# QOL REVIEW: SPATIAL PATTERNS, ANALYSIS AND LIMITATIONS

***Gillian Maree and Bonolo Mohulatsi, Gauteng City-Region Observatory***

Briefing document

This document was prepared as part of the Gauteng City-Region Observatory's Quality of Life survey ten year review process. It is a lightly edited version of the document provided to participants in preparation for the sampling workshop hosted on 14 February 2019.

Written by: Gillian Maree and Bonolo Mohulatsi



Published by the Gauteng City-Region Observatory (GCRO), a partnership of the University of Johannesburg, the University of the Witwatersrand, Johannesburg, the Gauteng Provincial Government and organised local government in Gauteng (SALGA).

Recommended citation: Maree, G. and Mohulatsi, B. (2020) QoL review: Spatial patterns, analysis, and limitations. Johannesburg: Gauteng City-Region Observatory. Available at https://gcro.ac.za/research/project/detail/quality-life-survey-10-year-review/.

## TABLE OF CONTENTS

# 1 Introduction

One of the goals of the QoL survey is to be able to profile areas, wards or municipalities and quantify their performance in relation to other areas. These results are often used as Maps of the Month and in many of the GCRO publications. The maps support government decision making and evidence based policy and often help target areas for intervention.

The choice of the QoL to be statistically representative at ward level has implications for spatial planning as wards are based on the number of people within an area meaning wards vary hugely in size (can we add the largest and smallest area). This presents a number of considerations for spatial analysis and the types of conclusions that can be inferred. The spatial distribution of QoL data collect is a factor that should be considered when mapping for doing spatial analysis.

Spatial statistics play an important role when mapping data points and visualising the data. The fundamental assumptions that we make with classical statistics, no longer hold for spatial data. Issues around spatial characterisation, spatial dependency/autocorrelation, scaling, sampling and boundary issues are all aspects of spatial statistics that are crucial to better understanding spatial analysis and the relationships that are derived from such. With spatial statistics, one is able to look at both the local and the global properties of the collected data, where the global properties apply to the entire dataset and local properties relate to each observed unit.

This paper looks at the spatial form of the QoLV data and unpacks some of the implication for spatial analysis. There are two key considerations here, first the implications of concentrated sampling in core areas and second the implications of mapping data at a ward level. No matter the choice of sampling method there will always be implications for mapping. This paper focuses on the implications of sampling for evidence based planning, rather than recommendations for changing the sampling method.

# 2 Overview of the spatial form of the QoL data

The QoL data is surveyed to be statistically relevant at a ward level. While the target interview sites are randomly selected the number of interviews per ward is set at 30 per ward in non-metropolitan wards and a minimum of 50 per metropolitan wards. This means that the data collection favours the more densely populated areas. The Municipal Demarcation Board (MDB) delineates wards for South Africa based on a formula for the number of councillors for each municipality which is set out in section 18 of the Municipal Structures Act, 1998. Once the number of councillors is published, this information is used by the MDB to calculate the number of wards by dividing the number of councillors for the municipality by two. A norm for the number of registered voters per ward is then calculated by dividing the number of registered voters in the municipality by the number of ward councillors. The legislation allows for a deviation of 15% above or below the norm. These numbers are then used during the spatial configuration of ward boundaries to ensure that each ward in a municipality has more or less the same number of registered voters, as required by Schedule 1 to the Structures Act, 1998.

Ideally, for data analysis it would be preferable to retain stable ward boundaries, it is in general not possible due to an increase or decrease in the number of registered voters which impacts on the number of councillors, and thus on the number of wards. In Gauteng the number of wards in 2006 was 423 and in 2016 there were 529, so there are 106 new wards to consider across the decade of QoL data collection.

Spatially, this means that the area of a ward varies considerably in Gauteng, with the smaller wards in the dense urban areas and larger wards on the periphery. The smallest ward is only 22ha (Central Joburg around Joubert Park, Ward 79800059) and the largest over 108 000 ha (Tshwane - Roodeplaat Ward ID 79900099, 56 sample points).

## 3 Implications for spatial analysis

If spatial analysis of QoLV data is done by a spatial area (e.g. ward level) there are considerations that need to be account for  a result of the data distribution across and within wards. This is a particular problem when ranking or comparing areas to one another as is often the case with QoL indicators.

### *3.1 Modifiable unit area problem (MAUP)*

MAUP refers to a distorted or changed picture that is caused by varying sizes or areas of demarcations of spatial units (like wards) or changing/aggregating the scale of analysis (Naude 2008, Openshaw, 1984).  This has been noted in spatial analysis in South Africa before, the NSDP mapping work referred to this as the 'Gordonia problem' (Naude et al, 2008). A ward is not a geographic unit of analysis, but rather a political one and this has implications for the zone and scale of analysis.  The scale problem involves outputs that change based on data that are analysed at higher or lower levels of aggregation or changing the number of units in the analysis. For example, evaluating data at the local government level vs. a ward level. The zonal problem involves keeping the same scale of research (say, at the ward level) but changing the actual shape and size of those areas.

The work done for the March 2018 GCRO Map of the Month on vulnerability illustrates this problem. Figure 1 shows the Vulnerability Index mapped at a ward scale. The darker the colour, the higher the vulnerability. Figure 2 uses a Empirical Bayesian kriging (EBK) model to generate a smooth surface from the individual respondent data. EBK is a statistical interpolation model that predicts a smooth surface based on individual data points and their relationships to one another.

Figure 1: Vulnerability mapped at a ward level

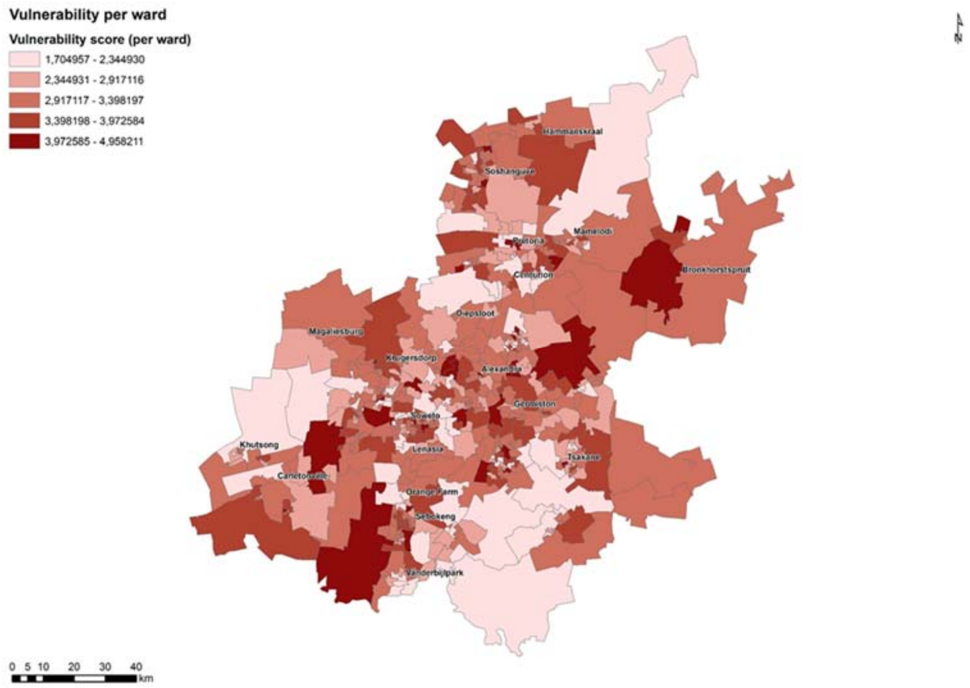
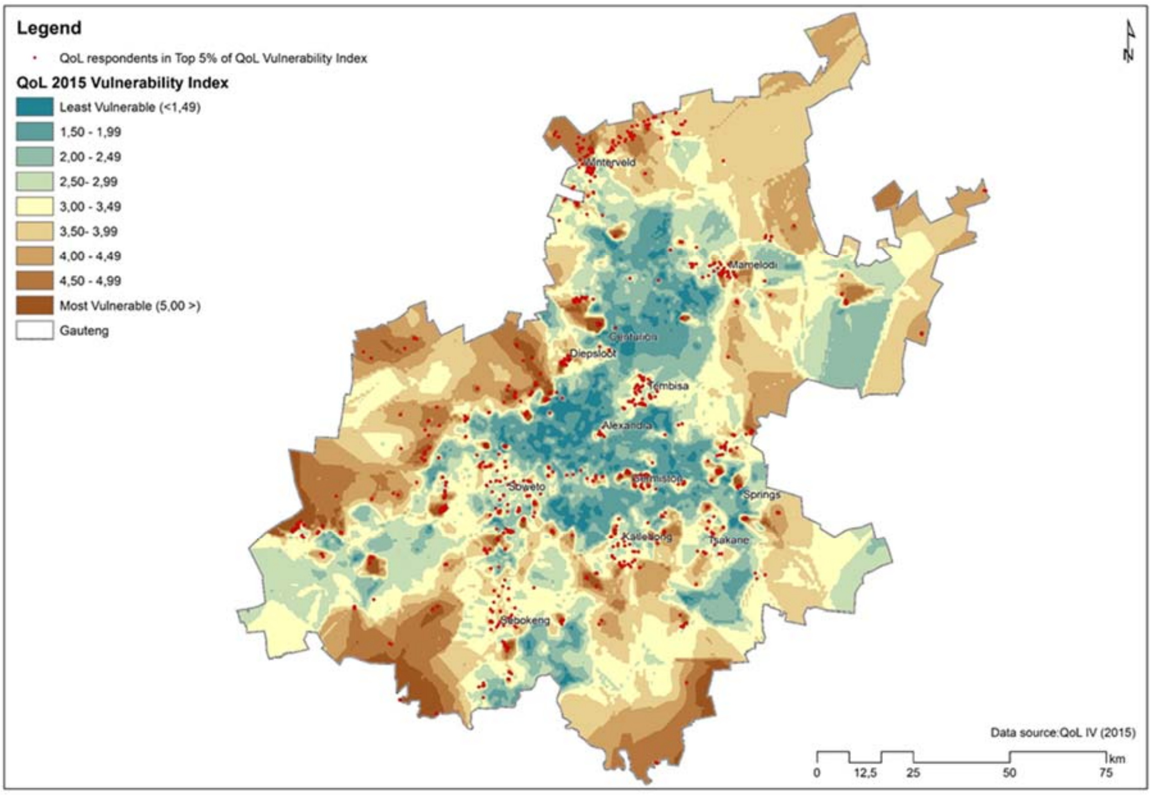
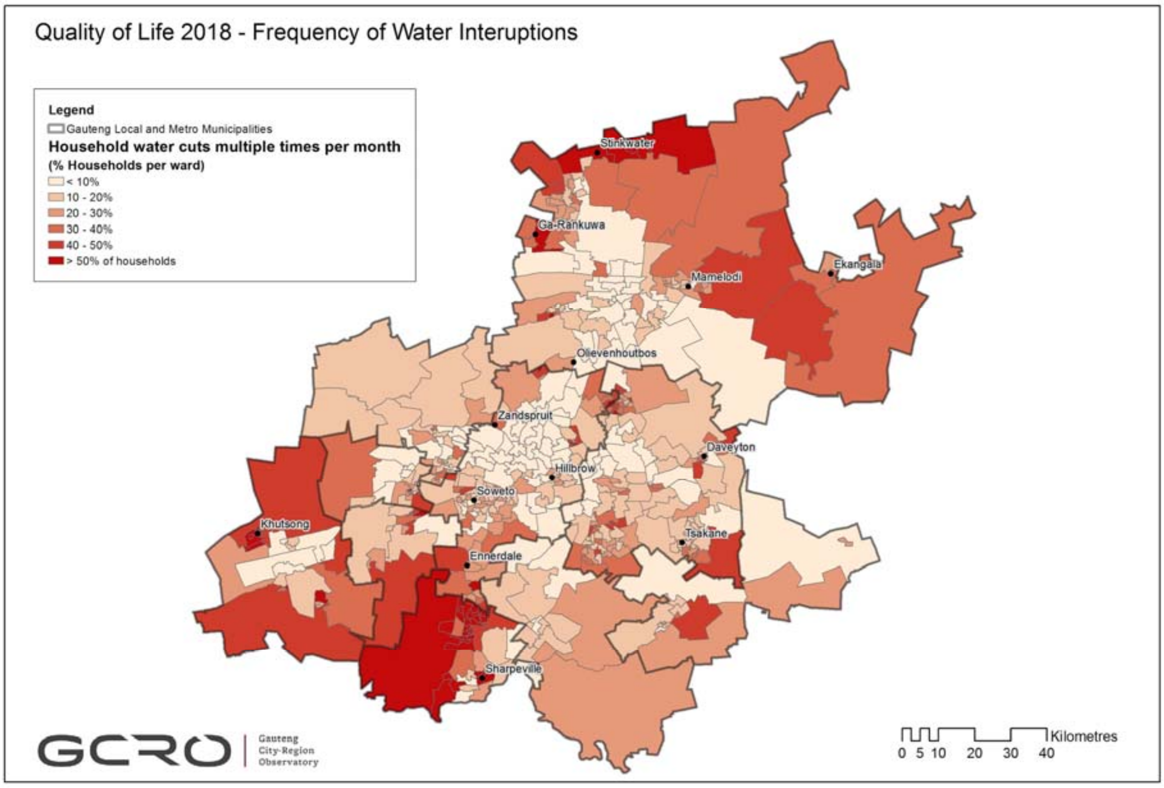



Figure 2: Vulnerability mapped at a respondent level using Empirical Bayesian Kriging




A further example maps water interruptions from the QoL V data. Figure 3 maps the data at a ward level, and Figure 4 maps the respondent data points. If a decision needed to be made of where to focus on maintenance of water infrastructure, the two maps would direct the user to very different areas. The ward level map shows the households on the periphery are the most in need, while the point map focuses attention on the most densely populated areas. Nether map presents the full picture, but rather are a result of the spatial scales chosen the spread of data collection.

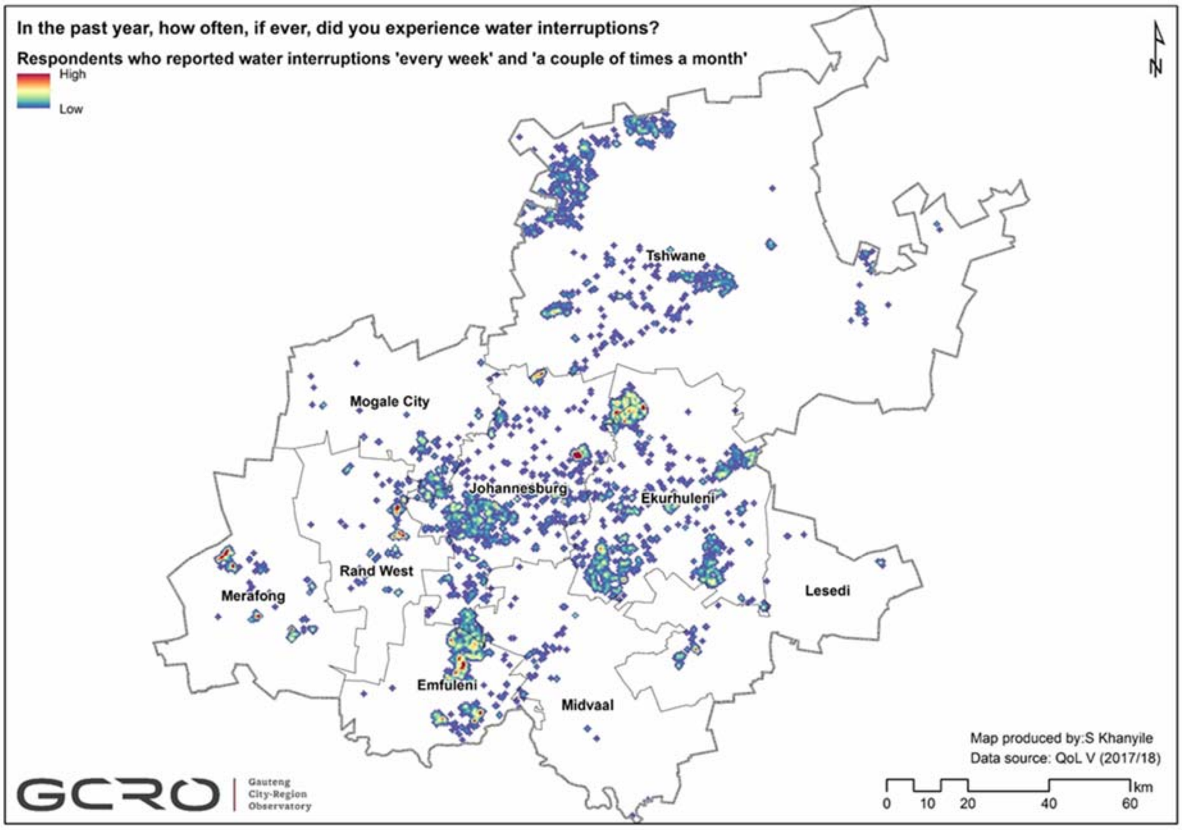
Figure 3: Frequency of water interruptions at a ward level




The 'Gordonia Problem' identified by Naude et al (2008) in the Gauteng case is a 'Goldilocks' in the GCR. Bronkhorstspruit is too big a ward and Thembisa is too small an area. In the case of Bronkhorstspruit ward the extremely large area and very high internal heterogeneity means that the area shows on a map as an area of significance. In the case of Thembisa, very small wards with high population densities does not show up as easily on a map. In the case of the vulnerability map (Fig 2) this problem becomes visible when points of the 5% most vulnerable were overlaid on the spatial surface model. When the point data and the EBK model were overlaid then the specific areas of significance can be seen. These issues are a particular problem when a variable has a very localised effect (e.g. exposure to flash flooding or segregation) and cannot be summarised across large wards. Mapping water interruptions is different as this variable is likely to affect whole neighbourhoods within the same infrastructure grid. In this case mapping at a sub place or a neighbourhood level will provide a better reflection.

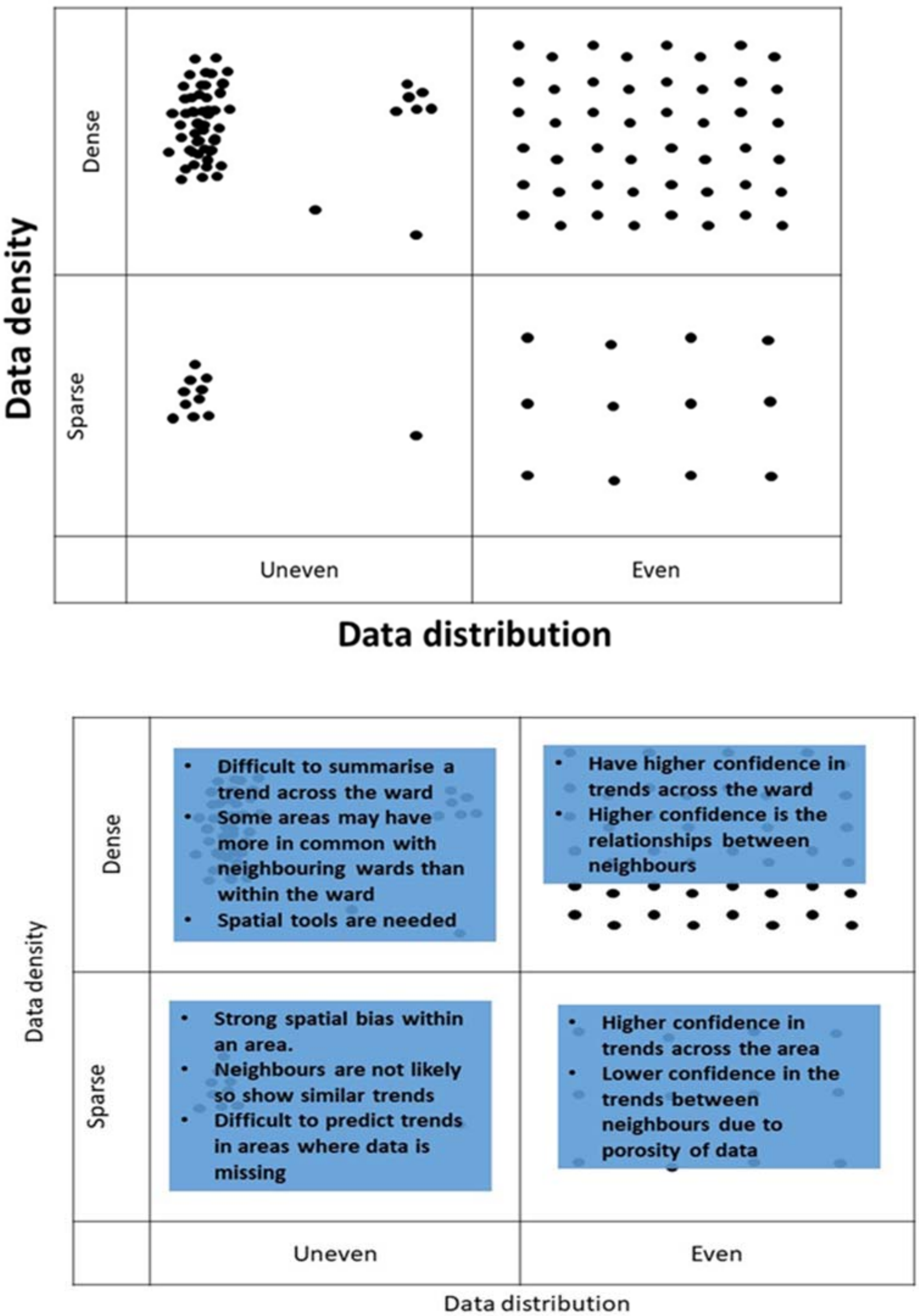
A key consideration here is that the wards are much smaller in densely populated residential areas and far larger smaller in agricultural or economic zones. This is not a problem is if the analysis is about understanding residential issues, but becomes an issue when considering economic and natural resource questions. A map can easily appear to distort the importance of a variable based on a large ward with few data points (and one or two outliers) versus a small residential ward with a large sample.

Figure 4: Water interruptions mapped at an individual respondent level




Boundary issues occur when data are clustered within one corner of a ward (or overlaps with a neighbouring ward) and an average of that variable is mapped as smooth surface across the whole ward. When mapped it is implied that the whole ward carries a similar value. However, clustering may mean that part of that ward may have more in common with a neighbour rather than within the ward (Figure 5). This is demonstrated by Bronkhorstspruit in the vulnerability maps where the ward level map indicates Bronkhorstspruit as being very vulnerable, but the data is actually clustered only around the town and informal settlements and the outer area of the ward have very low vulnerability scores.

Figure 5: Spatial implications of data distribution




## *3.2 Tobler's First Law of Geography*

Tobler's First Law of Geography states that all things are related to one another, but that objects that are near one another have a greater relationship than those that are further – this is the foundation of spatial autocorrelation and/or spatial dependency (Miller, 2004). Through Tobler's First Law, we are able to gather information that reveals spatial relationships between geographical entities and how, to a certain degree, there is an intrinsic uniqueness at every location (Miller, 2004). The consideration of Tobler's First Law may appear simplistic in nature, but it is sufficient in identifying local interactions between geographical entities that can produce complex behaviours at a global level (Sui, 2004). The implications of Tobler's Law allow us to better understand that for GIS and mapping purposes, one gets a better understanding of spatial relationship because these relationships are non-stationary.

Therefore, for modelling and visualisation purposes, there are spatial and statistical properties that do not remain constant and are not deterministic (Miller, 2004).

# 4 Implications for evidence based decision making

The choice of sampling method for QoL is done specifically to collect the best statistical sample in residential urban areas. Mapping average ward level data does can skew the outputs depending on the level of sampling, size of the ward and where the boundaries have divided the data. It would be impractical to collect an evenly spread statistical sample across the whole of Gauteng so the recommendation is that mapping of QoL data should take the following into consideration:

- Understand spatial bias may exist in the very large and very small wards. It is a good idea to check the levels of heterogeneity within wards or normalise the results by area to see if there is a spatial bias to better ensure that the sample better fairly represents the collected data
- Where variables may have a local impact and there is uncertainty about averaging data at a ward level, then make sure of spatial statistical checks or make use of the respondent level data to check for significant spatial relationships or hotspots.
- Spatial dependency or autocorrelation that accounts for how spatial areas are related to one another and have an effect to one another. This applies to neighbouring areas where the results have a negative or positive spill over effect on one another as well with the degree of association between the two (Wang et al, 2012)

In many cases QoL data allows us to the map relationships between data and respondents, rather than focusing on containing data in an artificial boundary like a ward.

## *4.1 Spatial sampling*

Depending on the variable that is being analysed, there are a number of ways and reasons to select a particular type of sampling. In spatial sampling, there are a number of samples that are obtained to establish the attributes of a larger geographic area (Wang, 2012). Factors such as frequency and distribution are calculated based on the predetermined sample region. If an attribute differs at various points, the area is heterogeneous. Heterogeneous areas are difficult to sample as when an area is missed out and not sampled, there are conclusions that are drawn for the entire dataset that may be inaccurate and not speak to the dataset at a local level (Wang, 2012).

There are two proposed sampling frameworks that may work to minimise such biases: Design-based sampling and Method- based sampling.

1. Design-based sampling
    - Sampling independent, identically distributed populations
    - Sampling considering spatial autocorrelation

- Sampling considering spatial heterogeneity

2. Model-based sampling
    - Minimising the estimation error
    - Equal spatial coverage
    - Equal coverage of feature space

The important consideration for QoL is the need to move from a 'container' based approach (what is within a ward) to a relational approach to sample collection. That means not focusing on what lies within a ward, but how data points they relate to each other, or how individual respondents relate to each other over space. No matter the choice of sampling method, there will be considerations of how that data is used and analysed in a GIS environment, and this is the most important consideration particularly when the maps are to be used to guide policy and decision-making.

## 5 References


Li, J. and Heap, A.D., 2011. A review of comparative studies of spatial interpolation methods in environmental sciences: performance and impact factors. *Ecological Informatics*, *6*(3-4), pp.228-241.

Miller, H.J., 2004. Tobler's first law and spatial analysis. *Annals of the Association of American Geographers*, *94*(2), pp.284-289.

Sui, D.Z., 2004. Tobler's first law of geography: A big idea for a small world?. *Annals of the Association of American Geographers*, *94*(2), pp.269-277.

Wang, J.F., Stein, A., Gao, B.B. and Ge, Y., 2012. A review of spatial sampling. *Spatial Statistics*, *2*, pp.1-14.