005-3. PMID:

10997203; PMCID: PMC3164799.

Hay, S. I., Noor, A. M., Nelson, A., & Tatem, A. J. (2005) *Europe PMC Funders Group: The accuracy of human population maps for public health application*, Trop Med Int Health, 10(10), 1073– 1086. https://doi.org/10.1111/j.1365-3156.2005.01487

Hodgson, D. L. (2017) *Demographic Change in the Upper Tsitsa Catchment: The Integration of Census and Land Cover Data for 2001 and 2011* (Master's thesis, Rhodes University, South Africa). Retrieved from: http://vital.seals.ac.za:8080/vital/access/manager/PdfViewer/vital:26870/SOURCE1?viewPdfInternal=1

Housing Development Agency (HDA) (2012) *South Africa: Informal settlements status*, Research Reports, Johannesburg, 1-64, accessed at: http://www.thehda.co.za/uploads/files/HDA_Informal_settlements_status_South_Africa.pdf

Ibrahim, S., Hamisu, I., & Lawal, U. (2015) *Spatial Pattern of Tuberculosis Prevalence in Nigeria : A Comparative Analysis of Spatial Autocorrelation Indices*, American Journal of Geographic Information System, *4*(3), 87–94. https://doi.org/10.5923/j.ajgis.20150403.01

Jankowska, M., Aldstadt, J., Getis, A., Weeks, J., & Fraley, G. (2008) *An amoeba procedure for visualizing clusters*, GIScience, (May 2014).

Kapwata T, Morris N, Campbell A, Mthiyane T, Mpangase P, Nelson KN, et al. (2017) *Spatial distribution of extensively drug-resistant tuberculosis (XDR TB) patients in KwaZulu-Natal, South Africa*. PLoS ONE 12(10): e0181797. https://doi.org/10.1371/journal.pone.0181797

Kiani, B., Raouf Rahmati, A., Bergquist, R., Hashtarkhani, S., Firouraghi, N., Bagheri, N., Moghaddas, E. & Mohammadi, A. (2021) *Spatio-temporal epidemiology of the tuberculosis incidence rate in Iran 2008 to 2018*. BMC Public Health 21, 1093. https://doi.org/10.1186/s12889-021-11157-1

Kracalik, I. T., Blackburn, J. K., Lukhnova, L., Pazilov, Y., & Hugh-, M. E. (2012) *Analysing the spatial patterns of livestock anthrax in Kazakhstan in relation to environmental factors : a comparison of local (Gi*) and morphology cluster statistics*, Geospatial Health, 7(1), 111–126.

Kootbodien, T., Iyaloo, S., Wilson, K., Naicker, N., Kgalamono, S., Haman, T., Mathee, A., & Rees, D. (2019) *Environmental Silica Dust Exposure and Pulmonary Tuberculosis in Johannesburg*, South Africa. International journal of environmental research and public health, 16(10), 1867. https://doi.org/10.3390/ijerph16101867

Lam, N. S. (1980) *Spatial Interpolation Methods: A Review*, The American Cartographer, 10:2, 129-150

Langford, M. (2006) *Obtaining population estimates in non-census reporting zones: an evaluation of the 3-class dasymetric method*, Comput Environ Urban Syst. 30:161–80

Lawson, A.B. (2018) *Bayesian Disease Mapping: Hierarchical Modeling in Spatial Epidemiology*, Third Edition (3rd ed.),Chapman and Hall/CRC, https://doi.org/10.1201/9781351271769

Loveday, M., Mzobe, Y.N., Pillay, Y., & Barron, P. (2019) *Figures of the dead: A decade of tuberculosis mortality registrations in South Africa*, South African Medical Journal = Suid-Afrikaanse Tydskrif Vir Geneeskunde, 109(10), 728–732. https://doi.org/10.7196/SAMJ.2019.v109i10.14073

Lurie, M.N., & Williams, B. G. (2014) *Migration and health in Southern Africa : 100 years and still circulating*, Health Psychology & Behavioural Medicine, *2*(1), 34–40. https://doi.org/10.1080/21642850.2013.866898

McIntosh AI, Jenkins HE, White LF, Barnard M, Thomson DR, et al. (2018) *Using routinely collected laboratory data to identify high rifampicin-resistant tuberculosis burden communities*

*in the Western Cape Province, South Africa: A retrospective spatiotemporal analysis*. PLOS Medicine 15(8): e1002638. https://doi.org/10.1371/journal.pmed.1002638

Melo, V. (2016) *The Production of Urban Peripheries For and By Low-Income Populations at the Turn of the Millennium: Maputo, Luanda and Johannesburg*, Journal of Southern African Studies, 42:4, 619-641, DOI: 10.1080/03057070.2016.1196955

Mennis, J. (2003) *Generating Surface Models of Population Using Dasymetric Mapping*, The Professional Geographer, 55:1, 31-42, DOI: 10.1111/0033-0124.10042

Mitchell, A. (1999) *The ESRI Guide to GIS Analysis*. New York: ESRI Press.

Mokgalaka, H., Mans, G., Smit, J. & McKelly, D. (2013) *GIS-based accessibility analysis- a mixed method approach to determine public primary health care demand in South Africa*, Proceedings Global Geospatial Conference (Combined Africa GIS and GSDI 14 Conference), Addis Ababa, Ethiopia, November 2013.

Mokgalaka, H. (2014) *GIS-based accessibility analysis: An approach to determine public primary health care demand in South Africa*. Global Spatial Data Infrastructure Association. http://hdl.handle.net/10204/7323

Mokgalaka, H., Smit, J., & Mans, G. (2015) *GIS-based analysis of spatial accessibility: An approach to determine public primary healthcare demand in metropolitan areas*, Department of Geomatics, 85.

Mubiwa, B. & Annegarn, H. (2013) *Historical spatial change in the Gauteng City-Region*, Occasional paper 04, Gauteng City-Region Observatory (GRCO), Johannesburg.

NDoH (2011) *National core standards for health establishments in South Africa*, Pretoria: National Department of Health.

Nelson, J.K. & Brewer, C.A. (2017) *Evaluating data stability in aggregation structures across spatial scales: revisiting the modifiable areal unit problem*, Cartography and Geographic Information Science, 44:1, 35-50, DOI: 10.1080/15230406.2015.1093431

Ngidi, M., Mans, G., McKelly, D., & Sogoni, Z. (2017) *Using a hybrid methodology of dasyametric mapping and data interpolation techniques to undertake population data (dis)aggregation in South Africa*, South African Journal of Geomatics, 6(2), 232. https://doi.org/10.4314/sajg.v6i2.8

Nteta T. P., Mokgatle-Nthabu M., & Oguntibeju, O. O. (2010) *Utilization of the Primary Health Care Services in the Tshwane Region of Gauteng Province, South Africa*. PLoS ONE 5(11): e13909. https://doi.org/10.1371/journal.pone.0013909

Obaromi, D., Ndege, J., & Yongsong, Q. (2019) *Disease mapping of tuberculosis prevalence in Eastern Cape Province , South Africa*. 241–248.

Pieterse, E., and Owens, K. (2018) *Johannesburg: Confronting Spatial Inequality*, World Resources Report Case Study, Washington, DC: World Resources Institute. Available online at www. citiesforall.org.

Queiroga, R.P.F. de, Sá, L.D. de, Nogueira, J. de A., Lima, E.R.V. de, Silva, A.C.O., Pinheiro, P.G.O. D., & Braga, J.U. (2012) *Spatial distribution of tuberculosis and relationship with living conditions in an urban area of Campina Grande – 2004 to 2007*, Revista Brasileira de Epidemiologia, 15(1), 222–232.

Richards, T. B., Berkowitz, Z., Thomas, C. C., Foster, S. L., Gardner, A., King, J. B., Royalty, J. (2010) *Choropleth map design for cancer incidence, part 1*, Preventing Chronic Disease Public Health Research, Practice and Policy, 7(1), 1–6.

Rush, M. & Vernon, S. (1975) *Remote Sensing and Urban Public Health*, Photogrammetric Engineering and Remote Sensing, Vol. 41, No.9, September 1975, pp.1149-1155

Rushton, G. (2003) *Public Health, GIS, and Spatial Analytical Tools*, Annual Review of Public Health, 43–56. https://doi.org/10.1146/annurev.publhealth.24.012902.140843

SAHO (2011) *Johannesburg the Segregated city*, South African History Online, https://www.sahistory.org.za/article/johannesburg-segregated-city#Town%20&%20regional%20planning%20in%20Johannesburg

Schoeman, F., Newby, T.S., Thompson, M. W., & Van den Berg, E.C. (2013) *South African National Land-Cover Change Map*, South African Journal of Geomatics, *2*(2), 94-105–105. https://doi.org/10.4314/sajg.v2i2

Shaweno, D., Karmakar, M., Alene, K.A., Ragonnet, R., Clements, A.C., Trauer, J.M., & McBryde, E. S. (2018) *Methods used in the spatial analysis of tuberculosis epidemiology: A systematic review*, BMC Medicine, 16(1), 1–18. https://doi.org/10.1186/s12916-018-1178-4

Sinai, I., Cleghorn, F., & Kinkel, H.F. (2018) *Improving management of tuberculosis in people living with HIV in South Africa through integration of HIV and tuberculosis services : a proof of concept study*, BMC Health Services Research, 18(711), https://doi.org/10.1186/s12913-018-3524-9

Sleeter, R. (2004) *Dasymetric mapping techniques for the San Francisco Bay Region, California*, Urban and Regional Information Systems Association, Annual Conference, Proceedings. Reno, Nevada: November 2004.

Stallworthy, G., Dias, H.M., & Pai, M. (2020) *Quality of tuberculosis care in the private health sector,* Journal of Clinical Tuberculosis and Other Mycobacterial Diseases, Volume 20, August 2020, 100171.

Stats SA (2011) *Statistical Release (Revised): Census 2011*, Statistics South Africa, Pretoria, South Africa, available at https://www.statssa.gov.za/publications/P03014/P030142011.pdf

Stats SA (2016) *General Household Survey, 2016* [Online], www.statssa.gov.za

Steiner, P., & Paulus, G. (2007) *Dasymetric mapping for public health planning*, 10th AGILE International Conference on Geographic Information Science, 1–10.

Tadesse, S., Enqueselassie, F., & Gebreyesus, S.H. (2018) *Estimating the spatial risk of tuberculosis distribution in Gurage zone, southern Ethiopia: A geostatistical kriging approach*, BMC Public Health, 18(1), 1–10. https://doi.org/10.1186/s12889-018-5711-3

Talbot T.O., Kulldorff M., Forand S.P., Haley V.B. (2000) *Evaluation of spatial filters to create smoothed maps of health data*, Statistics in Medicine, 19:2399–408

Tanser, F. & Wilkinson, D. (1999) *Spatial implications of the tuberculosis DOTS strategy in rural South Africa: a novel application of geographical information system and global positioning system technologies*, Tropical Medicine and International Health, 1999 Oct;4(10):634-8. doi: 10.1046/j.1365-3156.1999.00469.x. PMID: 10583895.

Tanser, F., Hosegood, V., Benzler, J., & Solarsh, G. (2001) *New approaches to spatially analyse primary health care usage patterns in rural South Africa*. Tropical Medicine and International Health, 6(10), 826–838.

Tanser, F. (2006) *Methodology for optimising location of new primary health care facilities in rural communities: a case study in KwaZulu- Natal, South Africa*, J Epidemiol Community Health 2006;60: p 846–850. doi: 10.1136/jech.2005.043265

Tanser, F., Gijsbertsen, B., & Herbst, K. (2006) *Modelling and understanding primary health care accessibility and utilization in rural South Africa: an exploration using a geographical information system*, Social Science and Medicine, 2006 Aug;63(3):691-705. doi: 10.1016/j.socscimed.2006.01.015. Epub 2006 Mar 29. PMID: 16574290.

Tao, R., & Thill, J.C. (2019) *flowAMOEBA: Identifying Regions of Anomalous Spatial Interactions*, Geographical Analysis, 51(1), 111–130, https://doi.org/10.1111/gean.12161

TB Diagnostics Market Analysis Consortium. (2015) Market assessment of tuberculosis diagnostics in South Africa, 2012 - 2013. International Journal for Tuberculosis and Lung Disease 2015;19(2):216-222. https://doi.org/10.5588/ijtld.14.0565

Thompson, M. (2019) *DEA E1434 Land-Cover: South African National Land-Cover 2018 Report & Accuracy Assessment*, public release report prepared by GeoTerraImage SA Pty Ltd and CSIR on behalf of Department of Environmental Affairs and Department of Rural Development and Land Reform.

Tobler, W.R. (1979) *Smooth pycnophylactic interpolation of geographical regions*, Journal of the American Statistical Association 74, 519–530.

Weir-Smith, G. (2016) *Changing boundaries: Overcoming modifiable areal unit problems related to unemployment data in South Africa*. South African Journal of Science, 112(3-4), 1-8. https://dx.doi.org/10.17159/sajs.2016/20150115

WHO (2019) *Global Tuberculosis Report*, Geneva: World Health Organization; 2019. Licence: CC BY-NC-SA 3.0 IGO

WHO (2020) *Rapid communication on systematic screening for tuberculosis*, Geneva: World Health Oragnisation; 2020, Licence: CC BY-NC-SA 3.0 IGO.

Wekesa, E., Mulenga, H., Luabeya, A.K.K., Shenje, J., & Hatherill, M. (2020) *Regional changes in tuberculosis disease burden among adolescents in South Africa ( 2005 – 2015 ),* 1–12. https://doi.org/10.1371/journal.pone.0235206

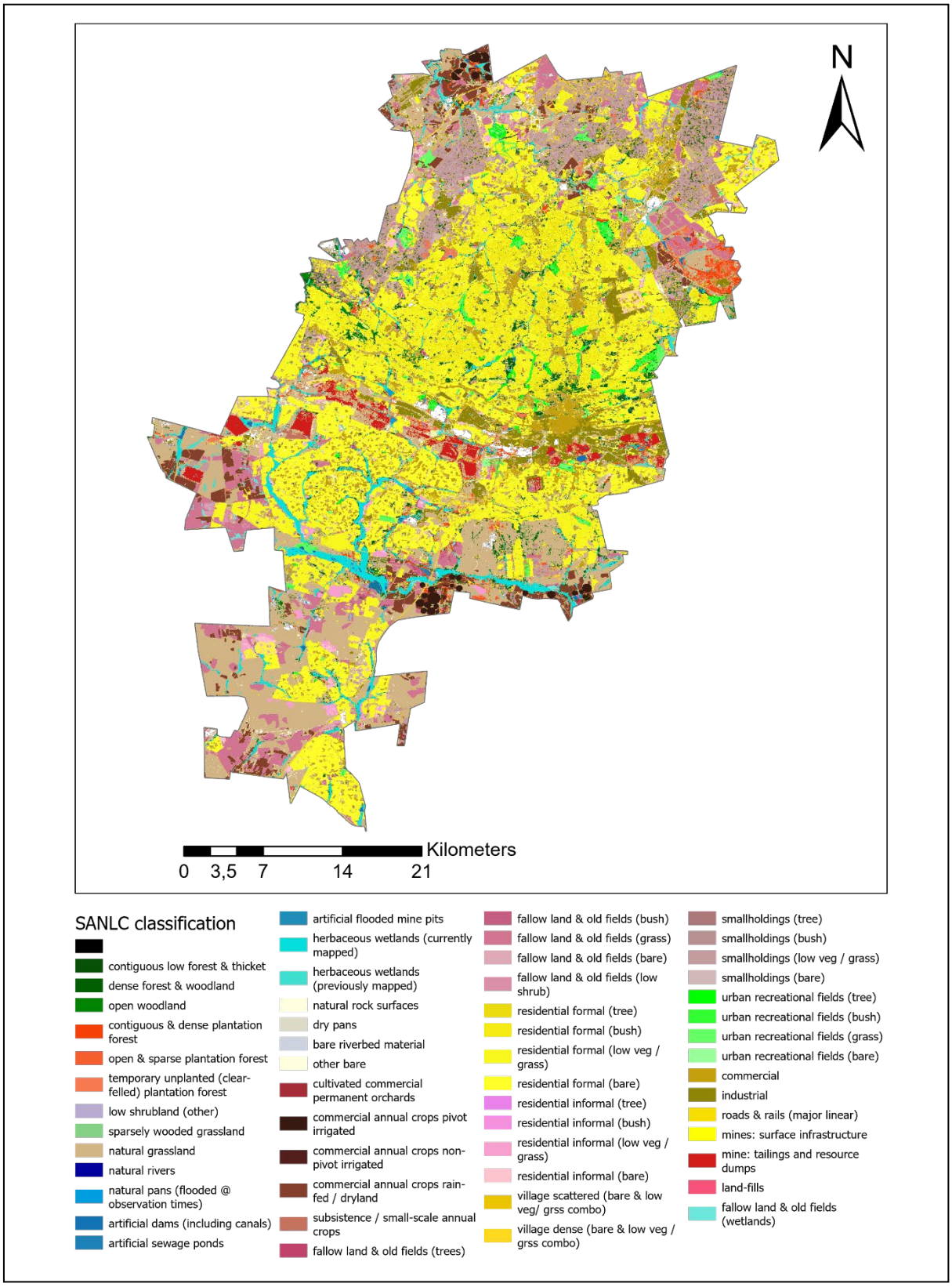
Wilkinson, D., & Tanser, F. (1999) *GIS/GPS to document increased access to community-based treatment for tuberculosis in Africa*. Geographic information system/global positioning system. Lancet, 354(9176), 394–395.

Wright, J.K. (1936) *A method of mapping densities of population: With Cape Cod as an example*, Geographical Review, 26(1), 103-110.

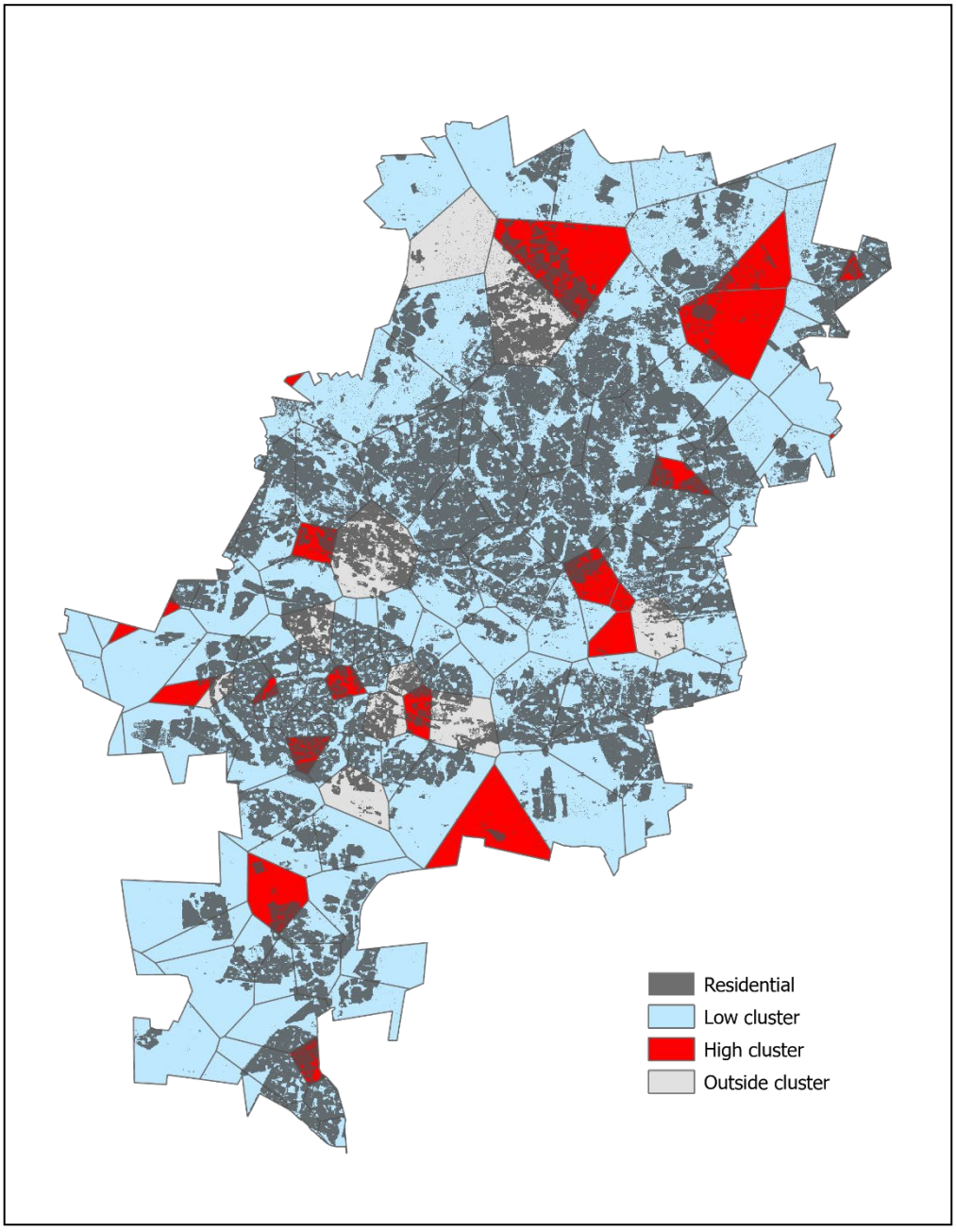
Wu, H., & Li, Z. L. (2009) *Scale issues in remote sensing: a review on analysis, processing and modelling*, Sensors (Basel, Switzerland), 9(3), 1768–1793. https://doi.org/10.3390/s90301768

# Addenda

## 1. SANLC Classification Map



## 2. AMOEBA overlayed with residential pixels



## 3. Ethical Clearance Certificate

UNIVERSITY OF THE WITWATERSRAND, JOHANNESBURG

**SCHOOL OF GEOGRAPHY, ARCHAEOLOGY & ENVIRONMENTAL STUDIES ETHICS COMMITTEE**

**CONSTITUTED UNDER THE UNIVERSITY HUMAN RESEARCH ETHICS COMMITTEE (NON-MEDICAL)**

**CLEARANCE CERTIFICATE** **PROTOCOL NUMBER: W-GAES-2021/GEOG-02**

**PROJECT TITLE**: Dasymetric Estimation of Population Density and Areal Interpolation of Facility-Level Tuberculosis Data

| | |
|---|---|
| **INVESTIGATOR** | Ms Beth Crankshaw |
| **SCHOOL/DEPARTMENT OF INVESTIGATOR** | Geography, Archeology and Environmental Studies |
| **DATE CONSIDERED** | 23 March 2021 |
| **DECISION OF THE COMMITTEE** | Approved unconditionally |
| **RISK LEVEL** | NO RISK |
| **EXPIRY DATE** | 30th December 2021 |

**ISSUE DATE OF CERTIFICATE**: 23rd March 2021 **CHAIRPERSON:**
(Professor Mulala Danny Simatele)

cc: Supervisor:

**DECLARATION OF INVESTIGATOR**

To be completed in duplicate and **ONE COPY** returned to the Chairperson of the School/Department ethics committee.

I fully understand the conditions under which I am are authorized to carry out the abovementioned research and I guarantee to ensure compliance with these conditions. Should any departure to be contemplated from the research procedure as approved I/we undertake to resubmit the protocol to the Committee.

**Signature** **Date** 23/03/2021

**PLEASE QUOTE THE PROTOCOL NUMBER ON ALL ENQUIRIES**