A new similarity measure for spatial linear networks

by

Mila Coetzee

17297363

Supervised by Inger Fabris-Rotelli

Co-supervised by Renate Thiede and Rene Stander

Submitted in partial fulfilment for MSc Coursework of the requirements for the degree

MSc Advanced Data Analytics

in the Faculty of Natural and Agricultural Sciences

University of Pretoria

July 2023

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Abstract

A linear network is a combination of line segments, or edges, that run between their defined endpoints, or nodes. They have become increasingly prevalent within spatial statistics due to the potential for representing systems from various fields as linear networks. One specific area of study within linear networks is understanding how they interact with one another and whether spatial similarity correlates with any underlying causal relationships. This line of research, however, remains limited due to the lack of a robust spatial similarity test suited for linear networks. This mini-dissertation therefore develops a new linear network spatial similarity test that specifically takes into account the spatial context of two linear networks and allows for spatially dependent variations in similarity. Different characteristics of the new test are demonstrated in two separate simulation studies. The first simulation study tests the overall performance while also illustrating how parameters can be optimised. The second simulation study shows the benefit of the newly proposed test compared to an alternative method. Finally, the test is applied to real-world informal road and mobility networks across northwestern Namibia to test whether mobility routes in rural areas are similar to existing infrastructure, and how the degree of similarity varies across regions. Subanalyses are also conducted to investigate the effect of road conditions, seasons and road density on the spatial similarity between the two networks.

Declaration

I, *Mila Coetzee*, declare that the mini-dissertation, which I hereby submit for the degree *MSc Advanced Data Analytics* at the University of Pretoria, is my own work and has not previously been submitted by me for a degree at this or any other tertiary institution.

_____

Mila Coetzee

_____

Date

Acknowledgements

I would like to start with thanking my supervisors Prof. Inger Fabris-Rotelli, Renate Thiede and Rene Stander for your unwavering willingness to always provide guidance with my mini-dissertation. Your help is especially appreciated since you always managed to accommodate my chaotic schedule despite your own busy workloads. My greatest stroke of luck during this degree was having three supervisors who are as kind as they are brilliant.

I would also like to thank Dr Ashley Hazel from the University of California for providing the mobility data set and sparking the initial idea for this research.

Finally, and most importantly, I want to thank my family for all their love, support and patience during the past eighteen months. Your combination of gentle encouragement and tough love have helped me see this degree through, and I am so grateful for your unconditional belief in me.

Furthermore, the financial support of the DSI-NRF Centre of Excellence in Mathematical and Statistical Sciences (CoE-MaSS) under grant #2022-018-MAC-Road is acknowledged and greatly appreciated. Opinions expressed and conclusions arrived at are those of the author and are not necessarily to be attributed to the CoE-MaSS. The research was also, in line with departmental procedure, approved by the Faculty of Natural and Agricultural Sciences under the ethics number NAS344/2022.

Contents

List of Figures

Chapter 1

Introduction

A linear network is a combination of line segments, or edges, that run between their defined endpoints, or nodes [13]. The edges take the shortest path between their respective nodes and are therefore linear. Additionally, linear networks within spatial statistics are also planar, which means that each data point along the edges and nodes can be embedded within either a 2D or 3D physical Euclidean space [198].

Linear networks are prevalent in many fields, and the spatial comparison between a linear network and other spatial phenomena is particularly worth investigating. Linear networks often represent complex systems of different scales and levels of complexity such as river ecosystems, congested urban streets and microscopic vein mappings within coronary arteries [71, 103, 87]. These systems have many intricate, interconnected components that not only vary across space and time, but may have dependencies on difficult-to-control latent variables [78]. Given the existence of these systems in a real physical space, the systems are rarely isolated and often overlap as they occur in close proximity to other related systems. For example, power grids and proposed transport routes for electrical vehicles occur in the same urban areas [81]. Alternatively, the systems may be part of a sequential chain of spatial phenomena such that the systems' underlying structures are dependent. An example would be neurological network structures, where the structure of neurons and their networks dictate the neurons' function and subsequently how they interact with surrounding neurons [182].

It is therefore beneficial to understand how the underlying dynamics of the systems may interact with any surrounding external systems such as other linear networks. While many of these highly technical cases require field-specific expertise to explain the actual dynamics of the system relationships, spatial statistics is able to provide its own insights. Specifically, by performing a spatial similarity test between the two linear networks, one would firstly be able to know exactly where these systems are similar. Additionally, by performing spatial comparisons for multiple time periods, the temporal aspect of any possible system relationships could also be modelled. Field experts would subsequently be able to investigate the specific measurements within those similar areas and during the included time periods, and explain the relationships from a theoretical point of view.

In the case where the surrounding external system is a point process, there exists literature which demonstrates the relationship between a linear network and a point process. Statistics to measure the spatial structure

between linear networks and point processes include a second-order statistic derived from the Ripley's K-function [39], a weighted K-function to account for the network's geometry [8] and inhomogeneous higher-order statistics [42]. Other related work includes Kriging and Voronoi-based intensity estimation of the point processes [196], non-separable first-order spatio-temporal intensity estimation [126], linear hotspot detection for a point pattern [133], pair correlation functions [193], relative risk estimation [127], parametric and non-parametric modelling [154], and point process modelling [40].

On the other hand, when the surrounding external system is a linear network, there is currently limited literature on the comparison of two linear networks. Previous work focuses on applying graph theory to quantify measures such as centrality, connectivity and spread of the individual networks and then compare the respective indices [191]. These indices, however, lose information by reducing the entire complex network down to a single value. They also do not fully consider how these individual characteristics interact. In the special case where linear networks may have the same nodes, known node-correspondence (KNC) methods have been proven to be accurate [187]. These methods include calculating the Euclidean distance between the two linear networks' adjacency matrices [65], computing the Jeffries−Matusita distance between the similarity matrices [98] and measuring the cut distance between the two graphs [111]. These methods, however, all require the recurring calculation of large distance matrices, which quickly become computationally expensive and restrict accurate use on larger networks [187]. They also do not have any application on networks with semi- or non-corresponding nodes.

EAnother previously applied method is trajectory similarity. The trajectory along linear networks describes the path a body such as a car or individual is likely to move along and is represented as a sequence of discrete positions [190]. This can be done using dynamic time warping, in which two linear networks are divided into equal discrete positions, and the Euclidean distance between each pair of points is calculated as a similarity measure [138]. Algorithms for the longest common subsequence have also been applied to see whether large segments of the two linear networks are in fact similar [77]. These two methods, however, both have a computational complexity $\mathcal{O}(n^2)$ which makes them unsuitable for any real-world application across large areas of road networks, for example. Neural networks have been used for comparing trajectories once they have been modelled as low-dimensional embedded vectors [110]. This addresses the computational limitations but fails to consider the effect of location and spatial context within spatial linear networks.

One last method for linear network comparison is the comparison of the linear features themselves [107]. Linear networks are deconstructed into vector lines which are mapped to embeddings without any feature engineering or format conversions. A Siamese neural network architecture is applied to minimise contrastive loss by arranging points with similar linear features close together and dissimilar features far away. The Euclidean distance of these learned embeddings is then calculated to estimate similarity. Similarly, linear networks can also be compared by being represented as a graph from which graph-embedded invariant attributes, such as node degree and relative edge distance, are calculated and compared. Both of these methods are computationally linear, but disregard spatial context and, in the case of the Siamese neural network, even disregard the comparison of nodes, which are integral in the composition of the linear network as a whole

[110, 107].

Given the breadth of application of linear networks in important fields such as engineering, medicine and environmental conservation, the need to compare linear network systems is clear [81, 159, 93]. The aim of this mini-dissertation is therefore the development of a spatial similarity test specifically for linear networks. An existing generic spatial similarity test for other spatial data has shown that representing spatial data as pixel images and then measuring similarity as the proportion of pixels that overlap has been highly successful [94]. The newly proposed linear network spatial similarity test builds on the theory of this existing spatial similarity test to extend to linear networks. This approach addresses the above-mentioned limitations of comparing linear systems. The test also allows for both global and local evaluation of similarity to better understand where and to what extent similarity varies.

The illustrating example of two connected linear systems considered for the remainder of this mini-dissertation is that of a social mobility network and the surrounding informal roads network. The social mobility network maps how individuals move around. The network nodes represent origin and destination villages, and the edges represent the shortest route between these origin-destination pairs. The informal road network maps all informal roads where edges are road segments and nodes are road intersections. As shown in Figure 1.1, villages are spread across regions with many different informal roads stretching between them. The two linear systems are potentially connected. Either the roads are formed to follow along mobility routes between villages, or villages are placed, and so mobility routes are formed, according to existing road networks. Social mobility is chosen as the demonstrative example in this mini-dissertation due to both the

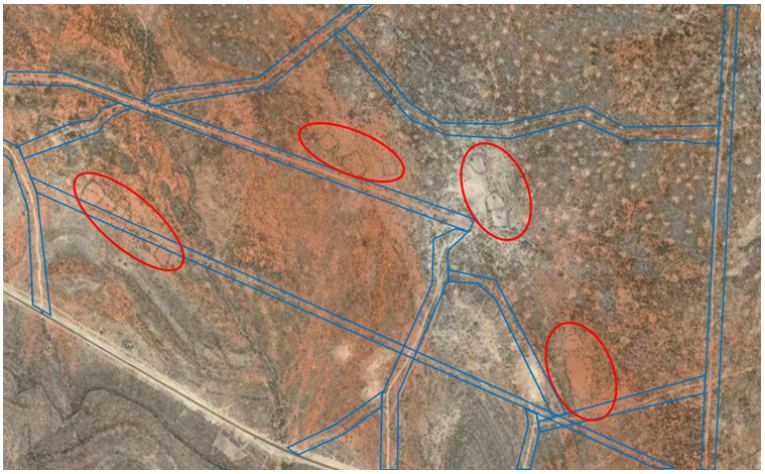


Figure 1.1: A satellite image of the surveyed area containing the social mobility network and the informal road network. The social mobility network consists of origin and destination villages marked in red. The informal road network includes the well-defined non-tarred informal roads between and around each village shown in blue [121].

wide applicability of mobility data as well as its diagnostic contribution to the spatial similarity test. Firstly, in a world of ever-increasing connectivity and dependency, social mobility data has many relevant applications. Geographical, infrastructural and social systems can all occur in tandem with social mobility. Examples include rivers, roads and familial ties which can be represented as linear networks. If the spatial relationship between all these systems is quantified, then it could help understand where people within the network are likely to move to next, the amount of people moving in each direction and possible variables affecting those decisions

[1]. An understanding of the dynamics and underlying processes driving mobility could lead to new and efficient applications in fields such as computational health, optimal urban planning, traffic control, refugee displacement, epidemic control, location services and emergency management [21, 214, 116]. One example is the 2010 Hawaiian earthquake, where mobile phone data was used to optimise first response allocation [116]. The optimal placement of public facilities such as airports, hospitals and shopping centres also depends on understanding spatial behaviour [58].

The beneficial application of mobility data has thus far been hindered, since most mobility data was sourced from traditional surveys [202]. This limited modelling since data often lacked granularity and was too expensive to scale. With modern technology, however, the immense amount of georeferenced data available has made data-driven modelling possible. A lot of this data comes from GPS-enabled smartphones, vehicles and geotagged social media posts [166]. These different data sources not only allow for larger data volumes but also for increased data variety. Vehicle-based data, for example, may explain how people travel whereas social media could be more helpful in understanding why people travel. Mobility data, however, still presents problems in modelling, specifically in the difficulty of pre-processing [166]. Technological data points are often represented as raw trajectories that not only take a long time to transform but can often include underlying noise as a result of poor signal reception. The process of filtering, constructing and comparing data points is also restricted by certain privacy laws [46]. Furthermore, data is not always consistently gathered across all modes of transport and thus different temporal periods and reporting conventions can make accurate modelling more difficult.

Limitations concerning social mobility studies are further exasperated in rural communities. Most studies have focused on urban mobility despite the fact that in certain countries like Namibia, rural travel accounts for 60% - 70% of all travel [130, 62, 206, 44, 5, 55, 30]. The primary reason for the lack of rural mobility studies is the availability of data. Far more mobility data is produced by urban inhabitants due to better Internet connection (and as a result, increased use of geotagged posts like Twitter and Instagram), closer proximity to cell towers and higher disposable income to afford vehicles and mobile phones with GPS transmission. Rural mobility remains mostly unmodelled because the little data available such as survey and census data does not allow for detailed analysis. Furthermore, it has been proven that urban mobility models cannot simply be applied to rural areas [130]. More difference has been observed between mobility in urban and rural areas of the same country than urban areas across different countries [161, 62].

Geographic morphology of an area is one factor that has been shown to significantly influence mobility [105]. As such, rural areas have different spatial patterns, which affect land use and street topology. Previous urban studies show how the regularity of urban streets affects mobility metrics like average shortest path, travel waiting times and trip durations [20, 43, 117]. Assumptions between urban and rural areas also vary. The assumption of scale-free networks, for example, does not generally hold in urban areas because scale-free networks expect an increase in locations over time, which is generally infeasible due to urban density. The assumption may, however, hold in rural areas due to the open spaces. Additionally, informal roads and settlements increase much faster than government-controlled roads and towns, given that they are created

by people as and when they need them. The assumption of scale-free networks is important as it affects the distribution of locations in a network, the expected growth of the network and the overall centrality and connectivity of the locations within the network [31].

Socio-economic factors also affect rural mobility. With comparatively lower infrastructure investment, rural areas do not have as much access to fast and easy transport as urban areas. This may reduce the number of recreational trips taken by rural inhabitants or at least delay them until several tasks can be completed during a single trip rather than taking a trip for each task as is often the case in urban areas. On the other hand, due to the decreasing number of people living in rural areas [169], these areas experience less congestion, so that trip duration may be shorter compared to trips in urban areas of the same distance [130]. Urban areas also often experience abnormal mobility behaviour at points of interests, such as airport pick-ups [206]. Rural areas do not have these particular locations, but they may have their own points of interests with their own effects on mobility. Differences in socio-economic circumstances, especially in disposable income and severity of work responsibilities, also affect people's behaviour and will affect their propensity to either stick to a predictable mobility pattern or otherwise explore new locations [155].

A few studies have attempted to model rural mobility using a modified gravity model [222]. The modifications address the tendency of the original gravity model to overestimate rural mobility. The modified models include parameters for socio-economic and geographic factors like education, economics, gender, environment and proximity to urban areas [57]. Certain models even account for the regional heterogeneity of the impact of these factors on mobility by estimating each of the parameters per region and per urbanicity[1]. Few models, however, consider how rural mobility is explicitly affected by location and the surrounding infrastructure. In particular, very little is known of how the informal road system affects rural mobility. Opening up analysis into rural mobility by means of a more intuitive yet accurate representation, as is intended with the novel spatial similarity test, would be very beneficial for application in current and future work.

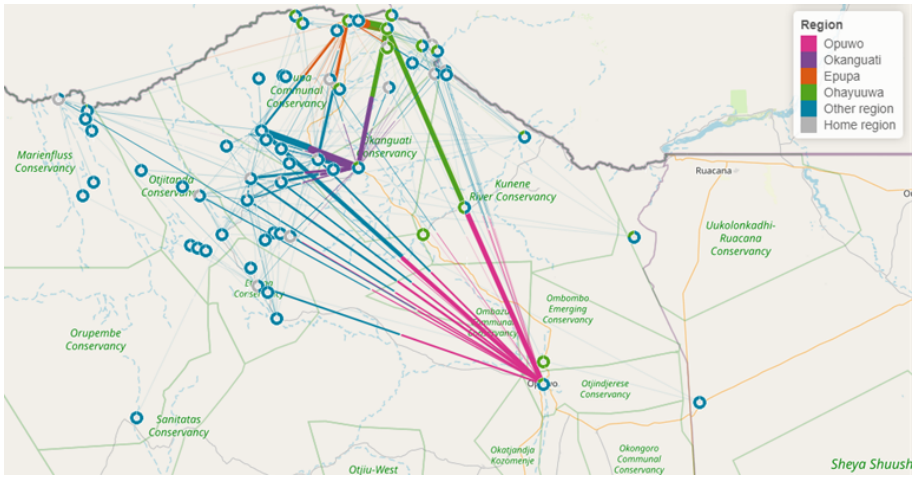


Figure 1.2: An example of a social mobility network represented as a linear network mapped over the Kunene region in Namibia.

Additionally, analysis into rural mobility can also be considered from a temporal perspective. A common example is tracking mobility across different seasons [223]. Seasonal changes affect mobility behaviour. For

[1] Urbanicity is defined as the type of traversal between areas and includes rural-rural, rural-urban, urban-rural and urban-urban [200].

example, the rainy season may decrease people's tolerance for travelling long distances [38]. Alternatively, dry seasons may dry up rivers, which increase the accessibility of certain roads. Even the choice of destination may be affected by the season as people may, for example, travel to villages with a clinic more often during winter due to increased cases of sickness [142]. The mobility data can be subdivided using simple methods like survey data collection. The changes in mobility behaviour can then be compared across time to identify any trends, cycles or outliers for mobility network characteristics like volume, location, centrality and complexity.

The second reason for choosing social mobility data is that it is a perfect example of a linear network that is difficult to accurately and thoroughly analyse. Mobility data is commonly not represented as a network but instead analysed using either interaction models like the gravity model or stochastic models such as the vacancy models or Markov mobility models [33, 212, 56]. These models are generally easy to understand, are well-founded in the principles of entropy-maximising [213] and provide sufficient modelling of the volume of mobility. However, they all lack spatial context and do not account for any spatial dynamics [24]. They do not allow for more in-depth analysis whether the spatial behaviour underlying the volume of mobility is affected by aspects such as the proximity to existing infrastructure, the conditions of the roads, means of transport or the reason for travelling [53]. These methods focus on modelling the nodes, or destination, and altogether neglect to analyse how the characteristics of the mobility routes such as length, placement, orientation and density, may correlate to the mobility volume at each route's destination. Additionally, Figure 1.2 shows that mobility networks can span over vast areas, which can quickly become computationally expensive. The newly proposed spatial similarity test can be considered useful if it addresses these limitations. It therefore needs to take into account the spatial context of the 2D plane the data is mapped onto, allow for spatially-varying or regional-specific factors that may affect the characterisation of the mobility network, and remain computationally feasible.

The informal road network was chosen as the second linear network to compare using the new spatial similarity test. Previous work has given a preliminary indication that mobility data and informal road networks are related under certain circumstances [72, 49, 118, 114]. This relationship, however, requires further investigation as the similarities have been shown to vary geographically. Additionally, most of the previous work has been done from a qualitative perspective. Furthermore, it is beneficial to demonstrate a new method of analysing informal roads given their increasing number of applications ranging from environmental sustainability to increasing transport accessibility to urbanisation [100, 99, 101].

As an added benefit, the informal road network also presents additional challenges that would make it difficult to properly compare to other linear networks. These challenges serve to further prove the robustness of the new linear network spatial similarity test. As previously mentioned, graph theory is often used to analyse networks. However, in the case of road networks, the number of nodes and edges tend to be very high due to the number of terminating side roads, the curves around geographical features and the distance of the area being covered. Comparing such a complex network with a much simpler network would result in misleading network indices [191].

Another point to consider is the compatibility of data granularity and reliability of the two linear networks

being compared. A simpler network has fewer nodes and so fewer data points need to be collected and verified. Mobility data, for example, is often manually surveyed and so each data point is usually checked for reliability and precision [151]. Informal road data, however, is often not as thoroughly verified. An informal road, by definition, is a road formed through natural human movement rather than government-driven construction[2]. As a result, most informal roads are not included in official databases [189].

Remote sensing is a popular method of collecting large amounts of road data as it scales and overcomes the issues of surveying difficult terrain [205]. Established methods of object detection in road extraction, however, cannot always be directly applied to informal roads owing to their varying widths, discontinuity, irregular textures, low contrast and undefined boundaries [132, 146]. Neural networks, on the other hand, have successfully been used in informal road extraction as they are able to consistently recognise patterns regardless of the patterns' placement or transformation [129, 134, 175, 50]. The neural networks rely on well-labelled training data sets and optimise their respective classification algorithms via back-propagation [170].

Despite the theoretical merits of neural networks, in practice there are still some limitations. A neural network would initially require an enormous amount of region-specific training to ensure that the neural network was familiar with as many examples of informal roads as could be expected. Different terrains may include native vegetation, region-specific soils, a different positioning in relation to the sun which affects shadows and varying levels of urbanisation. This will all affect the performance of the neural network [146]. Not only is this technically difficult due to a lack of high resolution satellite imagery, but the time it would take to collect and label the training data, build and train the neural network, and finally test and optimise the neural network falls outside the scope of this project. Additionally, it is uncertain what level of coverage the neural network would attain. This level of uncertainty could inadvertently affect the spatial similarity test.

The informal road data is collected by manually digitising the informal road network in ArcGIS from Google Map images [121]. This method is intended to standardise the data reliability between the two linear networks. Previous studies have shown that manual digitisation often maximises coverage and accuracy compared to automated methods [3, 29]. It should be noted that while it is important to collect data that is as thorough and accurate as possible, this also now means that the similarity test will need to be able to accommodate large coverage without losing any granularity.

In addition to the informal road network itself, the road conditions for all the roads can also be identified. Road conditions refer to the state of the road and is most often classified based on texture [184]. For example, the condition of regular, evenly paved roads can easily be distinguished from the condition of jagged, rocky roads. Different methods for identifying road condition labels exist, including aviation vehicle-based photogrammetry [217], ground-based data collection using smartphones [52], vehicular sensor networks [128] and point laser sensors [109]. For large networks of informal roads, on the other hand, these methods would either not be available or not be able to scale. Additionally, ground truth labels for the road conditions are not available to train appropriate supervised classification models. As a result, an automated, accurate and unsupervised texture classification model is required.

Despite the challenges, it is worth classifying the informal road network according to road condition for

[2]https://wayleave.tshwane.gov.za/document/download/335975/ Accessed: 2022/05/23.

further in-depth analysis. Road conditions have been proven to have a significant effect on route selection [201, 158, 216]. This is because road conditions affect factors such as accessibility, safety and comfort which all, in turn, affect how likely people are to travel along a particular road. Mobility routes and clear, even, stable roads are so much more likely to be similar, since people perceive them as the path of least resistance. Conversely, mobility routes and obstructed, rough, unstable roads will probably deviate more frequently. To best understand this effect of terrain and geography on mobility, it is best to consider the individual road condition linear networks for optimal spatial comparison.

In summary, this mini-dissertation proposes a new linear network spatial similarity test based on a recently developed a generic spatial similarity test [94]. The test aims to compare two linear networks in a robust manner that takes into account the spatial context of the 2D data, allows for spatio-temporal variation, and accommodates networks with distinctly different characteristics like shape and complexity. Additionally, the test must be detailed without becoming too computationally expensive. To fit the existing framework, the first step of the test is to convert the linear networks into suitable point patterns. The second step is to convert the point patterns into pixel images. The two respective pixel images are used to generate a local similarity map using the structural similarity index measure (SSIM) in the third step. Finally, a global similarity index is calculated. Specifically, Andresen's S-index, a non-parametric area-based spatial point pattern test, with a non-binary input is used to compare and quantify the similarity of the two different pixel images [6].

With the primary aim of developing and evaluating this new test, the objectives of this mini-dissertation are as follows:

1. Outline the methodology of the novel linear network spatial similarity test by extending the application of an existing spatial similarity test in [94].

2. Conduct the first simulation study to assess the overall performance of the new spatial test as well as demonstrating a method for optimising the test parameters.

3. Conduct the second simulation to prove the efficiency of the new method compared to an alternative method.

4. Apply the spatial similarity test to the social mobility network and digitised informal road network to determine the overall similarity between the two linear networks.

5. Apply the spatial similarity test to the social mobility network and each of the road condition linear networks to determine the effect of road condition on mobility behaviour.

6. Apply the spatial similarity test to the digitised informal road network and each of the seasonal mobility networks to determine the effect of seasons on route selection.

7. Apply the spatial similarity test to an updated digitised informal road network and the social mobility network to decrease the bias of disproportionately inflated road densities.

Chapter 2 provides the necessary background theory on linear networks. Chapter 3 conducts an exploratory analysis of the mobility and informal road network data sets to gain insights into the networks' individual

characteristics and provides a preliminary assessment of their apparent similarity. The new linear network spatial similarity test with all necessary theory is outlined in Chapter 4 while the simulation studies are included in Chapter 5. Chapter 6 includes the final application of the new spatial similarity test on the rural Namibian mobility and road network data as well as the road condition, seasonal and road density-corrected subanalyses. Finally, the conclusion and future work are described in Chapter 7.

Chapter 2

Background Theory

The purpose of this chapter is to provide an overview of linear networks. Section 1 defines a linear network and discusses all relevant concepts. Linear structures and their different applications are discussed in Section 2. Section 3 explains network analysis, given its prevalence in linear network analysis. Finally, the various texture classification methods and their application to linear networks are discussed in Section 4. Each section provides insight into different possible methods that could be used as well as highlights the limitations of methods not implemented in this mini-dissertation.

2.1 Linear network notation

The linear networks within the scope of this mini-dissertation are graphs embedded in an appropriate Euclidean plane. This is often chosen as $\mathbb{R}^2$ when a simple x-y coordinate system is used, but it can also be applied to $\mathbb{R}^3$ in the case of elevation-dependent data. The graph consists of a finite set of vertices $v \in V = \{v_1, ..., v_m\}$ and edges $e \in E = \{e_1, ..., e_n\}$.

The edges can be denoted as $e = [v, v\prime]$ since a vertex always occurs where two edges intersect and contains at most one point [7]. More formally, the edges are considered line segments and are defined as the shortest path between the vertices v_i, $v_j \in \mathbb{R}^2, v_i \neq v_j$. They are expressed as

$$e = \{tv_i + (1-t)v_j : 0 \leq t \leq 1\}. \tag{2.1}$$

The length of each line segment is denoted as

$$|e| = \|v_i - v_j\|, \tag{2.2}$$

where $\|.\|$ represents the $\mathbb{R}^2$ Euclidean norm.

A linear network can formally be defined as the union of its respective edges [7]

$$L = \bigcup_{i=1}^{n} e_i \tag{2.3}$$

for the line segments $e_1, ..., e_n$ where $1 < n < \infty$.

2.2 Linear networks and their application

A spatial linear network is a linear network where all nodes and edges occur in either a two- or three-dimensional space and Euclidean distance is used to measure physical characteristics of the linear network [17]. Before comparing spatial linear networks, it is important to understand when they are used and how they are represented. Spatial linear networks are generally applied to two types of spatial processes, each modelling a specific spatial phenomenon.

The first case considers a spatial point process. A spatial point process is formally defined as a stochastic process within a bounded area $A \in \mathbb{R}^2$ in which finite random subsets $\{\mathbf{x}_1, ...\mathbf{x}_n\} \subset A$ are realised [136]. Each of the points $\boldsymbol{x}_i = (u_{1i}, u_{2i})$ represents a known event. The spatial phenomenon is therefore a set of discrete points or locations. The focus, however, is not on the spatial distribution of the points themselves, but instead on any relationships between the points. A common application is origin-destination data, where the relationship between these points is of interest in order to understand underlying dynamics such as which locations generate the most traffic, which travel routes are most common, and what effect distance has on travel [64]. It should be noted that in this case the relationships between points are theoretical and therefore artificially derived based on certain assumptions, such as the shortest path between points [120]. The linear network edges are, as a result, abstract and do not correspond to any real spatial phenomenon. The advantage of using linear networks for these use cases is that the abstract relationships are now easily visualised, and the relationship links can be better compared and classified based on quantitative metrics like edge length, orientation and density.

A concise manner of modelling the specific case of spatial data containing origin and destination locations is the bivariate-linked point process [11]. The event $\mathbf{x}_i$ is described as $(\mathbf{x}_i, \mathbf{y}_i)$ where $\mathbf{x}_i$ and $\mathbf{y}_i$ are the i^{th} origin and destination locations, respectively. The process models any spatial dependency between the origin and destination locations, which is what makes it especially suited for migration and mobility data sets [115].

The Gibbs model with pairwise interaction functions was originally proposed for the bivariate linked point process [11]. It is assumed that there is a distribution which is conditional on the two unordered sets of the origin and destination locations, $\mathbf{x}_1, ...\mathbf{x}_n$ and $\mathbf{y}_1, ...\mathbf{y}_n$. The unordered set of events containing all possible origin-destination linkages is denoted as $\pi\varphi$. It is constructed by shuffling the destinations by some arbitrary permutation of indexes $(\pi_i, ..., \pi_n)$. A potential function ϕ is also specified to take into account any suspected constraints within the spatial domain A that the points occur. A common example is the assumption that travel likelihood is directly inverse to travel distance. Given this potential function, a random permutation is realised from the unordered set $\pi\varphi$ as a result of either accepting or rejecting the permutation based on a specified transition rate.

The bivariate-linked point process finally converges with a probability density function expressed as

$$C \exp\left(g\left(\mathbf{x}_1, \ldots, \mathbf{x}_n\right) + h\left(\mathbf{y}_1, \ldots, \mathbf{y}_n\right)\right) \sum_{\pi} \exp\left(-\sum_{i<j} \phi\left(\left(\mathbf{x}_i, \mathbf{y}_{\pi i}\right), \left(\mathbf{x}_j, \mathbf{y}_{\pi j}\right); \boldsymbol{\theta}\right)\right) \tag{2.4}$$

where C is a constant, g is the marginal probability density function for the origin point process, h is the marginal probability density function for the destination point process and $\sum_{\pi}$ denotes the summation of all possible destination permutations. Additionally, θ is a parameter measuring the degree to which the origin and destination processes are dependent. When $\theta > 0$, the density function is assumed to attribute more probability mass to instances where origin points with close proximity are likely to also have corresponding destination points with close proximity. Origin-destination pairs are therefore clustered. When $\theta < 0$, the opposite occurs, and nearby origin points have far away destination points. If $\theta = 0$, the origin and destination points are independent.

Another method of point process representation of origin-destination includes modelling the origin and destination locations as two separate point processes. Any spatial dependency is subsequently modelled using the joint intensity $\lambda(\mathbf{x}_i, \mathbf{y}_i)$ across all location pairs [180]. A marked point process can also be used where, each origin location x_i in the spatial pattern $\mathbf{x} = x_1, ...x_n$, has an associated random structure $\mathbf{m}_i = m_{i1}, ...m_{in} \in [0, 1]$ representing all possible connections between the n destination locations $\mathbf{y} = y_1, ...y_n$ [82]. For example, if there is a connection between the i^{th} origin location x_i and the j^{th} destination location y_j then $m_{ij} = 1$ while no connection would be denoted as $m_{ij} = 0$. Similarly, the destination point process can also be marked by considering connections between each respective destination location y_i and all n origin locations $\mathbf{x} = x_1, ...x_n$.

The resulting point pattern can be used to construct the linear network. This is done by defining each of the points as nodes [191]. It should be noted that origin-destination data does not always first have to be physically represented as a point process. Defining nodes directly from origin-destination matrices is common. This, however, does not negate the underlying point process. A distinct difference is being made here between spatial and non-spatial linear representation of systems, where the latter is often approached from either an algebraic or geometric perspective [91, 75]. Once the nodes have been noted, a path between each bivariate pair of origin-destination points is added as the network edges [191]. In the case of linear networks, the assumption of shortest path between points is used. This is a reasonable assumption but given that spatial linear networks occur within the Euclidean space, physical constraints and obstacles should at the very least be noted in any analysis [145].

In the second case, linear networks are applied to line processes. A line process is defined as a random collection of lines [183]. The lines, similar to the point process, are realised by means of a stochastic process from a finite random subset of line segments $\boldsymbol{l}_1, ..., \boldsymbol{l}_n$. The line process is bounded within the area $A \in \mathbb{R}^2$. Here, each of the line segments l_i also represents a spatial event, specifically the occurrence of a physical linear structures such as roads or rivers. The focus is to extract and estimate the best linear representation of these physical structures. The linear network edges, unlike the previous case, therefore, do correspond to real spatial phenomena.

The task of line extraction often falls within the fields of GIS and computer vision [164, 131]. The linear structure is initially represented as an image. It is skeletonised to extract the medial axis while maintaining information of all relevant geometric and topological properties [173]. In general, this can be done from a geometric, curve propagation or digital perspective [174].

The first category utilises Voronoi diagrams in which the medial axis is regularised by attributing each component of the axis a measure of prominence and stability based on Euclidean distances and connectivity metrics. While Voronoi diagrams remain invariant to geometric transformations and noise, they tend to produce a large number of unwanted line segments [148]. The second method uses partial differential equations to optimise the medial axis, but is not always topologically connected [10]. The final method applies a thinning algorithm in which geometric and topological rules are applied to dictate the iterative removal of pixels within the defined digital grid [173]. These include kernel-based, distance transform and iterative boundary peeling algorithms [9]. Parallel iterative algorithms are also used in which fully parallel algorithms apply the same deletion criteria simultaneously [83]. Additionally, subiteration parallel algorithms alternate the deletion criteria while subfield parallel algorithms subdivide the image and apply one designated deletion criteria to each subdivision [221, 61]. Common parallel thinning algorithms include the ZS, Tarabek, RIEPTA and OPTA algorithms [221, 188, 108, 28]. These all aim to approximate the medial axis with a one-pixel thick line while preserving both endpoints and connectivity while ignoring slight noise near the boundary [174]. In more complex cases where the image is not binary and the linear structure is composed of multiple classes, alternative vectorisation methods such as constrained Delaunay triangulation, Bayesian estimation theory, contour following, junction analysis and maximum threshold morphology are used [141, 18, 195, 37, 84].

The preliminary linear structure that has been extracted consists of lines structured according to some underlying physical, contextual and functional constraints [32]. The manner of representation determines how much of that structural information is retained, including information pertaining to centrality, connectivity, hierarchy, circuitry and topology [124].

There are two primary ways of representing a linear structure, namely structural and graphical [123]. In structural representation, also referred to as route structure analysis (RSA), routes are chosen to represent the most continuous paths of movement through a junction. This can often be optimised in many different ways. Different methods include continuing roads with the same classification if any previous classes (like road type or purpose) are known, adhering to junction priority where certain roads are legally required to yield, continuing physical alignment by giving priority to the longest roads, and using information such as street names [122]. As can be seen in Figure 2.1 a), the linear structure is represented as 5 continuous routes. On the other hand, a traditional graph network, shown in Figure 2.1 b), divides the continuous routes into 9 non-intersecting segments. This distinction is important as it means structural representation preserves continuity through intersections, and therefore does not lose information on continuation and termination conditions [122].

Additionally, RSA prioritises structural analysis of the road network overflow analysis of people who move along the roads [122]. This can be restated as saying RSA is link-centric as opposed to graph networks which are predominately node-centric [215]. Link-centric means the focus of the analysis is on properties of the

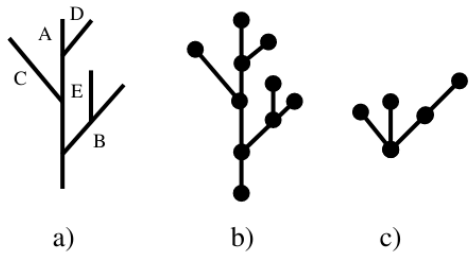


Figure 2.1: A visualisation of the different line structure representations, namely a) a route structure, b) a primary graph, and c) a dual graph.

roads themselves. Three common properties include continuity, connectivity and depth. Continuity considers the number of links in a particular route, while connectivity looks at the number of routes that connect to a particular route. Depth refers to the distance a route is from a baseline datum route and is measured in number of steps of adjacency. The datum routes are assigned a depth of 1 and are chosen strategically to represent the most important road, whether that be due to location or capacity. One very useful application of these properties is in deriving new road types. The relative value of each metric is calculated and represented as γ, χ and δ for the relative continuity, connectivity and depth, respectively. The values are then plotted on a routegram along three axes. As is seen in Figure 2.2, the values are classified according to which sector of the routegram they lie in. Classification can also be done for the whole network by summing the metrics for all routes and plotting the relative values along a similar netgram. While this does not aid in calculating the spatial similarity of linear networks, it is a good method of quantifying structural similarity.

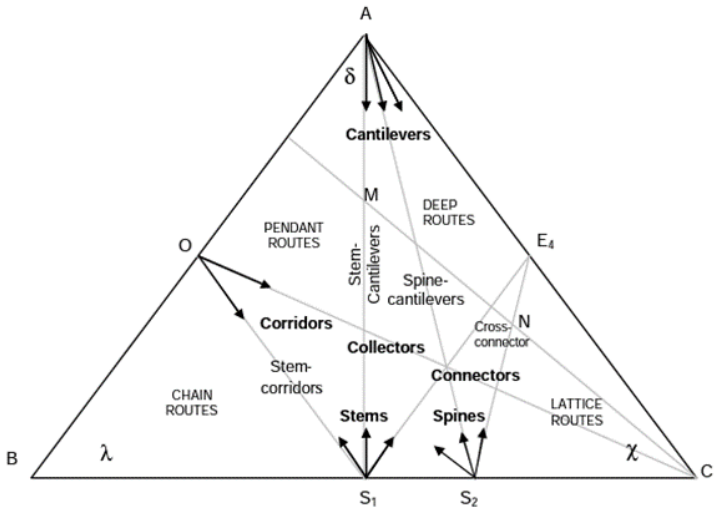


Figure 2.2: A routegram demonstrating how road classification can be done using γ, χ and δ values [122].

Other properties that can be calculated include junction type and cardinality. Junction type classifies junctions according to their shape and number of connections. The hierarchy of the road network can also be quantitatively measured by calculating the cardinality, which measures how high on the road hierarchy a particular road is based on how often it yields to other routes.

As a preliminary analysis of spatial similarity between two linear networks, each of these characteristics can be individually and independently considered. For example, the more continuous and connected roads are, the

more likely people are to travel along those roads as opposed to roads experiencing discontinuity [220]. Traffic flow along a particular road is driven by the convenience, comfort and consistency of the road. This applies to both the quality of the road as well as the effort of having to transfer to either a new level in the road hierarchy or to a new transport mode [215]. The cardinality of the road network can help segment road networks into subnetworks according to hierarchy and mobility can be analysed for each of the different roads to determine if behaviour varies across hierarchies. People are also less likely to travel along routes with higher depths if the baseline route is chosen as a route they use most often, like their route to and from home. Junction type can be analysed to see if certain junction types lead to higher congestion, higher use of pedestrians compared to cars and increased traffic flow to certain sectors of the network [122].

A graphical representation, on the other hand, represents data as a generalisation of traditional graph diagrams. This remains the most widely used linear network representation as it is easy to construct and remains both conceptually and spatially intuitive. The graphical network consists of nodes, and edges that are defined depending on the graph's context [191]. Primal graphs are most common. The vertices of the line segments are represented as the nodes, and the line segments themselves are the graph edges. Admittedly, this does mean that primal graphs are node-centric and that topological analysis of the network focuses on the distribution of edge intersections rather than the edges themselves. Only a few edge-specific metrics, such as average path length, are commonly included in network analysis [36]. Other analysis is highly dependent on detailed region-specific data such as travel time and link load capacity [179, 34]. Despite the limited number of metrics focused on network edges, a primal graph representation is appropriate in the case where both nodes and edges are of concern. While the structural representation prioritises the structural information of the edges, it completely excludes information on the nodes. This is specifically problematic when the nodes represent points of interest with intricate spatial dependencies, such as origin-destination locations.

One alternative to consider are dual graphs, as shown in Figure 2.1 c), where nodes represent line segments and edges represent intersections. This allows analysis of the road distribution, centrality and location entropy. It is also helpful as it differentiates between edges that continue through nodes, like people moving across road junctions, and edges that simply move between junctions, like a shuttle network where nodes represent point-to-point service stops [123]. As a result, mobility trends at intersection nodes and road endpoint nodes are differentiated. Comparisons can therefore be made between mobility behaviour in the middle of roads as people are still travelling and mobility behaviour at the end of roads once they have reached their destination [48]. Dual graphs, however, lack the visual resemblance to actual road networks and are thus less interpretable [19]. They would also lose spatial granularity by abstracting the origin and destination locations from a primal planar graph to a dual graph [198].

2.3 Network analysis

A network graph consists of vertices, or nodes, which represent locations visited and edges which represent the shortest path between these locations. Edges can be directed, in which case mobility only flows in one direction, or otherwise undirected, in which case movement between locations is mutual [152]. There are different types of

networks including regular, random and complex networks [15]. They are classified according to characteristics like node degree distribution and clustering coefficients. Regular networks have nodes arranged along square or cubic lattices where each node has the exact same number of edges. Random networks, on the other hand, consist of N nodes and $\frac{N(N-1)}{2}$ edges, where the likelihood of any two nodes being linked follows some underlying probability distribution [167]. Random networks are thought to reflect many real-world systems and are particularly noted to have a steep cost of adding nodes, which corresponds to real-world restrictions such as limited land area [15]. Complex networks are also common and include small networks in which there is usually high transitivity and low average path lengths [209]. Scale free networks are also complex and typically contain a small number of highly connected nodes or hubs. Knowing the type of network is therefore useful in outlining the general expected behaviour along the network. Two methods are used to further investigate networks, namely geometric and graphical analysis.

2.3.1 Geometric analysis

Geometric analysis aims to provide a basic description of the spacing, shape, orientation, density and geometric pattern of the network [215]. Three common geometric metrics include heterogeneity, connection patterns and continuity. Firstly, heterogeneity considers the variation in road classification where roads are distinguished by properties such as traffic flow, functional class (e.g. divided motorways, undivided motorways, ramps and collectors) and operational performance (e.g. speed and riding comfort). Heterogeneity is measured using an entropy measure defined as [177]

$$H(X) = -\sum_{i=1}^{m} p_i \log_2 (p_i)\,, \tag{2.5}$$

where m represents the number of road classes and p_i denotes the proportion of roads in system X belonging to class i.

When entropy is close to 0, we have a homogeneous road network. This means that mobility will be relatively uniform across the network as the effects of infrastructure, traffic and travel regulations are constant.

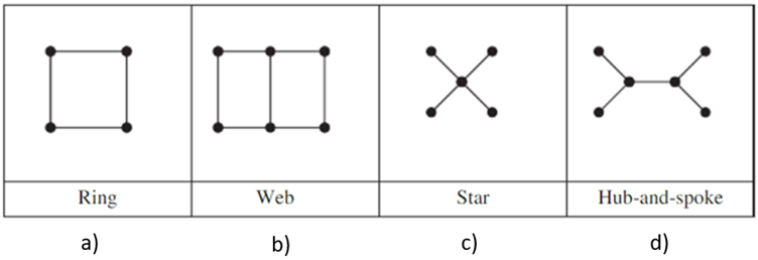


Figure 2.3: Different subnetwork connection patterns in networks [215].

Secondly, connection patterns consider the shape and geometric pattern of arterial roads, which form subnetworks. Since road patterns are often determined by surrounding infrastructure, these patterns can provide insight into any correlation between mobility and infrastructure [185]. As shown in Figure 2.3, there are four primary types of connection patterns. Figure 2.3 a) and b) are classified as circuits. A circuit is a closed path between nodes, or otherwise an arrangement of nodes in which there are at least two different paths to reach one node from another [215]. The first circuit pattern is a ring pattern, which consists of only

one circuit, while the second pattern is a web that has multiple conjoined circuits.

Figure 2.3 c) and d) show examples of branching patterns, in which paths follow a tree structure without completing a circuit. Branching patterns are either stars, which have a single hub from which other paths branch, or hubs-and-spokes, which have multiple hubs [215]. Branching patterns have no circuits. The proportion of different circuit patterns in the network can further be quantified by calculating the degree of ringness (Φ_{ring}), webness (Φ_{web}), circuitness ($\Phi_{circuit}$), and treeness (Φ_{tree}) as follows:

$$\Phi_{ring} = \frac{\text{length of arterial roads in ring patterns}}{\text{total length of arterial roads}} \tag{2.6}$$

$$\Phi_{web} = \frac{\text{length of arterial roads in web patterns}}{\text{total length of arterial roads}} \tag{2.7}$$

$$\Phi_{circuit} = \Phi_{ring} + \Phi_{web} \tag{2.8}$$

$$\Phi_{tree} = 1 - \Phi_{circuit} \tag{2.9}$$

Lastly, continuity considers that traffic flow is directly related to users' convenience and comfort when using roads as well as the consistency of the roads. The inconvenience of encountering discontinuity and having to either transfer to a different class of road along the same route or transfer to a new route altogether leads to user resistance [219]. Mobility is therefore likely to decrease in these areas and be diverted to continuous subnetworks. The discontinuity for a trip along the shortest path is defined as

$$Y(P) = \sum_{a \in \{P\}} y_a, \tag{2.10}$$

where $y_a = |k_1 - k_2|$, and k_1 and k_2 represent the hierarchy levels of the first and second class of road between which travel is transferred.

The discontinuity of the whole network is defined as

$$Y = \frac{\sum_{\text{all } (R,S)} Y\left(\{P_{RS}\}\right) \times q_{RS}}{\sum_{\text{all } (R,S)} l\left(P_{RS}\right) \times q_{RS}}, \tag{2.11}$$

where P_{RS} is the shortest path between R and S, $l(P_{RS})$ is the length of the shortest path and q_{RS} is the number of trips between R and S.

2.3.2 Graph analysis

Graphical analysis focuses on the underlying relationship between the graph's nodes and edges [25]. This allows the quantification of abstract concepts such as centrality, connectivity and spread [191]. Firstly, centrality is beneficial in identifying the most influential nodes within the graph. These would either correspond to the road intersections with the most branching traffic, or the locations with the highest visitation frequency, depending on the network being analysed. The importance of a node can be defined, and subsequently measured, in many different ways. Degree centrality measures the number of adjacent nodes, which provides insight into

how connected a node is [191]. Closeness centrality considers the inverted distance of the average shortest path from one node to all other nodes [172]. It depends on the geographical position of the node within the network, given that nodes in the middle of the network will have noticeably higher centralities. This makes it a suitable measure for spatial networks. Another measure of centrality is betweenness centrality, which is measured as the total number of shortest paths within the whole network that pass through a specific node [191]. This indicates which areas of the network are most efficient. All three of these measures of centrality provide insight into which parts of the network are most frequented as well as the overall distribution of activity across the network. The calculation of these measures, however, are dependent on the size of the network and can therefore not be used for comparing networks of different sizes.

A second concept addressed by graphical analysis is connectivity, which is largely measured with indices. These graphical indices have the distinct advantage of representing complex, inter-dependent structural properties, since they are calculated as a ratio of one measure to another [25]. Common connectivity indices include the α-, β- and γ-index. The α-index $\alpha = \frac{u}{2v-5}$ considers the number of cycles u compared to the maximum number of possible cycles in the graph. The number of cycles is estimated as $u = e - v + p$ where p is the number of subgraphs within the overall graph. The β-index $\beta = \frac{e}{v}$ measures the ratio between the number of edges and vertices. The γ-index $\gamma = \frac{e}{3(v-2)}$ considers the ratio between the number of links and the maximum number of possible links. The higher these indices are, the more complex and connected the network is. This is an important measure, as complex networks follow different node and edge distributions compared to simple networks with lower connectivity [191]. It is also linked to related measures such as accessibility, which can be calculated with edge and node densities [113]. These measures of connectivity are important as they indicate to what degree each node is responsible for network branching and how network development is distributed. These measures, however, are restricted since single indices only capture a limited amount of information about a complex structure [191].

Spread can also be measured with graphical indices [191]. The π-index considers the ratio of the total length of the network compared to the diameter. On the other hand, the η index calculates the number of network arcs. Spread is useful in determining how developed a network is as well as the estimated length of each edge. This can be used in further analysis since mobility behaviour has previously been shown to differ depending on the length of the road being travelled [191]. However, these again have the same informational limitations as mentioned above.

Additionally, community detection can be applied in graphical network analysis to determine which nodes are more closely connected and clustered together. A common community detection algorithm is the Louvain algorithm which aims to maximise modularity [162]. Modularity compares the degree of connectivity within communities to the degree of connectivity between communities [144, 45]. This can be used in mobility flow prediction as people are more likely to move along roads within the same community.

Overall, networks are helpful as they allow analysis of location importance, network activity distribution and the effect of distance on visitation frequency. Networks enable the comparison of mobility flows between different locations as well as flows for the same area across different time periods [48]. They aggregate mobility

flow for both individuals and larger populations, thereby allowing simple flow prediction between locations [53]. Tailored analysis is possible by weighting edges in line with additional data such as road width, load capacity, and safety and performance indices [74]. Previous studies have also used accessibility, direct connections and distance matrices to analyse internal structures of road networks [185]. These methods, however, still do not take into consideration the distance or orientation of line segments within the context of real physical space.

2.4 Texture classification

Given how vast and complex a linear network can be, it is useful to analyse subgraphs within the overall network [143]. A subgraph is defined as a graph with a vertex and edge set that is defined as a subset of the overall network's vertices and edges [168]. Commonly, subgraphs are extracted using geometric and topological methods. However, for the purpose of this mini-dissertation, we do not want to classify the subgraphs based on structural characteristics but rather on textural characteristics. This will give the best approximation of road condition classes.

Broadly speaking, texture refers to the variation across a surface [171]. More specifically, it can be defined as the spatial arrangement of grey levels of pixels within a region of an image [153]. It can represent different degrees of smoothness, coarseness, depth and regularity within a specific region. These textures will then carry meaning within the context of the image. For example, the texture of a road indicates its condition. The analysis of texture itself can be approached in many ways, but is generally divided into three main stages, namely pre-processing, feature extraction and texture classification [125].

The first step in pre-processing is converting red, green and blue (RGB) imagery to greyscale imagery since many feature extraction methods rely on the pixels' greyscale distribution. Additionally, greyscale images also reduce computation as the image is reduced from three dimensions to one dimension [147]. The equation for greyscale conversion is given as [102]

$$G(i,j) = \frac{B(i,j) + Gr(i,j) + R(i,j)}{3}, \tag{2.12}$$

where $B(i,j)$, $Gr(i,j)$ and $R(i,j)$ denote the values for blue, green and red of the pixel in row i and column j. The values will vary between 0 and 255 where 0 denotes black and 255 represents white.

The second step in pre-processing is image segmentation. The image being analysed is segmented into regular, pre-defined squares. The size of the regions of interest needs to be kept consistent and a balance needs to be found between smaller regions that are more accurate and larger regions that are more scalable [181]. This is a sufficient method of segmentation if texture is not expected to change drastically or quickly. In cases where there are, however, drastic texture changes, the image would need to be partitioned into regions of homogeneous texture as part of texture discrimination [88].

Feature extraction, derived originally from pattern recognition, refers to extracting significant features from an image among irrelevant details that are able to numerically describe and distinguish texture properties [176, 125]. The precision of classification will depend on extracting only the most discriminant features and

keeping the number of features to a minimum [197]. These features can include histograms characteristics, second-order statistics and estimated parameters to statistical models. To extract these features, we can approach texture from a structural, statistical, model-based or transformed-based perspective [125].

Structural methods define texture as the spatial arrangement of primitives in recurring patterns, such as regularly spaced parallel lines [66]. Textures are distinguished by the properties and placement of each of these patterns within an image. Structural analysis, however, has been noted to only really be reliable when used with regular synthetic textures and therefore has limited application to natural textures [106].

Model-based methods structure grey level values as inputs in models such as autoregressive models, moving average models and Markov Random Field Models [80]. Parameters are subsequently estimated for these models, and images with similar estimated parameters are considered similar. These methods are generally rotationally invariant, are able to model isotropic and anisotropic textures, and allow for efficient parameter estimation. However, depending on the complexity of the model, computational complexity begins to increase exponentially.

Transform-based methods include filter banks, Gabor decomposition-based approaches, Fourier transform-based approaches and wavelet-based approaches [80, 70]. They are beneficial in the case of transformation variance as well as any scenario requiring robust multi-resolution decomposition. These methods, however, are generally not rotationally invariant. They can also be non-orthogonal, which leads to computationally wasteful and redundant features.

Statistical methods, on the other hand, focus on the spatial distribution and relationship between pixel grey levels. Second-order statistics are used, which means analysis is done on pairs of pixels rather than individual pixels [66]. These methods have already been shown to outperform structural and certain transformed-based methods [211]. The focus of this mini-dissertation will be on two common statistical methods, namely the grey level co-occurrence matrix (GLCM) and linear binary patterns (LBP) with a justification for this choice discussed below.

2.4.1 Grey level co-occurrence matrices

Grey level co-occurrence matrices (GLCM) define texture as a statistical function of the varying pixel grey levels [67]. The GLCM matrix is constructed so that the number of rows and columns total the number of grey levels, G, within the image. An example is shown in Figure 2.4 a) of a 5×5 image with four grey levels. Each pixel contains one grey level value. A relative frequency matrix P is constructed where the element $P(i, j \mid d, \theta)$ represents the relative frequency with which two pixels, one with a grey level intensity i and the other with a grey level intensity j, occur within the same neighbourhood while separated by a pixel distance d and rotated at an angle θ. We can formally define the $M \times N$ image neighbourhood that contains G grey levels ranging from 0 to $G - 1$. The intensity of sample m along line n is defined as $f(m, n)$. As such, we can express the relative frequency matrix as

$$P(i, j \mid \Delta x, \Delta y) = WQ(i, j \mid \Delta x, \Delta y) \tag{2.13}$$

where

$$W = \frac{1}{(M - \Delta x)(N - \Delta y)}$$
$$Q(i, j \mid \Delta x, \Delta y) = \sum_{n=1}^{N-\Delta y} \sum_{m=1}^{M-\Delta x} A \tag{2.14}$$

with

$$A = \begin{cases} 1 & \text{if } f(m, n) = i \text{ and } f(m + \Delta x, n + \Delta y) = j \\ 0 & \text{elsewhere} \end{cases}. \tag{2.15}$$

This can more plainly be explained by saying that each left-to-right horizontal pixel intensity pair is considered for the final relative frequency in Figure 2.4 c). These pairs are shown in Figure 2.4 a) by any two adjacent pixels with the same colour. In the example, there are 16 different pairs ranging from $(0, 0)$ to $(3, 3)$. The corresponding frequency of each pair is entered into the frequency matrix, shown in Figure 2.4 b). For example, in Figure 2.4 a) the pair $(2, 3)$, denoted in dark purple, occurs 5 times. This is shown in Figure 2.4 b) where 5 is assigned to the pixel corresponding to $(2, 3)$. It should be noted that the matrices are not symmetrical, so $(i, j) \neq (j, i)$. Figure 2.4 b) demonstrates this as the pixel corresponding to $(3, 2)$, denoted in light purple, is assigned 2 and not also 5. Finally, the relative frequency is derived by dividing by the weight W. Figure 2.4 c) shows that in this example, all the tallies were divided by $W = 20$.

To ensure the texture classification is rotationally invariant, the matrix is usually calculated for four different distances, namely $d = \{1, 2, 3, 4\}$ where $d = 1$ represents pixels right next to each other and $d = i+1$ represents pixel pairs with i pixels between them. Similarly, the matrix is also calculated for $\theta = \{0°, 45°, 90, 135°\}$ where $\theta = 0°$ represents the original image and each subsequent rotation rotates the image anti-clockwise. The relative frequency matrix $P(i, j \mid d, \theta)$ is calculated for all 16 combinations of the d and θ parameters, and the mean for each pixel is taken.

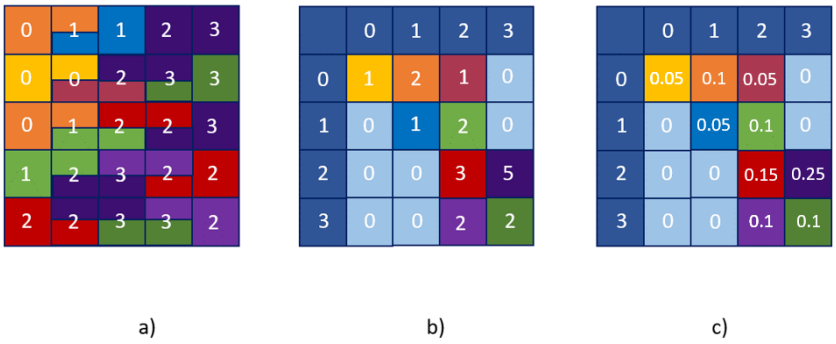


Figure 2.4: GLCM process explained with the a) original input image, b) frequency matrix, and c) normalised matrix [165].

From the final relative frequency matrix $P(i, j)$, n features can be derived by various transformations and summations of the matrix elements, where $1 \leq n \leq 14$. Common features include energy, entropy, contrast, homogeneity and dissimilarity [67, 160]. This method has the advantage of allowing different combinations and weightings of features to be tested [165]. The efficiency of GLCM in remote sensing has been proved despite GLCM originally being used for 2D textures [79, 204]. There are, however, still certain limitations to this method. While the GLCM method is rotationally invariant by taking the average frequency matrices across

multiple θ values, there is not an equivalent approach to ensure scale invariancy. For example, if two images of rocky roads are compared at different scales and resolutions, one might be mistakenly misclassified as a gravelled road instead. The second limitation is computational cost [165]. Given that large matrices (usually 256×256) need to be calculated 16 times for each subimage and certain data sets can contain thousands of subimages, this method is known to be slow and inefficient on large data sets [163].

2.4.2 Linear binary patterns

The method of linear binary patterns (LBP) is computationally simple. It has been reported as the foremost statistical method for texture analysis used in remote satellite imagery [199, 12]. Depending on the chosen resolution of the image M, the image, assumed to be processed as a square, consists of M^2 pixels. Each of the M^2 pixels are sequentially selected to be the centre pixel. Each centre pixel x_i is assigned a subimage y_i with dimensions $(2R+1) \times (2R+1)$ where R denotes the radius of a neighbourhood circle. The chosen radius determines the number of neighbourhood pixels $P = R^2 - 1$ that will be used to calculate the final LBP values. A common radius used is 1. For example, in Figure 2.5 a), the chosen radius is $R = 1$, the number of neighbouring pixels is $P = 8$ and the subimage is a 3×3 matrix.

LBP defines the spatial pattern of a texture by comparing the greyscale value of the central pixel with the greyscale value of its neighbouring pixels [149]. Neighbouring pixels form a binary pattern by assigning pixels with a smaller greyscale compared to the centre pixel a 0 and pixels with a greater greyscale a 1. Figure 2.5 b) demonstrates this with the top left corner being assigned 0 because $12 < 50$ while the top right corner is assigned 1 given that $70 > 50$. The LBP value for the centre pixel x_i is calculated as

$$LBP_{P,R}(x_i) = \sum_{P=0}^{P-1} s\left(g_p - g_0\right) 2^p \tag{2.16}$$

where

$$s(x) = \begin{cases} 1 & if\ x \geq 0 \\ 0 & if\ x < 0 \end{cases}$$

with g_0 representing the greyscale of the central pixel and g_p representing the greyscale of the neighbouring pixels.

The LBP value for the example in Figure 2.5 c) is calculated by multiplying the binary pixel value assigned in 2.5 b) with 2 raised to the power of the pixel's index. The pixels are indexed starting at 0 in the top left corner and then moving clockwise. The final LBP value is therefore

$$LBP_{8,1} = 0 + 0 + 4 + 0 + 16 + 32 + 0 + 0 = 52 \tag{2.17}$$

This process is repeated for all the pixels x_i, $i = 1, ..., M^2$. However, from the above example, it is clear that the LBP feature will vary if the neighbouring pixels are rotated. A uniform and rotation invariant descriptor

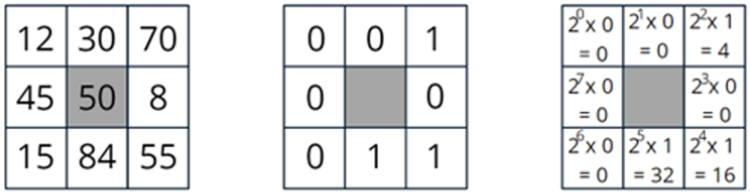


Figure 2.5: An example of the LBP process. The original pixel values are segmented into a) the input square tile. Next, b) binary values are assigned to each pixel where neighbouring pixels lower than the shaded centre pixel are assigned 0 and 1 when they are higher. Finally, c) the neighbouring pixel LBP values are calculated as the pixel's value multiplied by 2 raised to the power of the pixel's index.

$LBP^{riu2}_{P,R}$ for the LBP value is thus defined as [149]

$$LBP^{riu2}_{P,R} = \begin{cases} \sum_{p=0}^{P-1} s\left(g_p - g_c\right) & \text{if } U\left(LBP_{P,R}\right) \leq 2 \\ P+1 & \text{otherwise} \end{cases} \tag{2.18}$$

where

$$\begin{aligned} U\left(LBP_{P,R}\right) &= \left|s\left(g_{p-1} - g_c\right) - s\left(g_0 - g_c\right)\right| \\ &+ \sum_{p=1}^{P-1} \left|s\left(g_p - g_c\right) - s\left(g_{p-1} - g_c\right)\right|. \end{aligned} \tag{2.19}$$

The $U(\cdot)$ refers to the uniformity of a neighbourhood of pixels. Uniformity is defined as the number of bitwise transitions made either from 0 to 1 or 1 to 0 within the neighbourhood. If $U \leq 2$ then the pattern is determined to be uniform [150]. The method is based on the theory that uniform patterns exist within the quantization of the angular space across any spatial resolution and that scale invariance can be accomplished by using a combination of multiple multi-resolution analysis operators [104]. Additionally, rotational invariant variance measures used to describe the grey level contrast of the images have also been proved effective in attaining rotational invariance [104].

The uniform and rotational invariant LBP descriptor is thus used to assign an LBP value to each pixel x_i. The M^2 LBP values are plotted as a histogram to summarise their distribution. The number of bins used depends on the level of proposed classification granularity. More bins would allow for more detailed discrimination between various textures and result in more texture classes [90].

Features can be derived from the histogram, namely the mean, standard deviation, median, skewness, kurtosis and entropy [137]. Alternatively, similarity between the two histograms could be determined by distance. The χ^2-statistic, the Wasserstein-Kantorovich distance metric or dynamic time warping [112, 89, 2] could all be used. Support vectors have also been used to classify texture classes based on a multivariate vector of histogram features [35]. Methods of 'histogram intersection' have specifically been used as an appropriate kernel function for the support vector machines [16]. Additionally, the discriminative power of the histogram features themselves can be increased, by transforming the features using the Dirichlet Fisher kernel [95]. Other methods include deriving subsequent features like feature ratios [156].

Despite the prevalence of the LBP method in remote sensing, one limitation to note is its hindered

performance in shadowed regions [90]. This affects real-world applications as most remote sensing imagery includes shadows and a lack of mitigation could lead to many cases of texture misclassification. A more robust texture extraction method is therefore needed.

2.4.3 Feature concatenation and K-means clustering

A widespread approach in texture analysis today is combining features from multiple feature extraction methods. This is effective for varying texture surfaces, such as roads with intermittent shadows [51]. It also aids in reducing significant computing time for marginal gains in accuracy [220]. One study demonstrated that combining LBP with the GLCM contrast and energy features improved non-shadow classification from 90% to 100% and shadowed contrast from 85% to 97%. In addition to GLCM features, visual perception features defined by Tamura et al. [186] could also be considered for combination. These include contrast, regularity, roughness, coarseness, directionality and line-likeness, regularity, and roughness.

Once the features have been extracted, they need to be categorised. Given the characteristic lack of data on informal roads, there is no ground truth on the actual number of road conditions, and therefore neither the number of texture classes, nor which line segments correspond to which class. As such, unsupervised classification is used by means of K-means clustering [69]. The number of preferred texture classes can be specific as k. Different k values can be tested and the optimal number of classes can be determined using the Akaike's information criterion (AIC), Bayesian inference criterion (BIC), elbow method or cross-validation [96]. Data points are defined in K-dimensions, where K denotes the number of features. Next, K arbitrary centroids are chosen, each data point is assigned to the closest centroid and then the K centroids are recalculated based on the aggregate location of the newly assigned data points. This is repeated until the centroids no longer move, and the algorithm has minimised the objective function

$$J = \sum_{i=1}^{N} \sum_{k=1}^{K} w_{ik} \left\| x_i - \mu_k \right\|^2$$

$$w_{ik} = \begin{cases} 1 & \text{if } x_i \text{ is in cluster } k \\ 0 & \text{otherwise.} \end{cases} \tag{2.20}$$

where N denotes the number of observations and μ_k is the centroid of cluster k.

2.4.4 Conclusion

In this chapter, the necessary background theory on linear networks was discussed. A linear network was defined and the two different spatial processes for which a linear network can be applied, as well as how the spatial process determines the definition of the network edges and vertices, was explained. The different methods of representing a linear structure were also discussed, and it was shown why the graphical method was preferable. Given the preference for a graphical representation of a linear network, common network analysis metrics were discussed and various limitations to the existing analysis methods were highlighted. Finally, two

well-proven methods of texture classification were described to lay the foundation for how this can later be applied to linear networks.

The next step in Chapter 3 is to consider two linear networks, namely the social mobility network and the network of informal roads. These are examples of linear networks with an underlying point process and an underlying line process, respectively. The linear networks will either be constructed or extracted based on the relevant theory. Network analysis will be performed on each linear network and the shortcomings of simply comparing spatial characteristics will be demonstrated.

Chapter 3

Data

3.1 Mobility data

3.1.1 Data description

Social mobility data in northwestern Namibia was collected from a 2008-2009 survey with the purpose of better understanding mobility behaviour in rural areas. The focus of the data set is on how mobility behaviour, specifically in terms of travel path and frequency, is influenced by the existing informal road networks. The mobility data indicates which villages people frequently travel between and where the mobility routes likely lie. The data tracks the movements of 1300 participants and includes answers to questions such as 1) which village and region the participants originated from, 2) which village and region they travelled to, 3) which season they travelled in, 4) their means of transport, 5) their reason for travelling, and 6) how long they stayed at each particular location.

The original data included 27 origin and 266 destination villages, but data points that either did not include coordinates or did not have a searchable village name were excluded. Ultimately, 405 data points were removed, and 195 villages were excluded altogether. The average visitation count for each of the excluded villages was 2.05, so the majority of mobility flow is still included in the final data set. The final data used includes 27 origin villages and 70 destination villages within the Kunene region, with some villages being both origin and destination locations. The distribution of these villages is shown in Figure 3.1.

Figure 3.1 shows that the origin villages are all clustered relatively close together. Many of the destination villages, on the other hand, are spread out. When considering which trips are likely to follow informal roads more closely, it can be expected that destination villages close to origin villages will be relatively spatially similar since there is less time and distance to stray from the short, efficient paths. On the other hand, when villages are far away, individuals may encounter many hindrances such as rivers, mountains and cliffs that force them to take detours and deviate from the linear mobility route.

Figure 3.2 demonstrates that there are clearly routes that experience higher volumes of flow. While the majority of villages, shown in red, have under ten visits, the few villages that have more visits show an exponential increase. There are 19 villages with between 10 and 50 visitations, as seen in green, and 3 villages

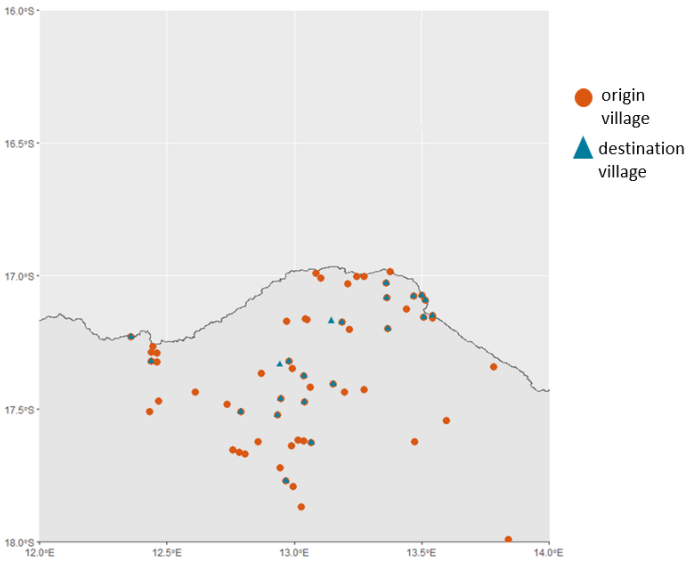


Figure 3.1: The distribution of origin and destination villages, visualising both the number of villages as well as their clustering patterns.

with more than 50 visitations, as seen in orange. The most visited village Opuwo far outperforms all the other villages with 156 visits. The frequency of visits may also affect the degree to which the mobility path adheres to the road network. As more people travel between two locations, the more efficient travelling paths are likely to become over time as people will learn from one another and eventually find the fastest way to reach their destination. This optimal path, apart from unavoidable deviations due to immovable geographical features, should approach the theoretical shortest path represented by the mobility data.

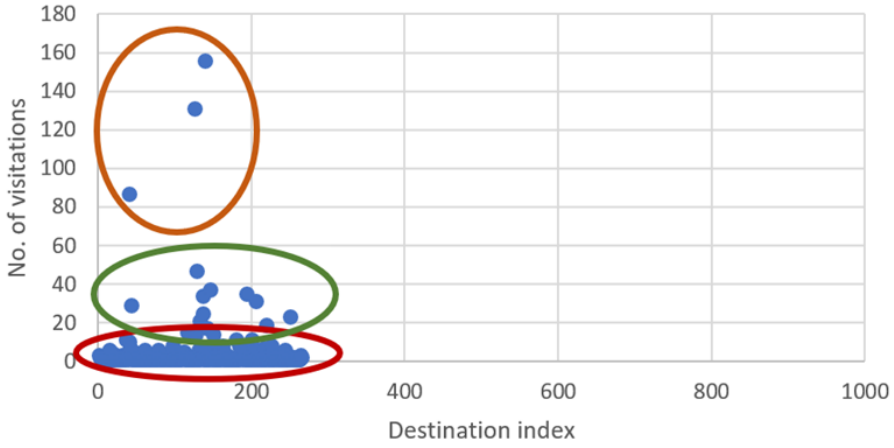


Figure 3.2: A scatter plot of the number of visitations per destination, showing three distinct types of villages. The first type of village, shown in red, occurs most frequently, with 49 villages having less than 10 visits. The second type of village, shown in green, represents 19 villages with between 10 and 50 visits each. The third type of village, shown in orange, only occurs three times and represents the most visited villages with more than 80 visits each.

Specific questions were also asked concerning travel to the three larger cities Opuwo, Okanguati and Epupa which could facilitate more in-depth spatio-temporal analysis of these areas. The map in Figure 1.2 highlights these three cities in addition to another large city, Ohayuuwa. From the map, it is clear that most of the movement occurs between these four villages, as seen by both the number and width of the edges. The thickest lines are seen to emanate from Opuwo and to Ohayuuwa, representing that the highest volume of

people are travelling to and from these villages, respectively. Most travel to Opuwo occurs over long distances whereas the flow of mobility for Okanguati, Epupa and Ohayuuwa seem to occur mostly over short distances. Figure 1.2 also shows that most of the villages around the fringes have only one thin edge. In most instances, only one person visited that particular village. It is possible that these routes will be significantly dissimilar from road networks since there is likely very little infrastructure, including informal roads.

Another consideration is that most people surveyed have vulnerable means of transport, as shown in Figure 3.3 a). The majority of people walk, ride animals like donkeys and horses, or drive in vehicles. As opposed to cars that are constrained to wide roads that cannot traverse different kinds of geography, walking and riding animals allow for more flexibility in cases where the shortest path between two locations may require crossing difficult terrain and circumventing features such as mountains and rivers.

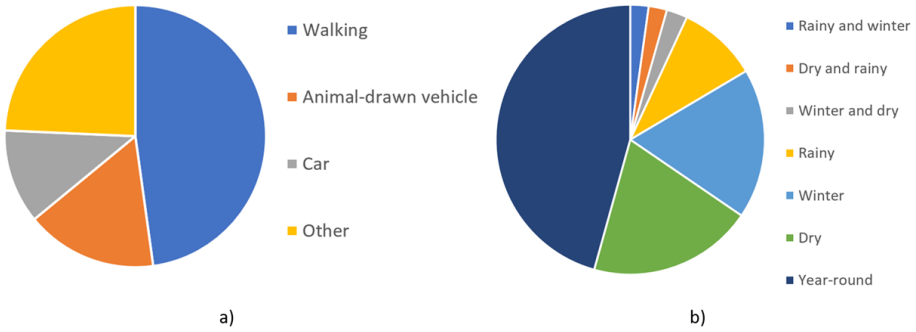


Figure 3.3: Pie charts describing the demographic trends in a) means of travelling, and b) seasons for travelling.

The effect of seasons should also be considered. Seven different seasonal conditions, predefined by the social scientist conducting the surveys, were included in the survey, namely 1) rainy, 2) winter, 3) dry, 4) year-round, 5) rainy and winter, 6) winter and dry, and 7) dry and rainy. These seasonal groupings are assumed to have idiosyncratic meanings within the rural northwestern region of Namibia. People in the rural communities do not necessarily divide the year equally into the four traditional seasons, but instead likely define time by other more intuitive, practical markers such as harvesting months, flooding months and seasonal job months. These divisions were then best approximated by defining them in terms of temperature and precipitation. Figure 3.3 b) shows that over half the surveyed individuals travelled either year-round or during multiple seasons. These individuals can be considered non-seasonal travelers as their trips were not confined to a single season. . This initially suggests that different seasons, and their subsequent effect on road conditions, do not greatly impact the volume of social mobility traffic. It should be noted that while the fact that people decide to travel remains generally unchanged by the seasons, this does not necessarily mean that the seasons do not change how and where they travelled.

Table 3.1 shows that the underlying characteristics of the seasonal routes differ significantly. Many people travelled year-round, but they travelled a comparatively shorter distance. Longer journeys may not always be feasible year-round due to geographical barriers like river crossings during the rainy reason. People's tolerance for travelling long distances also generally decreases with prolonged exposure to inhospitable weather conditions like hot and dry, cold or wet. This is especially when travelling by foot or animal. Additionally, there are many different distinct year-round routes across the whole Kunene region, which suggests there is no regional

	Rainy	Winter	Dry	Year-round	Rainy and winter	Winter and dry	Dry and rainy
Contribution to overall mobility volume (%)	9.6	18	19.8	45.7	2.2	2.5	2.2
Number of distinct mobility routes	3	96	88	140	12	14	10
Average length of mobility routes (km)	16.3	47.4	56.7	40.9	60.5	59.8	51.1

Table 3.1: Descriptive measures of the seven different seasonal routes.

effect on people's seasonal resilience. On the other hand, a moderate number of people indicated that they only travelled during the winter or dry seasons. There were many different routes during this particular time period, but these trips were, on average, longer than the year-round trips. It is possible that these trips are only feasible once the river beds dry up in the dry seasons or the unstable muddy roads harden during the cold winter seasons. Furthermore, a smaller number of people stated that they exclusively travelled during the rainy season. It can be seen that there were only a few short-distance trips travelled during this season. The restricted travel is likely due to the difficulty of travelling on dirt roads in the rain. Lastly, a very few number of people only travelled during the rainy and winter, winter and dry, and dry and rainy seasons. There were only a small number of trips, and they were, on average, long distance. These routes may again be driven by seasonal accessibility. Alternatively, the routes may be determined by mandatory reasons for travel like medical attention and are in reality perhaps independent of the seasons. The sample sizes for these three seasonal routes, however, are simply too small to draw any more reliable conclusions.

Further analysis is required to understand what effect seasonal changes have on the mobility routes themselves. This will show if seasons cause changes in underlying mobility behaviours such as choice of destination, general tolerance of long distances, and use of weather-dependent access like river crossings during the dry season.

3.1.2 Network analysis

The mobility data also includes the geographical coordinates for each location mentioned in the survey for the purpose of mapping. As shown in Figure 1.2, the locations of the origin and destinations are mapped across the Kunene region and the corresponding mobility routes are interposed. When representing the data as a linear network, there are 71 nodes and 281 edges which can clearly be seen in the graphical representation of the linear network in Figure 3.4. The nodes are the unique subset of both the origin and destination villages while the edges are the shortest distance path between these villages. The nodes and edges can be used in many ways to derive characteristics of the network.

Firstly, node degree and betweenness centrality can be considered, as shown in Figure 3.5 a) and b). The degree of a node denotes the number of connections made to that particular node [26]. The higher the node degree, the more people either travel from or to that particular village. The distribution of node degrees corresponds with the mapping in Figure 1.2 since the nodes with the highest degree align with the most populous villages seen with the heaviest traffic branching out from them.

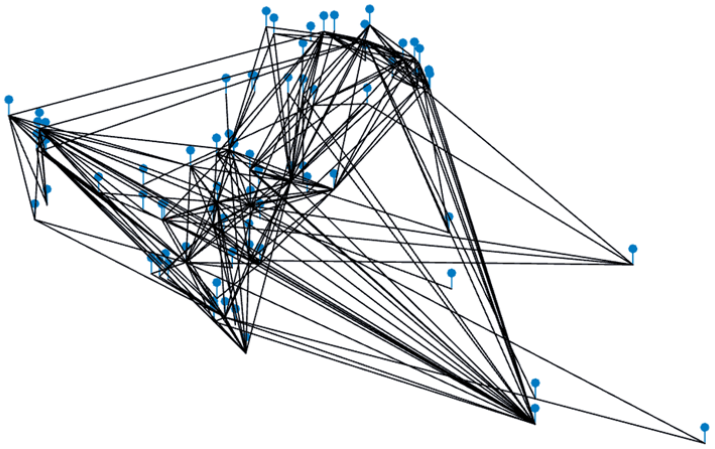


Figure 3.4: Mobility data as graphical linear network with locations marked.

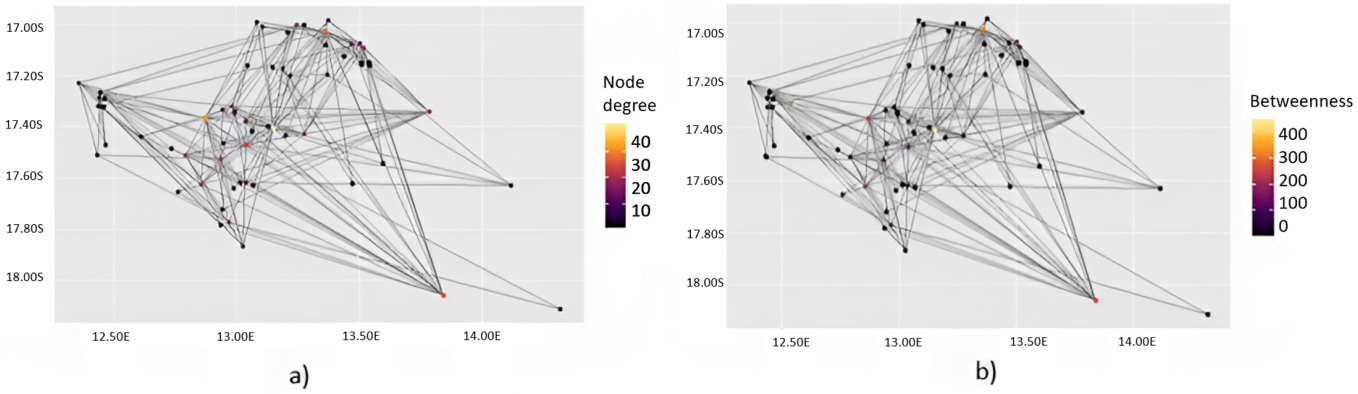


Figure 3.5: Analysis of social mobility network including a) node degree distribution, and b) node betweenness centrality for mobility data.

We also see that a majority of the remaining nodes have very low degrees, which correspond to infrequently visited villages far away from the four busy villages. These are therefore singular destinations which are out of the way and less likely to be part of people's regular commute. This is visualised in Figure 3.6 a). Due to the increased distance, the routes to these destinations are more likely to cross at least some areas of inconvenient terrain. This could lead to detour routes deviating from the shortest path. On the other hand, villages that are more interconnected can be expected to be visited throughout the year as they either have better infrastructure (since most interconnected villages correspond to the more populated villages) or they are an intermediary stop towards multiple other destinations. As can be seen in Figure 3.6 b), villages that are more interconnected also tend to be more centralised. Similarly, the betweenness centrality very closely matches the node distribution for most nodes. The centrality represents the number of shortest paths that pass through that particular node. It is an indicator of which locations, on average, are most likely to be passed through as part of a trip to a final destination. This again affirms which locations are most frequently travelled to and where we would expect a high density of roads to pass through.

The node degree and betweenness centrality should also be considered in light of which villages were surveyed and how many people participated in each village. The significant data points with a minimum of at least 25 participants are shown in Figure 3.7. It can be seen that, with the exception of Opuwo in the bottom right, all the areas with high node degree and betweenness centrality correspond very closely to the

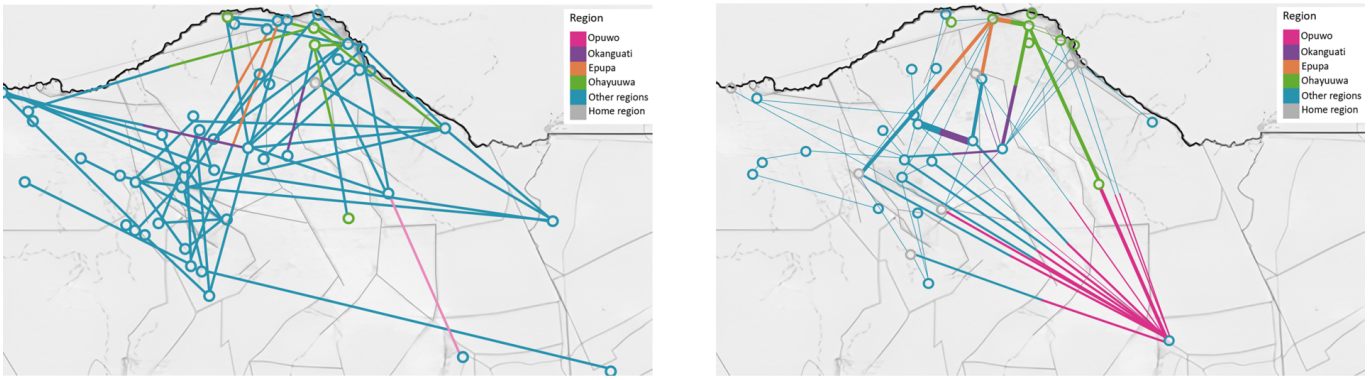


Figure 3.6: Examples of the social mobility network containing only a) single-visit destinations, and b) multi-visit destination.

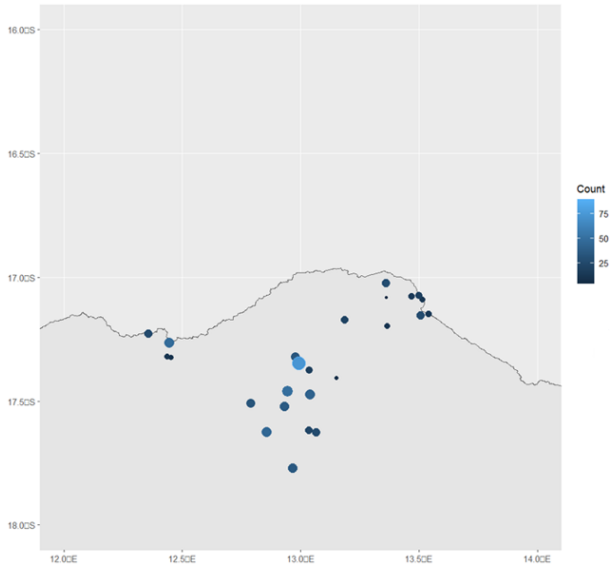


Figure 3.7: A mapping of the varying density of surveyed origin villages across the Kunene region.

areas with the most participants. The exception of Opuwo is due to the fact that Opuwo was not surveyed as an origin village. The focus of the mobility study was on people living in rural areas and so Opuwo is only included in the data set as a destination village. This could be considered as a form of sampling bias since high participation rates do appear to somewhat contribute to high connectivity. Given that the data was collected in rural areas with limited resources, this sampling discrepancy is expected. The effect of the bias, however, is partially limited. Many people often travel together or have similar mobility patterns and so the number of distinct mobility routes is usually far less than the number of participants. Varying participation rates are still expected to have an effect on the number and placement of mobility routes, but far less when compared to the effect on the frequency and weighting of mobility routes. Since weighted linear networks are beyond the scope of this mini-dissertation and the frequency of travel between origin-destination pairs is excluded, the bias is curbed. The bias, however, should still be taken into account during a final analysis to avoid any uninformed conclusions.

Lastly, mobility behaviour can also be analysed through community detection. Communities within network analysis are defined as clusters of nodes where the density of nodes within clusters are higher than the density between those nodes and the rest of the network [139]. The Louvain algorithm is a common community

detection algorithm [45]. The algorithm seeks to maximise the modularity of each community by prioritising sparse connections between nodes in different modules and dense connections between nodes in the same module. The community detection yielded 10 distinct communities. These graphical communities, as shown in Figure 3.8, may correspond to actual communities, as people who live closer together tend to behave more alike compared to people who live further away. This may be due to shared infrastructure and similar surrounding terrain. As such, the movements of various communities can be analysed to see whether certain groups behave more constantly throughout the year or are more dependent on the weather. Factors such as demographics, existing infrastructure and geography could then possibly be considered to determine what makes a certain community more or less constant in their mobility.

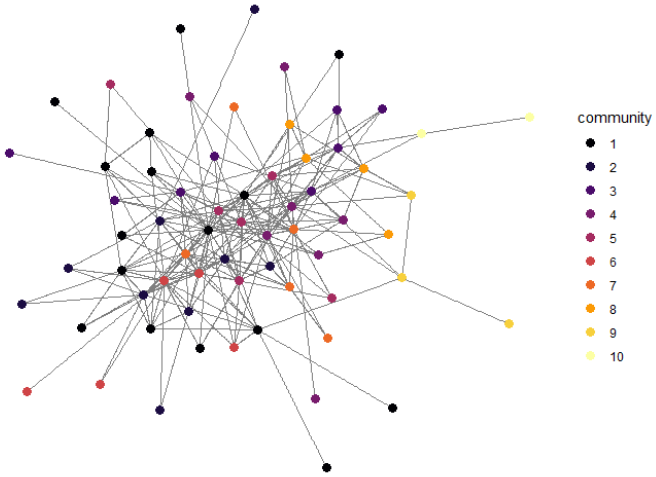


Figure 3.8: Communities within the social mobility network.

3.1.3 Geographically weighted Poisson regression model

Spatial models of mobility are one method of better understanding the underlying dynamics of the mobility data. The effect of location on mobility is taken into consideration by allowing variables that have been shown to have a significant effect on mobility to vary by location. This can be done using a geographically weighted Poisson regression model (GWPR). A GWPR is a spatial count regression model which specifically takes into account the effect of location on various predictor variables [4]. Previous studies have proven a localisation effect of infrastructure on mobility, where examples of infrastructure variables include the density of road networks, the number of car-worthy roads, and the presence of points of interest such as bus and taxi stations [154, 85]. The simplest response variable to model is the number of visits per location, but other approximations of mobility, depending on available data, have also been used, including the number of registered tourists, number of transport transfers and rates of disease spread [154, 85, 4].

For the purpose of this mobility data set, the number of visits per location is defined as the response variable. Three explanatory variables are considered, namely the size of the village, the density of roads within the village, and the distance of the village from the regional capital, and largest city, Opuwo. The final

location of a village is estimated by selecting the closest identifiable village to that village's given coordinates. The size of the village is estimated, as shown in Figure 3.9, by creating the smallest possible bounded area that still includes all visible village characteristics such as huts and cattle fencing. This method of bounding the area of the village according to consistent characteristics prevents inconsistent size estimates and also addresses the general tendency to overestimate the size of villages due to all the open surrounding area. Distance is calculated as the shortest Euclidean distance between Opuwo's coordinate pinpoint and that of the particular village. For road density, the total length of the road network within a circular search area centred at the village's coordinate pinpoint with a diameter of 1 km is first calculated. This total length is then divided by the area of the circular search area to derive the road density. The GWPR analysis is done in `R` using the `hoxo − m/gwpr` [140] package.

Figure 3.9: An example of how the area of a village is derived from remote sensing imagery.

The GWPR equation is expressed as

$$\mu_i \sim \text{Poisson}\left[N_i \exp\left(\beta_0\left(\mu_i, v_i\right) + \beta_1\left(\mu_i, v_i\right) x_{1,i} + \beta_2\left(\mu_i, v_i\right) x_{2,i} + \beta_3\left(\mu_i, v_i\right) x_{3,i}\right)\right] \tag{3.1}$$

where (u_i, v_i) represents the coordinates of location i while $\beta_0\left(u_i, v_i\right)$, $\beta_1\left(u_i, v_i\right)$, $\beta_2\left(u_i, v_i\right)$, $\beta_3\left(u_i, v_i\right)$ and N_i represent the intercept parameter, the coefficient of village area, the coefficient of village distance, the coefficient of road density and the offset variable at location i, respectively.

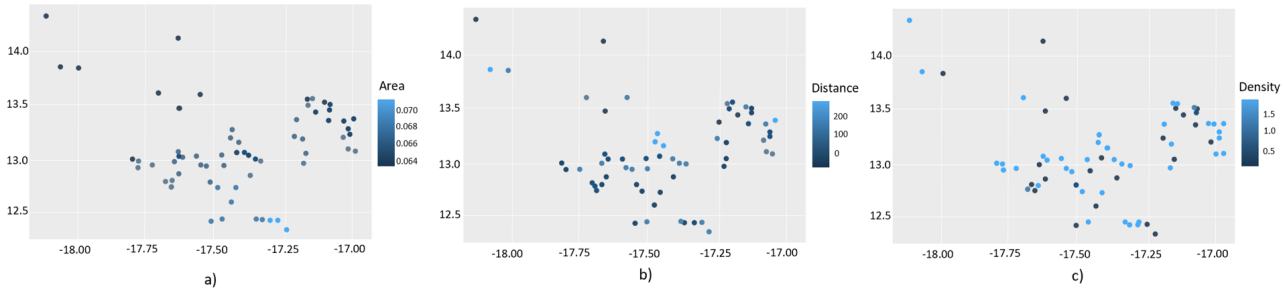


Figure 3.10: Scatter plots of geographically weighted values for a) the size of the area coefficient, b) the distance coefficient, and c) the road density coefficient.

From Figure 3.10 a), a clear geographical variation in the effect of village size can be seen with the effect of size being maximised in the southeast. This happens to be close to the regional capital Opuwo . In areas with the most populous villages, it is possible that smaller villages are more easily overlooked in favour of larger,

better developed villages. In more rural areas, however, there is no such comparison. On the other hand, there seems to be no consistent geographical variation in the effect of distance from the regional capital, as seen in Figure 3.10 b). Areas with a weak relation between distance and location visitation are marked in dark blue. When compared with the satellite imagery, it can be seen that many of these areas often correspond to mountainous or other harsh terrain. These areas are therefore more dependent on accessibility than distance while areas with flat, convenient terrain tend to have a stronger relation with distance. A similar effect can be seen in Figure 3.10 c) when looking at road density.

To demonstrate the advantage of the GWPR model, a generic Poisson model with the same response and explanatory variables was fitted. Goodness-of-fit for each of these models was compared using the biased-corrected AIC. With a bandwidth of b, the biased-corrected AIC is defined as [4]

$$\mathrm{AIC_c}(b) = D(b) + 2\left(K(b) + \frac{K(b)(K(b)+1)}{N - K(b) - 1}\right) \tag{3.2}$$

where D is the model deviance, K is the number of parameters and N is the number of observations.

Using an exponential kernel and a bandwidth of 1 based on similar previous work [4], the AIC_c for the GWPR model was 510.95, while it was 1142.6 for the generic Poisson model. The GWPR model therefore outperforms the simple Poisson regression model. The localisation effect of the variables has proven to lead to better predictive performance. Additionally, GWPR residuals have more spatial randomness, which indicates that the model takes into consideration the majority of spatial determinants on the response variable. The GWPR model, however, is limited to the data available. In this case, there are no official statistics on village size, road density or any other infrastructure metrics. All variables are derived from satellite imagery and locally surveyed coordinate data, which is not always very accurate. As demonstrated in Figure 3.11, the location data was sometimes incorrect and did not correspond to any nearby village. In these cases, approximations had to be made as to which closest villages should be used and which should be excluded. Another issue to be noted is the inevitable human error and inconsistency that may have occurred when compiling the informal road data set, which would in turn bias the road density variable. The GWPR model is also parametric and therefore, requires further optimisation when selecting parameters such as the bandwidth and kernel function. The GWPR model, therefore, while performing better than a generic Poisson model or other mobility flow models, still has problematic dependencies on non-robust variables and parameters.

3.2 Road network data

3.2.1 Data description

The road network data captures the most frequently travelled paths, either by vehicle or foot, between popular villages in the northwestern region of Namibia. Aerial photography of the Kunene region is analysed to extract an informal road network. As 67% of the Namibian population remains rural[1], the area chosen will provide a sufficient sample of elaborate and regularly used informal road networks. These include different types of

[1] https://www.fao.org/3/i9756en/I9756EN.pdf Accessed: 2021/03/21

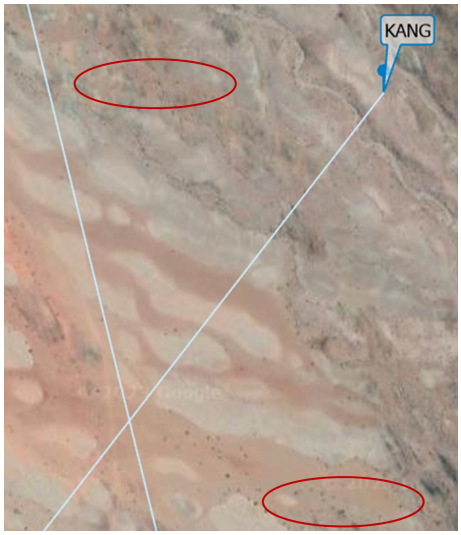


Figure 3.11: Example of incorrect location data for villages included in the mobility study.

roads, each with their own physical and environmental constraints. For example, roads in urban areas and villages curve around buildings, roads along forested areas follow the tree line for the sake of shade and roads in open areas are unobstructed as seen in Figures 3.12 a), 3.12 c) and 3.12 d), respectively. Physical features such as mountains and rivers, shown in Figures 3.12 b) and 3.12 f) also influence reasonable accessibility and placement of roads. These factors will affect to what extent road networks will be able to follow the shortest linear path between destinations. Additionally, areas close to villages and towns generally have more activity, and therefore more roads, when compared to remote transitional areas. This variation in road density will also affect how spatial similarity varies across an area since the complexity of the road network would be significantly structurally different from the sparser mobility networks.

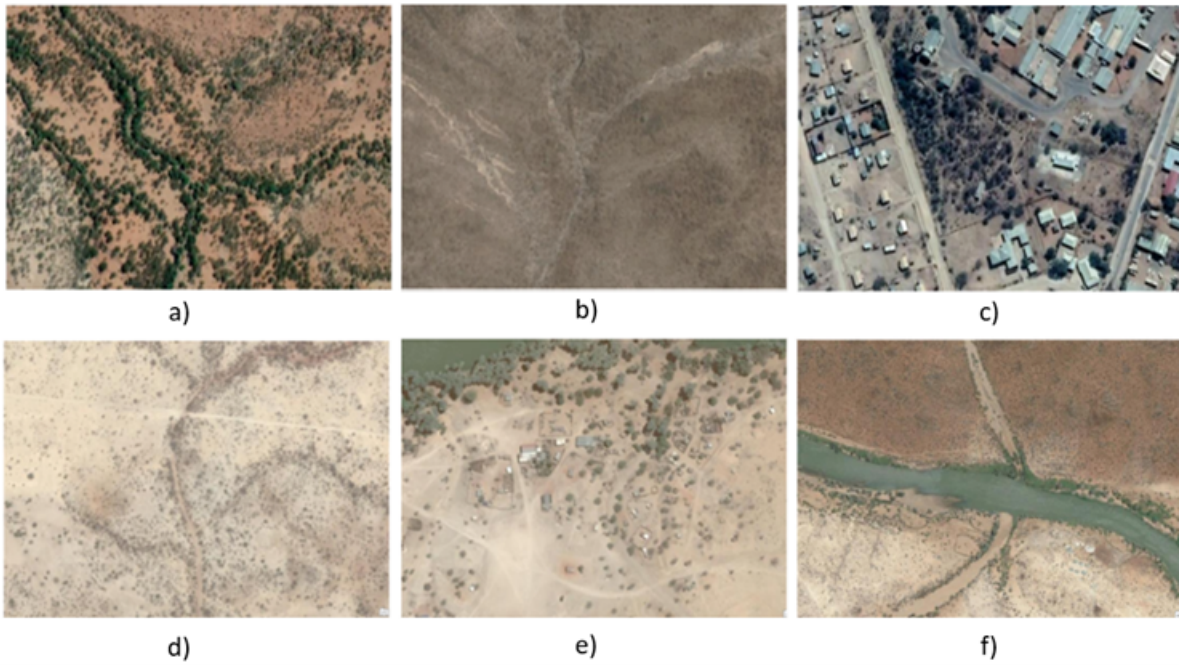


Figure 3.12: Satellite imagery of different types of roads including a) tree line walkthroughs, b) mountain passes, c) urban area driveable roads, d) rural area driveable roads, e) footpaths, and f) river crossings.

The informal road network is constructed by digitising all identified informal roads from Google Maps

images between a latitude of 16.9847°S and 18.1546°S and longitude of 12.3579°E and 14.3184°E [121]. The images are from the most recent 2023 Google satellite cycle. The coordinate bounding box is determined by the minimum and maximum coordinates of all included mobility data locations. The roads are digitised as polygons covering the total road surfaces at a map scale of 1:2000 m and a raster resolution of 1 m in ArcGIS Pro. The final network is derived by extracting the centre lines from these polygons. The full road network, shown in Figure 3.13, covers a total length of 13641.82 km^2 with approximately 10700 road segments.

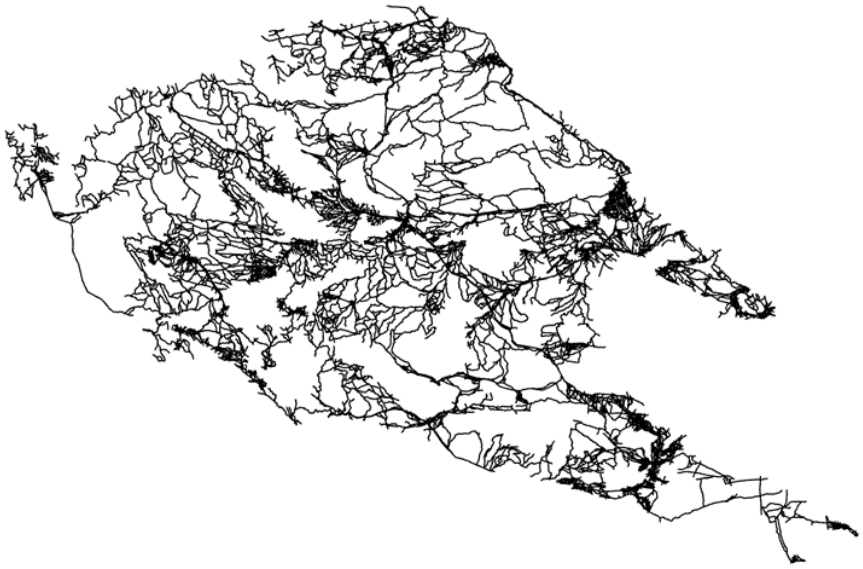


Figure 3.13: The final informal road network represented as a linear network.

3.2.2 Network analysis

The road network, once represented as a linear network in R, has 16121 nodes and 10689 edges. The nodes are either the intersections of the end points of roads, while the edges are the roads themselves. The fact that there are more nodes than edges already suggests that there are many branching roads that only lead to a singular destination without reconnecting to the overall road network. This can quantitatively be shown using node degree and betweenness centrality. As shown in Figures 3.14 a) and 3.14 b), the majority of nodes have both low node degrees and betweenness centrality. This means that most nodes in the road data are endpoints that are neither connected to many other nodes nor passed through by many edges. These endpoints are likely caused by footpaths to small villages which stem off from the main roads. Areas with a higher centrality in Figure 3.14 b) are shown in yellow and orange, and they correspond with areas of higher road density in Figure 3.13. Interestingly, these areas do not correspond with the largest, most populous villages. This is because the larger villages have multiple points of interest within the village and so the nodes are distributed across the village, as opposed to smaller villages that only have one point of interest. This is fundamentally different from the mobility data where each village is only assigned one node. Overall, Figures 3.14 a) and 3.14 b) show that the nature of the road network and its distinct definition of nodes result in such highly clustered and disconnected nodes that the linear structure of the network is almost entirely obscured. This would make a detailed and informative structural comparison with the well-defined linear mobility network far more difficult

and less intuitive.

Another characteristic to note about the road data is that most of the roads are naturally non-linear. The curves of the roads are best captured with many short road segments rather than a few large segments. This results in an exponential increase in nodes. The large number of unconnected nodes complicates community detection and makes it very computationally expensive. A one-to-one comparison between mobility and road data is therefore not always feasible. This limitation is exacerbated by other complications of digitising real-world roads such as certain informal roads appearing faded or being altered by seasonal factors like the river water level.

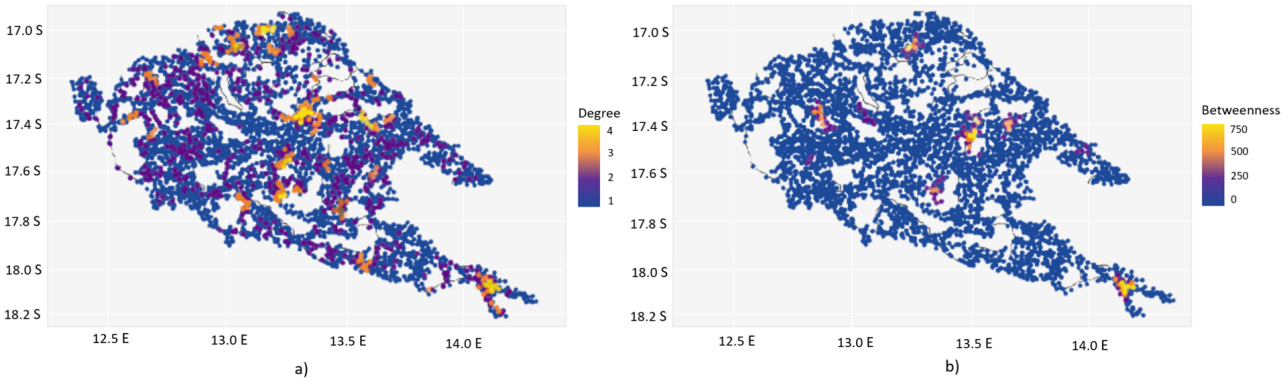


Figure 3.14: Analysis of informal road network including a) node degree distribution, and b) node betweenness centrality for mobility data.

3.3 Comparison of network analysis for mobility and road data

The limitation of comparing two linear networks with fundamentally different node definitions and distributions has been pointed out above. There are, however, a few indices that can be used to analyse the individual networks and compare the two networks based on how similar the various indices are. Certain measures of spread like the π-index, which indicates network shape and is a measure of distance per unit of diameter, as well as network density suggest that the two networks are relatively similar. This is simply due to the fact that the area of informal roads to be considered was specifically stipulated to cover the same area as the mobility data. Beyond that, however, it very quickly becomes apparent that two completely different types of linear networks are being compared.

	Connectivity			**Spread**		
	α	β	γ	π	η	**Network density**
Mobility data	0.099	4.080	0.127	52.096	47.433	0.947
Road data	0.000	0.663	$<$0.001	56.643	1.276	1.054

Table 3.2: Connectivity and spread indices for mobility data.

Firstly, there is a large disparity in the number of nodes and edges between the two networks with the road data having 227.06 times more nodes and 38.04 times more edges than the mobility data. The difference in the η-index values, shown in Table 3.2, also indicate that the average length of edge paths in the mobility data is much greater. Road data is therefore confirmed to be comprised of many more short, single-point connected

edges. The vastly different distributions of node degree and betweenness centrality confirm this. In fact, up to 68.9% of nodes in the road data have a node degree of 1 and a betweenness centrality of 0 while it is only 15.4% for the mobility data. Other connectivity indices yield similar results. The mobility data β-index is above 1 which indicates a complex network with each node having multiple edges. This implies increased connectivity, access and adaptability. The road data β-index, on the other hand, is below 1 indicating a simple network. Many nodes therefore only have one path connecting them to the rest of the graph. This limits accessibility and the connectivity of these roads are decreased. Additionally, both the α-index and γ-index values are higher for mobility data. These indices range from 0 to 1 with 1 representing a fully connected network. The mobility data is therefore again shown to be much more connected.

These indices, however, are limited in their use for sufficient spatial similarity comparison. The indices take an average across the whole network. There is therefore a loss of information concerning regional variations. The indices also tend to only consider one measure at a time, such as connectivity or spread. In reality, these measures tend to be dependent, but the indices fail to take any of these underlying interaction effects into consideration. It is due to all these limitations that a new spatial similarity test is required. Specifically, a successful new spatial similarity test would be robust towards node differences between linear networks, allow for regional variations and take into consideration multiple network characteristics at once.

3.3.1 Conclusion

In this chapter, an exploratory analysis of the social mobility and informal road network was performed. The data collection and linear network construction process was explained, including all assumptions made. Node degree and node betweenness centrality as well as connectivity and spread indices were calculated to evaluate the spatial characteristics of each network. The mobility network indicated moderate complexity with a few central, highly connected populous villages surrounding by many single-visit village in rural areas. The informal road network was much more complex due to the many little branching footpaths creating an exponentially increasing number of intersections. Comparing the two sets of network characteristics clearly demonstrated the limitations of network analysis for the purpose of spatial similarity tests. This was primarily due to the loss of information, the lack of spatial context and the inability to accommodate different types of networks simultaneously. Additionally, a geographically weighted Poisson regression model was fitted to the mobility data set. It confirmed that spatial events like visitation counts are dependent on geographically varying coefficients for variables like village size, road density and distance from the regional capital. Such a model, however, remains exploratory as it does not include any spatial characteristics of the linear network edges which are of primary interest.

Next, Chapter 4 will focus on outlining the theory of the novel linear network spatial similarity test which is based on a recently developed generic spatial similarity test. The test specifications will aim to address many of the issues encountered when comparing linear networks using traditional network analysis.

Chapter 4

Methodology

Spatial linear networks have many different characteristics used to describe and classify them. Certain characteristics may only consider geometric and structural properties like connectivity and spread, while other characterise focus instead on location and spatial proximity. In order to best understand whether two linear networks are connected, we need to compare many different characteristics simultaneously. Depending on the use case, it could potentially be misleading to conclude that two linear networks with a very similar structure are connected if they are a large distance apart. Similarly, two linear networks that occur very close together, but have completely different degrees of connectivity, centrality or spread are also not necessarily connected. A robust spatial similarity test is therefore needed to equally and simultaneously consider both the structural and spatial characteristics of the linear networks during a comparison.

A spatial similarity model is defined as any model which tests whether the spatial structure of two data sets are derived from the same stochastic process [27]. Previous similarity tests for linear networks, however, are not suited for the purpose of a spatial linear network. The tests are commonly limited by computational costs driven by large distance matrices. Additionally, the loss of information when using simplified network indices limits insights, while the inability to accommodate different types of networks simultaneously restricts analysis application. They also do not take the spatial context of the linear network into account, nor do they allow for spatial similarity to vary across different regions. The foremost objective of this mini-dissertation is therefore to develop and demonstrate the use of a novel linear network spatial similarity test which measures the spatial similarity between two linear networks. The primary aim of the test is to estimate the percentage of similarity between two linear networks, denoted as L_1 and L_2. The secondary aim of the test is to provide the local similarity map to visualise both the distribution and degree of similarity regionally.

The newly proposed spatial similarity test is developed based on the foundational theory of a recently introduced generic spatial similarity test [94]. In this mini-dissertation, the generic test is extended and optimised to now also accommodate linear networks. Additionally, the new test also allows for the optimisation of test parameters to better fit the respective linear networks. This work therefore not only contributes to existing statistical literature but also works towards creating an easily implemented and customised test for novel application cases in various fields of research.

The test consists of four steps. First, the two linear networks are each represented as equidistant point patterns. This is to standardise the data representation and link it back to the original spatial similarity test [94]. The second step is to convert the two point patterns to pixel images. A local similarity map based on the pixel images of L_1 and L_2 is generated in the third step, and in the fourth step, a global similarity index based on the local similarity map is calculated. Both local and global similarity are therefore derived to provide the most detailed analysis of spatial similarity.

4.1 Step 1: Convert linear network to point pattern

The generic spatial similarity test converts all spatial data sets to a standard pixel image representation [94]. Therefore, before the linear network spatial similarity test can be applied, the linear networks need to be represented as a data type that can be pixelated. The test has already shown reliable results for unmarked point patterns [94]. It follows that if the respective linear networks can accurately and reliably be represented as point patterns, the similarity test can successfully be extended to linear networks. This is the primary novelty of this new linear network spatial similarity test. Additionally, the linear networks in this case do not contain any additional class labels or volume values, so an unmarked point pattern would be appropriate.

In the process of converting a linear network to a point pattern, the objective is to lose as little structural information as possible. The most data efficient way to do this is to keep the entire vertices set V while foregoing the much larger edge set E. Based on Equations 2.1 and 2.3 formally defining a linear network and an edge as a line segment, the synthetic point pattern X_j is generated for every edge e_j for $j = 1, ..., n$. As such, X_j is defined as a finite set $x_j = \{x_{j1}, \ldots, x_{jn}\}$ of distinct points, where n represents the point density. The point density n determines the number of equidistant, discrete points interposed between two endpoints in order to estimate the continuous line. The i^{th} point on the edge e_j, $x_{j,i} \in e_j$ can be formally expressed as

$$x_i = [u_x + (i-1)(v_x - u_x); u_y + (i-1)(v_y - u_y)] \tag{4.1}$$

where u_x, v_x, u_y and v_y are the planar x and y coordinates for the u and v endpoints, respectively.

The resulting point pattern will therefore consist of equidistant points with a distance of $\frac{\|e\|}{n-1}$ between each point. The points are equidistant along each individual edge, but not throughout the whole network. Figure 4.1 shows an example of this resulting point pattern.

Given that the point pattern is an estimated representation of the line segment, as the point density n increases, it approaches the true line segment. However, the aim of the methodology is to test how spatially similar the general structural patterns of two linear networks are in terms of measures like network length, density and orientation. If the test is applied too strictly to the actual line segments, as opposed to approximations, the test may be skewed by simplifying assumptions made about the linear networks. Such assumptions include representations of varying road widths, vertices snapping during the road digitisation, estimated village coordinates based on missing or inaccurate data collection, and shortest linear paths for mobility routes. Given that the linear networks are, to an extent, approximations themselves, there is no need

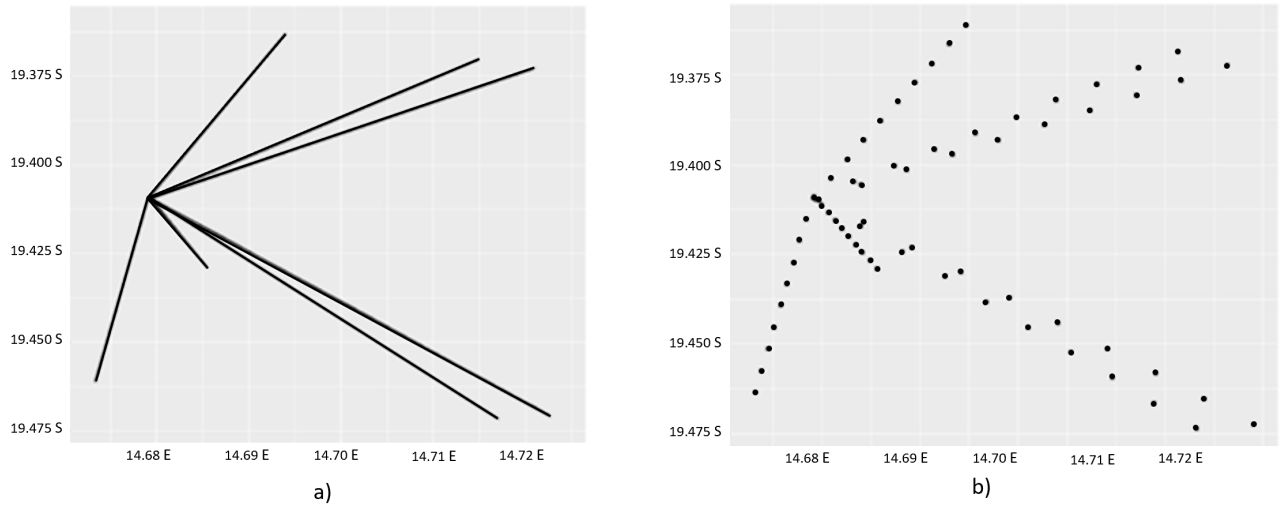


Figure 4.1: Example of converting a) a linear network to b) a point pattern with $n = 10$.

to strictly use the line segment. Furthermore, increasing n also increases the computational cost of the spatial similarity test and therefore starts to impede its efficiency and scalability. The optimal value of n therefore needs to be chosen by the user to avoid biased spatial similarity estimates by either adhering to closely to the line segment or using too sparse data. It also needs to remain computationally feasible. A method of optimising these optimal values is demonstrated during the simulation studies.

For the comparison of the two linear networks L_1 and L_2, we construct two regular, equidistant point patterns P_1 and P_2. The value of n is held constant between the two point patterns to avoid any measuring errors.

4.2 Step 2: Convert point patterns to pixel representation

The two point patterns P_1 and P_2 are now represented as pixel images I_1 and I_2. To achieve this, the two point patterns must first be represented in the same spatial domain, or window. The window is defined as the smallest bounding box containing both point patterns. This ensures that no edge is truncated or excluded. For simplicity's sake, the window is defined to be rectangular, but this methodology would similarly work for any convex shape. The window is represented as an $m \times m$ grid, where the value m represents the chosen resolution. Choosing a higher m will divide the spatial domain of the data set into smaller subregions. This has the benefit of higher granularity, as shown in Figure 4.2, which will allow for more local similarity and dissimilarity to be detected, but it also has the disadvantage of increased computational cost. Each of the $M = m^2$ pixels are then allocated centroids $p = \{p_1, p_2, \ldots, p_M\}$.

When converting a point pattern to a pixel image, the value of each pixel represents the intensity of the point pattern's underlying spatial point process. The intensity of the point process counts the number of points within a certain bounded area and is therefore a density measure [194]. The intensity function at any of the centroids, generalised as p, is expressed as [40]

$$\lambda(p) = \lim_{|d_p| \to 0} \frac{E[N(d_p)]}{|d_p|}, \tag{4.2}$$

where d_p is the infinitely small area such that $p \in d_p$, $|d_p|$ is the area of d_p, and $N(d_p)$ denotes the number of points contained in d_p.

It is, however, known that point processes are theoretically stochastic processes and so the intensity function cannot truly be known, but can instead be estimated [194]. Kernel density estimation applies kernel smoothing to estimate the intensity function [210]. Given that the estimation is based on the spatial point pattern and not on any parametric model, a non-parametric estimator is used [210]. Such an estimator should be unbiased and applicable to non-homogeneous data [94]. This means that the intensity of the point process is not constant but instead is spatially dependent [41]. Diggle's kernel estimator has been proven to meet these criteria [47]. Diggle's corrected density estimate is calculated for each centroid p_i for $i = 1...M$. The estimate is defined as [92]

$$\tilde{\lambda}^D(\boldsymbol{p}_i) = \sum_{j=1} \frac{1}{e(\boldsymbol{x}_j)} \kappa(\boldsymbol{p}_i - \boldsymbol{x}_j) \tag{4.3}$$

where the kernel $\kappa(.)$ is specified as a bivariate Gaussian density $f(\boldsymbol{d}) = \frac{1}{2\pi|\Sigma|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}\boldsymbol{d}\Sigma^{-1}\boldsymbol{d}'\right\}$ with the smoothing parameter $\sum$ = (Diggle's bandwidth) $\times$ I_2. Here $\boldsymbol{d} = \{d_1, ..., d_M\}$ represent the differences between the spatial point pattern observation $\boldsymbol{x}_j$ and the centroids p_i for $i = 1...M$.

Additionally, Diggle's edge correction factor is defined as [92]

$$e(x_j) = \frac{1}{\int_D \kappa(x_j - q_k)\, dq_k}. \tag{4.4}$$

The edge correction factor is important to optimise performance of the estimator around the window boundaries

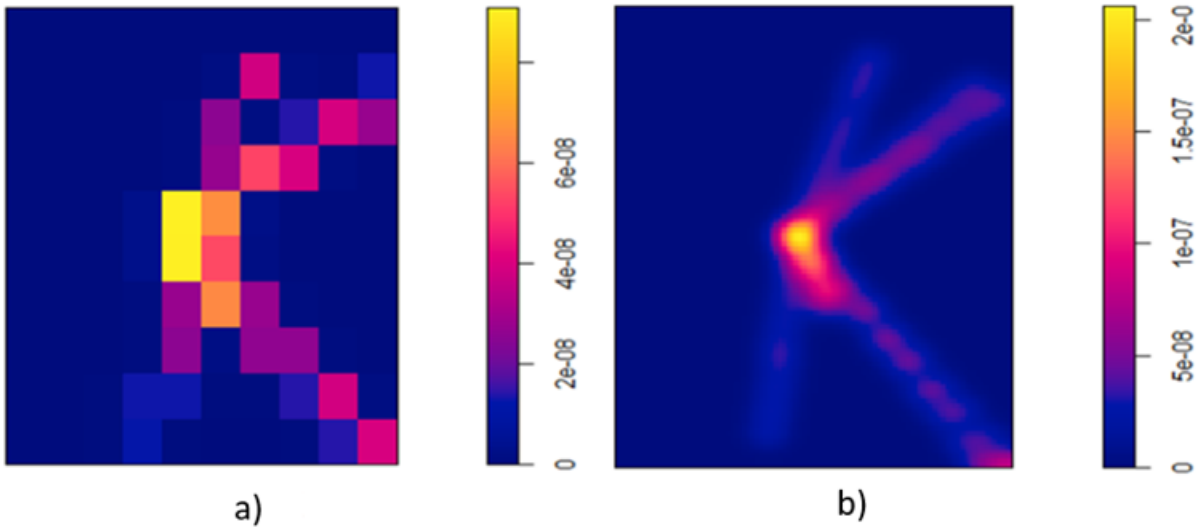


Figure 4.2: Pixel representation of point patterns using a) m=10 and b) m=100.

by addressing edge effects [59]. The correction factor accounts for the bias that the loss of points near the boundaries perpetuate [63]. This is done by separately subdividing the pixel images I_1 and I_2 into a secondary grid $r \times r$ with a higher resolution such that $r > m$. The centres of each $R = r^2$ grid cells are again defined as centroids q_k, $k = 1, ..., R$. For each observed point $x_j, j = 1, ..., n$ in the point pattern, the difference between that respective point x_j and each centroid q_k is calculated as $\boldsymbol{d}^* = \{d_1, ..., d_R\}$. The bivariate Gaussian density

estimate $f(\boldsymbol{d_k})$ is used again. The final edge correction factor can then be calculated as

$$e(\boldsymbol{x}_j) = \frac{\text{area}(D)}{R} \sum_{k=1}^{R} f(\boldsymbol{d_k}). \tag{4.5}$$

4.3 Step 3: Generate a local similarity map

A local similarity map S is derived from the pixel images I_1 and I_2 using the structural similarity index (SSIM) [207]. SSIM is a metric used to measure the similarity between two images by relying on concepts of the human visual system in order to discern any underlying deviation in the structural information of the two images. This method therefore prioritises the overall pixel structures. Other methods, like mean square error, on the other hand, simply aggregate the difference in pixel intensities and isolates the image colour over its content. While this is useful for other tasks like image restoration and quality assessment, a robust comparison of spatial similarity must focus on the underlying structure [178, 218].

The similarity index is calculated for each pixel using a sliding window with dimensions $w \times w$ centred at the specific pixel. We consider the two pixel images I_1 and I_2 where each image is an $m \times m$ matrix with $M = m^2$ individual pixels. The i^{th} pixel will therefore be assigned two vectors, namely $y_{1,i}$ and $y_{2,i}$ which represent the number of valid pixels within the sliding window centred at $I_{1,i}$ and $I_{2,i}$, respectively. The smaller the sliding window is, the more 'local' the derived similarity value for each pixel will be as it only considers information from the pixel's closest neighbours. This is preferable, as image statistical features have been shown to be significantly non-stationary [208]. The focus on only one local area at a time further mimics the natural human visual system that SSIM is based on [208]. Furthermore, localised similarity comparisons provided a higher resolution similarity map which can be used for more detailed analysis of the structural differences in the linear networks. A larger sliding window, on the other hand, would be more global and likely miss any subtle local variations in the network. Therefore, a small window is recommended. The actual number of pixels considered will also depend on where the pixel is located. Central pixels will consider $w^2 - 1$ pixels, border pixels will have $w^2 - w - 1$ neighbouring pixels and corner pixels will only take into account $w^2 - 2w$ pixels.

For each pair of vectors $y_{1,i}$ and $y_{2,i}$, we calculate comparison functions for three human visual system concepts using the values which fall within the sliding window [207]. The first component is luminance which measures the luminous intensity of light, or brightness, within the sliding area. As shown in Figure 4.3, the luminance measurement is calculated first so that the effect of luminance on both images can be removed given that it has no bearing on the underlying structure of the images. It is expressed as

$$\text{Luminance}: \quad \ell(y_{1,i}, y_{2,i}) = \frac{2\mu_{y_{1,i}}\mu_{y_{2,i}} + C_1}{\mu^2_{y_{1,i}}\mu^2_{y_{2,i}} + C_1}. \tag{4.6}$$

The second component is contrast, which measures the variation between the different greyscale values of all pixels within the sliding window. Similar to luminance, the effect of contrast is also separated from the final

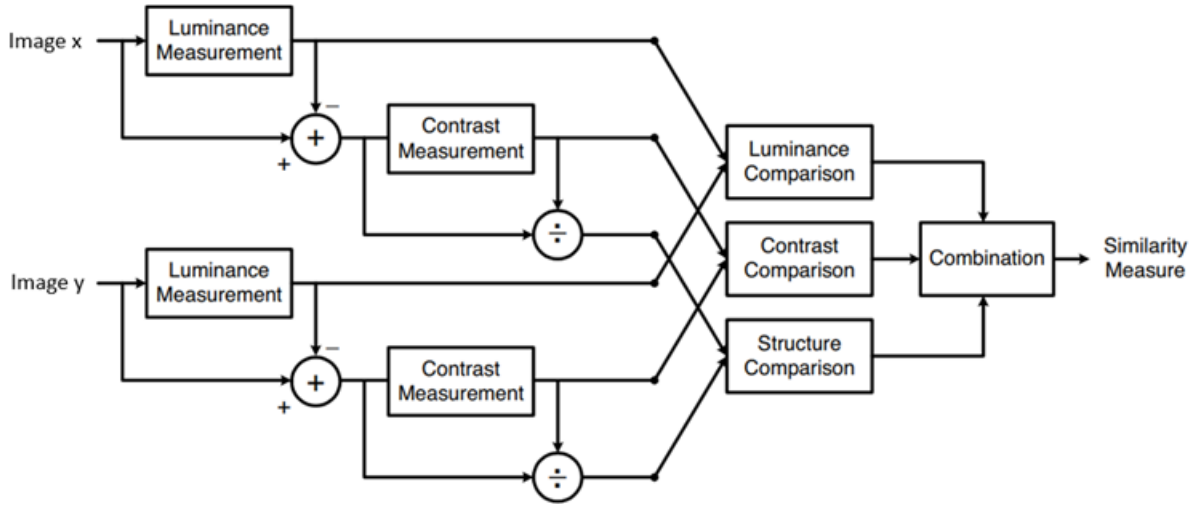


Figure 4.3: A visualisation of the structural similarity (SSIM) measurement process as included in the original paper [208].

structural comparison. The expression for contrast is

$$\text{Contrast}: \quad c\left(y_{1,i}, y_{2,i}\right) = \frac{2\sigma_{y_{1,i}}\sigma_{y_{2,i}} + C_2}{\sigma^2_{y_{1,i}}\sigma^2_{y_{2,i}} + C_2}. \tag{4.7}$$

The third component is structure, which considers the interdependencies between pixels that are spatially close together within the sliding window. It is expressed as

$$\text{Structure}: \quad s\left(y_{1,i}, y_{2,i}\right) = \frac{2\sigma_{y_{1,i},y_{2,i}} + C_3}{\sigma_{y_{1,i}}\sigma_{y_{2,i}} + C_3} \tag{4.8}$$

where

$$\mu_{y_{1,i}} = \frac{1}{N}\sum_{j=1}^{N} y_{1,i,j} \quad \mu_{y_{2,i}} = \frac{1}{N}\sum_{j=1}^{N} y_{2,i,j},$$

$$\sigma^2_{y_{1,i}} = \frac{1}{N-1}\sum_{j=1}^{N}\left(y_{1,i,j} - \mu_{y_{1,i}}\right)^2 \quad \sigma^2_{y_{2,i}} = \frac{1}{N-1}\sum_{j=1}^{N}\left(y_{2,i,j} - \mu_{y_{2,i}}\right)^2$$

and

$$\sigma_{y_{1,i},y_{2,i}} = \frac{1}{N-1}\sum_{j=1}^{N}\left(y_{1,i,j} - \mu_{y_{1,i}}\right)\left(y_{2,i,j} - \mu_{y_{2,i}}\right)$$

with N as the number of valid pixels in the sliding window, L as the range of pixel values, and the stabilising constants set as $C_1 = (0.01L)$, $C_2 = (0.03L)$ and $C_3 = \frac{C_2}{2}$. These specific values for the constants are recommended in the original paper with the purpose of minimising biased similarity indices in the special cases of null pixels where both the sum of the squared means and sum of the variances approximate 0 [208].

The final similarity index for the i^{th} pixel is calculated as

$$\text{SSIM}\left(y_{1,i}, y_{2,i}\right) = \left[\ell\left(y_{1,i}, y_{2,i}\right)\right]^{\alpha}\left[c\left(y_{1,i}, y_{2,i}\right)\right]^{\beta}\left[s\left(y_{1,i}, y_{2,i}\right)\right]^{\gamma}$$

where $\alpha = \beta = \gamma = 1$ to allow for equal weighting of the three visual system concepts [208].

To summarise, the common centre pixels, $I_{1,i}$ and $I_{2,i}$, from the two images being compared, I_1 and I_2,

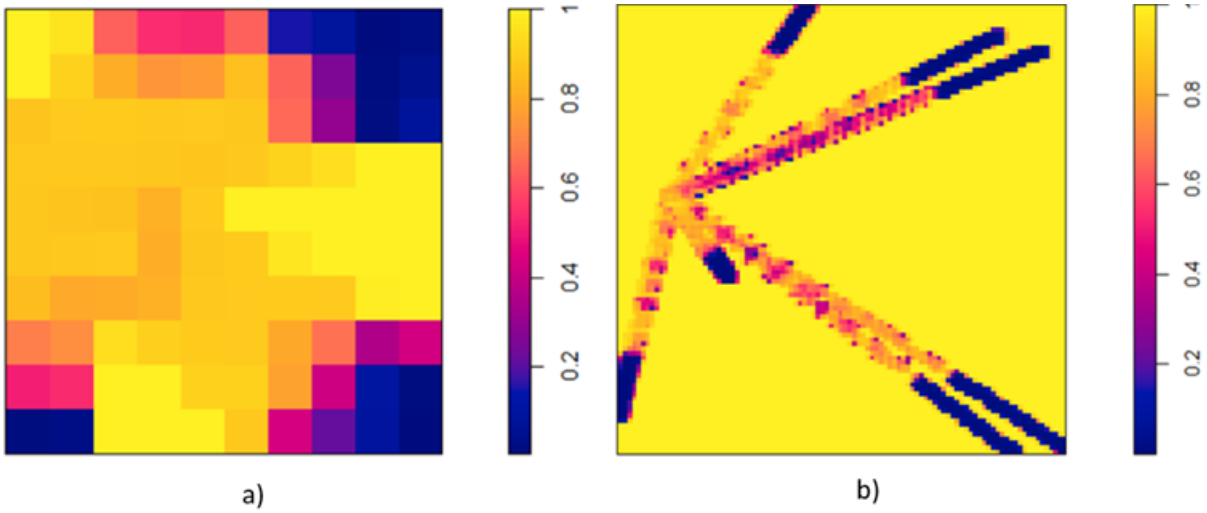


Figure 4.4: Similarity maps comparing the reference linear network in Figure 4.1 with another 70% similar linear network at a) m = 10 and b) m = 100.

are assigned sliding windows of the same dimensions. The mean, variance and covariance of the pixel values within the sliding windows are calculated separately for each image. Six values should therefore be calculated for each i^{th} iteration. These values are appropriately substituted into the comparison functions for luminance, contrast and structure. Finally, the three values for luminance, contrast and structure are substituted into the similarity index formula $\text{SSIM}\,(y_{1,i}, y_{2,i})$. The similarity index is calculated for each of the M pixels to construct the final $m \times m$ local similarity map S.

As shown in Figure 4.4, the similarity values range from 0 indicating absolute dissimilarity to 1 which corresponds with perfect similarity. Additionally, the choice of the resolution grid size m clearly has an effect on the local similarity map. While both maps do capture the general distribution of variations near the borders, the similarity map at $m = 100$ shows in much more detail that the variations are specifically due to a scaling transformation of each of the edges. The effect of grid choice is further demonstrated in Chapter 5 in the simulation study.

4.4 Step 4: Calculate the global similarity index

Once the similarity map has been generated, a global similarity index is calculated. The pixel values are continuous, which makes the adapted Andresen's S-index with a non-binary input used in the generic spatial similarity test an appropriate measure [6, 94].

In Andresen's spatial point pattern test, one point pattern is used as the baseline point pattern. The spatial domain of the baseline point pattern is subdivided into a grid of areal units and each point is assigned to a specific areal unit. An aggregate of the total number of points within each areal unit is derived. The proportion of points within each areal unit is calculated. The comparison point pattern is then subdivided into the same areal unit grid. A Monte Carlo simulation is used to derive a confidence interval of the proportion of points in each areal unit. During each simulation iteration, the proportion is estimated by selecting 85% of the total comparison point pattern using random sampling with replacement. For each individual areal unit, the

baseline point pattern proportions and the comparison point pattern confidence intervals are compared. The two point patterns are considered similar if the baseline point pattern proportion falls within the corresponding comparison point pattern confidence interval.

Initially, Andresen's S-index was only binary where similar areal units are assigned a 1 and non-similar units have a value of 0. The S-index has, however, been extended to allow for non-binary input in recent work by using the continuous pixel values of the local similarity map [94]. The similarity index is therefore calculated as the mean similarity across all the areal units and is expressed as [6]

$$GS = \frac{\sum_{i=1}^{M} \text{SSIM}\,(p_i)}{M}$$

where M is the total number of pixels in the local similarity map S and $SSIM(p_i)$ is the similarity index for the i^{th} pixel with centroid p_i.

The final index indicates the mean similarity value between the two linear networks. For the example used in Figure 4.4 the two global similarity indices for the respective grid sizes were 0.657 and 0.691. These values generally align with the known similarity of 70% and demonstrate the validity of the global index.

This test has the benefit of being independent of clustering, uniformity and randomness, which makes it apt for comparing linear networks with different levels of complexity and overall distribution [6]. Additionally, the example also shows that using the mean of local similarity values is more accurate than simply considering the proportion of similar pixels. For example, 4.4 a) shows that only 19 of the 100 pixels were exactly similar with a local similarity value of 1, denoted in yellow. If the similar pixels were only considered, spatial similarity would be severely underestimated. The majority of the common pixels are also the empty spaces between network edges that the two linear networks have in common and not the pixels representing the linear network edges themselves. This method would therefore not only be inaccurate, but it would also miss the fundamental objective of the test, which is to test the structural similarity of the linear network edges.

4.4.1 Conclusion

In this chapter, the methodology for the new linear network spatial similarity test was outlined. The spatial similarity test consists of four steps and depends on three user-defined parameters. In the first step, each of the two linear networks is converted to a point pattern. The regular, equidistant point pattern depends on the user-defined point density. The second step is to convert the two respective point patterns to pixel image representations using kernel estimation. The resolution of the pixel image is determined by the specified grid size. In the third step, a local similarity map is generated based on the two pixel images. The granularity of the spatial similarity depends on the user-defined sliding window size. The fourth step is to calculate the global similarity index based on the local similarity map values. The global index is derived from Andresen's S-index for spatial point patterns.

The next steps in Chapter 5 are to finally demonstrate the new linear network spatial similarity test in simulation studies. The first simulation study will assess the overall performance of the test. Additionally, the simulation study will demonstrate how to optimise each of the three user-defined parameters used in the test.

The second simulation study will verify the improved accuracy of this outline methodology compared to an alternative method.

Chapter 5

Simulation

In this chapter, two simulation studies are performed. The first simulation study assesses the overall performance of the test. A method of optimising the user-defined parameters is also demonstrated. The second simulation study compares the performance of the newly proposed linear network spatial similarity test with an alternative method to specifically demonstrate the advantage of including point patterns in the test's methodology. For both simulation studies, the test's performance is measured with bias, precision and accuracy metrics. The results are discussed and any relevant insights into the test's implementation are highlighted.

5.1 Introduction

A simulation study is carried out to demonstrate the proposed spatial similarity test's novel application on linear networks. The simulation study demonstrates the accuracy and reliability of the test in lieu of any theoretical mathematics to prove the validity of the test. This is done by applying the test to synthetic linear networks with known spatial variations. The reference linear network, seen in Figure 4.1, is a subset of the previously mentioned mobility data set with seven destination points. The destinations vary in distance, density and orientation and should therefore accurately simulate a real-world data set. The variations of the synthetic simulations are known by stipulating beforehand the percentage of the total edge length that needs to overlap between the reference and synthetic networks. This simulation study specifically considers different levels of spatial similarity at 70%, 80% and 90%. The motivation for choosing to only simulate highly similar networks is to highlight the precision of the test. In highly similar networks, there are very nuanced differences in either the number of orientation of nodes. We want to test whether we can determine both the degree and the location of small differences accurately. It is also assumed that the use of the novel spatial similarity test will primarily be to test two networks that are assumed to be similar but now needed a precise quantification of that similarity. Networks with very low similarity will very clearly not be structurally comparable and will therefore not require further testing. It should be noted that while networks with low similarity fall outside the scope of this simulation study, the spatial similarity test can still be applied to any network regardless of assumed similarity level.

To ensure the test is robust against all types of spatial dissimilarity, different types of variation are analysed

for each of these levels of spatial similarity. The first type of variation, shown in Figure 5.1 a), is re-scaling the network edges. The number and orientation of the edges are kept constant while the lengths are varied so that the total edge length of one network is 70%, 80% or 90% of the total edge length of the other network. This is particularly helpful with testing whether the similarity test is biased when one of the data sets is a subset of the other. The second type of variation, demonstrated in Figure 5.1 b), applies geometric transformations like rotations, reflections and translations to the edges. It also includes adding or removing entire edges. This tests how the spatial similarity test responds to linear networks with underlying differences in edge frequency, orientation and length. The only stipulation is that 70%, 80% or 90% of the reference network's total edge length must overlap with that of the simulated networks. Lastly, a combination of edge scaling and transformation, seen in Figure 5.1 c) is simulated to test how the test performs overall. At each variation level and for each variation type, 100 different iterations are simulated to account for as many types and combinations of edge variations. All simulations are generated using the `R` package `spatstat.linnet` [14]. It should be noted that Figure 5.1 is merely an illustration of one of the iterations of the simulation. The same starting point was used in this case to simply visualise what it means for 70%, 80% and 90% of the total network length to differ according to the different variation methods. Other simulation iterations do include networks with different nodes and starting points as well as varying node frequency and placement.

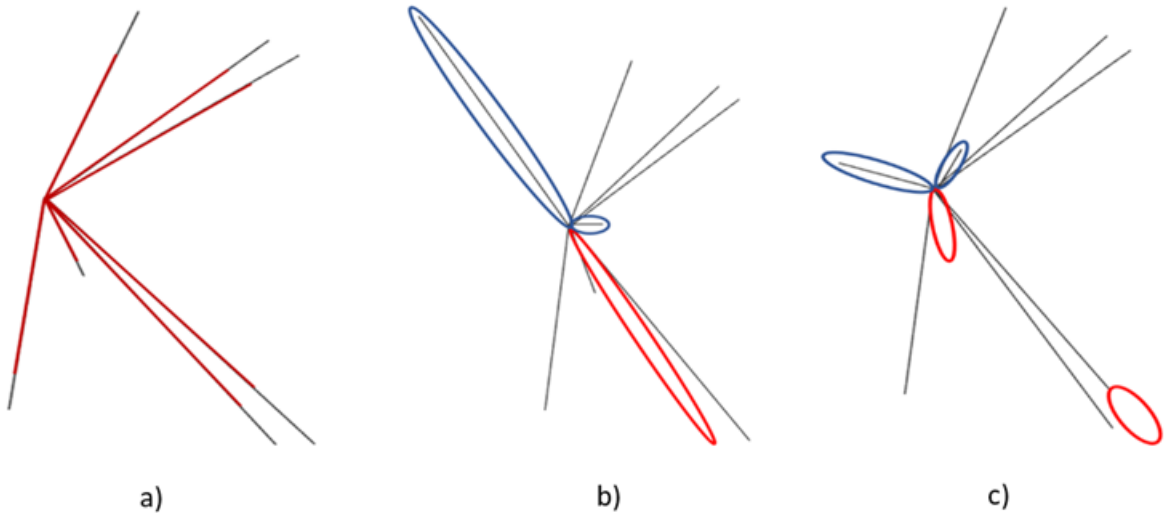


Figure 5.1: Examples of different linear network variations with a) edges rescaled to 70% of the original edge lengths, b) edges added (shown in blue) or removed (shown in red) to keep only 80% of the original edge lengths, and c) a combination of edge transformations to keep 90% of the original edge lengths.

Given that the synthetic linear networks are transformations of the reference linear network and the points are not generated from a Gaussian distribution, it is known that the data is not Gaussian and that the assumption of normality does not hold. Additionally, the distribution of both the x and y coordinates of the reference linear network were tested using the multivariate normality Henze-Zirkler test [73]. The test yielded a p-value of 1.262×10^{-9} which rejects the null hypothesis of multivariate normality. In this case, the difference of means can be tested using the Friedman test [60]. This is a non-parametric test to determine if the means differ significantly across multiple groups in which the same simulation iterations are included in each group. The difference of means is tested for all three parameters, namely the sliding window size, the resolution grid

size and the point density. If the means for a parameter are found to differ, then it means that the parameter has a statistically significant effect on spatial similarity within that particular variation type and level. These parameters can then be optimised to maximise the accuracy of the spatial similarity test.

5.2 Performance metrics

The true level of spatial similarity is, by design, known in the simulation studies and so can be used to derive statistical performance measures for the test's assessment. Particularly, we consider measures of bias, precision and accuracy.

Firstly, bias quantifies to what extent the test's estimate deviates from the true value [97]. In this case, the final global similarity index can be interpreted as an unbiased estimator of spatial similarity if, on average, it yields zero bias. The test can thus still be considered unbiased if it marginally under- or overestimated similarity in a few cases, as long as the overall bias is negligibly different from 0. The overall bias is measured using the average relative bias of the sample mean. The average bias of the j^{th} simulation scenario calculates the mean bias across for one of the nine combinations of the different variation level and type combinations. It is formally expressed as

$$\text{Average bias}_j = \frac{\sum_{i=1}^{N} \left(\frac{E\left[\hat{\theta}_{ji}\right] - \theta_j}{\theta_j} \right)}{N} \quad j = 1, ..., 9 \tag{5.1}$$

where θ_j represents the true spatial similarity for the j^{th} simulation scenario, $\hat{\theta}_{ji}$ denotes the estimated spatial similarity for the j^{th} simulation scenario and i^{th} simulation iteration, and N is the number of simulations run for each scenario.

Secondly, precision measures the variation of an estimator and whether there are any random errors in estimation [97]. When parameters such as the resolution and sample size of the data are held consistent, precision measures to what extent estimates deviate from one another. The primary precision measure used is the standard deviation. It measures the average distance between any one simulation's similarity estimate and the mean similarity for all simulations. A low standard deviation indicates a consistent and precise estimator. Additionally, if there is a high standard deviation, further analysis can be done to see which specific simulation conditions are deviating and therefore yielding particularly poor results. The j^{th} standard deviation is written as

$$\sigma_j = \sqrt{\frac{1}{N} \sum_{i=1}^{N} \left(\hat{\theta}_{ji} - \bar{\theta}_j \right)^2} \tag{5.2}$$

where $\bar{\theta}_j$ is the sample mean of the j^{th} estimator.

Another simple precision measure is the range, which is calculated as the difference between the maximum and minimum estimates within the simulation. A small range would indicate overall estimator consistency under reasonable conditions. On the other hand, a large range would indicate volatile estimates that would not be reliable enough for further analysis.

Lastly, accuracy is defined as the overall distance of an estimate from the true value across multiple

simulations [97]. The first accuracy measure considered here is the root mean square error (RMSE) and it is defined as the square of the mean distance between the estimate and true value. The square is taken since, like the standard deviation, it returns the measure to the original scale and makes interpretability easier. The RMSE incorporates both bias and precision, and it can be calculated as the addition of the variance of the estimator and the square of the bias. A low RMSE depends on minimising both bias and variance. Therefore, a high accuracy is accompanied by low bias and high precision. It can formally be expressed as

$$\text{RMSE}_{\ j} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\hat{\theta}_{ji} - \theta_{ji}\right)^2}. \tag{5.3}$$

It should be noted that since the RMSE is calculated using squared differences, it is highly affected by extreme outlying estimates. To address this issue, mean absolute error (MAE) is used as another accuracy metric. It is calculated as the mean absolute difference between the estimates and true values and can be expressed as

$$\text{MAE}_{\ j} = \frac{1}{N}\sum_{i=1}^{N}\left|\hat{\theta}_{ji} - \theta_{ji}\right|. \tag{5.4}$$

5.3 Optimisation of parameters

The similarity tests depend on three parameters, namely the sliding window dimension w, the resolution grid size m and the point density n. In the case where varying these parameters has a clear and statistically significant effect on the spatial similarity test's accuracy, the parameters need to be optimised prior to application. Optimising parameters means to select parameters that minimise the objective function. The sum of squares is a common example of an objective function and is written as

$$\arg\min \sum_{r\in R}\left(\hat{\hat{\theta}}_j - \theta_j\right)^2 \tag{5.5}$$

where R is the set of all parameter combinations and r represents a single parameter combination.

A grid search is one of the most widely applied methods of hyperparameter optimisation [86]. Cross-validation is used to evaluate the results of all the simulations, and the optimised parameter combination is determined by the combination with the highest performance metric. The grid search is often exhaustive which means all possible parameter combinations are considered and tested. Alternatively, the search could also be conducted using successive halving. Parameter combinations are compared and, in each iteration only the best half of combinations are retained for further comparison until finally only one optimised combination remains.

There are also other successful hyperparameter optimisation methods such as random searches, genetic algorithms, Bayesian optimisation and gradient-based optimisation [23, 203, 54, 22]. When optimising across large, continuous parameter spaces, these methods often perform better than the grid search. Additionally, these methods allow for further specifications and constraints when parameters must align with practical real-world values as opposed to purely optimised theoretical values [135]. However, given that the parameters of

the linear network spatial similarity test are discrete and do not have any restricting practical constraints, the grid search is a good choice. It is accurate, efficient and computationally simple.

An exhaustive grid search is therefore used to optimise the hyperparameters in the simulation studies. Given that there are three parameters to optimise in the first simulation study, the grid is constructed in $\mathbb{R}^3$. In the second simulation study, when only two parameters are optimised, the grid is constructed in $\mathbb{R}^2$. Cross-validation finds the optimal parameter combination by specifically comparing the MAE of the simulation results.

Once the optimised parameter combination has been selected, the improved performance of the test is demonstrated by calculating the relative decrease in RMSE and MAE between the optimised simulations and simulations without parameter optimisation. These metrics are denoted as RMSE* and MAE*, respectively, and are formulated as

$$\begin{aligned} \text{RMSE}^* &= 100\left(\left(\frac{\text{RMSE}_P}{\text{RMSE}_U}\right)^2 - 1\right) \\ \text{MAE}^* &= 100\left(\left(\frac{\text{MAE}_P}{\text{MAE}_U}\right)^2 - 1\right) \end{aligned} \tag{5.6}$$

where RMSE_P and RMSE_U represent the RMSE of the parameterised and unparameterised simulations, respectively, and MAE_P and MAE_U represent the MAE of the parameterised and unparameterised simulations.

It should be noted that the choice of an optimal parameter combination will vary if other characteristics of the data set are changed. Data resolution, the size and shape of the spatial domain and the shape of the sliding window all influence which parameters perform best. As such, the results of the parameter optimisation process are not universal for all linear networks. In practice, a very small subset of the data can manually be assigned a true level of spatial similarity and subsequently be used for the grid search's cross-validation. This procedure could be considered part of the data pre-processing step, and it does not negate the improved detail, efficiency and scalability of the spatial similarity test's analysis.

5.4 Design of the first simulation study

The purpose of the first simulation is to test how well the spatial similarity test performs. The performance of the test is initially evaluated without any prior knowledge of which user-defined parameters to use in the optimisation of the test. This sensitivity analysis is done to assess both how accurate and how robust the test is.

As such, the simulations are conducted for different sliding window sizes, namely $w = \{3, 5, 7, 9\}$. These smaller sizes are chosen since smaller window sizes are expected to be more granular and accurate. For each value of w, various combinations of point density n and resolution grid size m values are considered. The point density increases from 5 to 100 in increments of 5 while the resolution grid size increases from 10 to 100 in increments of 5. The minimum resolution grid size is chosen to accommodate the different w values. A sliding window with dimensions 9×9 cannot be applied to a pixel image with only a 5×5 resolution grid, for example. On the other hand, the minimum point density is chosen to avoid edge estimations that are objectively too sparse. In total, 1520 different parameter combinations are considered to ensure robustness. Additionally,

the large selection of parameter values is chosen to avoid the risk of selection bias. The ease of sampling lower parameter values in the simulations is evident in the exponential increase of computational time as the resolution grid size and point density increase. Therefore, a larger sample size is explicitly specified. All these parameter combinations are furthermore repeated for each variation level and type combination.

The spatial similarity test is then applied to the simulated linear networks. Initially, the estimated global similarity indices across the 100 iterations for each of the 1520 parameter combinations are all aggregated. This yields a mean estimated global similarity index for each of the nine variation level and type combinations. Appropriate bias, precision and accuracy metrics are used to evaluate the overall performance prior to any parameter optimisation. The parameters are then optimised with a grid search to improve the accuracy of the test. The estimated global similarity indices across the 100 iterations for the optimised parameters are again aggregated. Accuracy metrics are analysed based on the mean performance for each variation level and type, and the effect of parameter optimisation on accuracy is quantified.

5.5 Design of the second simulation study

The novelty of the new linear network similarity test is first representing the linear networks as point patterns in order to address aspects such as resolution robustness and computational efficiency. To prove that the point pattern representation is in fact responsible for the improvement in the spatial similarity test's performance, a similar simulation study is conducted on the unprocessed pixelated images of the linear networks.

The unprocessed pixelated images are based on the graphical representation of the simulated linear networks mapped directly in R. The graphs are represented in the JPEG format and saved as images. These images serve as the direct, unprocessed pixel representations of the linear networks. From this point onward, the spatial similarity test is applied just as before. A local similarity test is generated from the pixel images and a final global similarity index is calculated.

While the above method does not use point patterns and so does not specify a point density n, the method still depends on the resolution grid size m for the pixelation and the sliding window sizes w for the local similarity map. Therefore, to ensure a fair comparison between the newly proposed test and this alternative method, various combinations of m and w are simulated to find an optimised combination. Simulations are again conducted for the different sliding window sizes $w = \{3, 5, 7, 9\}$. For each w, different grid sizes increasing from 10 to 100 in increments of 5 are considered. In total, 76 parameter combinations are simulated. This is repeated for each combination of variation level and type.

The spatial similarity test is applied to the unprocessed pixelated simulated linear networks. The estimated global similarity indices across the 100 iterations for each of the 76 parameter combinations are all aggregated. This again yields a mean estimated global similarity index for each of the nine variation level and type combinations. Appropriate accuracy metrics are used to evaluate the overall performance of the unprocessed pixelation method prior to any parameter optimisation. The optimised sliding window size and resolution grid size are then chosen using a grid search. The estimated global similarity indices across the 100 iterations for the optimised parameters are again aggregated. Accuracy metrics are analysed based on the mean performance

for each variation level and type, and the effect of the point pattern representation on accuracy is quantified.

5.6 Results and discussion of the simulations

5.6.1 Evaluating the performance of the spatial similarity test without parameter optimisation for the first simulation study

Prior to parameter optimisation, the linear network spatial similarity test is applied to the simulated linear networks considering all parameter combinations. From Table 5.1, it can be seen that the test performs well with an average deviation of only 2.3% from the true level of spatial similarity. There is a tendency for the test to underestimate similarity across all variation types and level, as shown by the average relative bias, but the underestimation is generally limited. The test seems to perform marginally better for scaling-type variations. It also yields more accurate similarity estimates at higher levels of similarity. Additionally, the standard deviations appear low enough to conclude that the test is precise and therefore reliable enough.

		Performance metrics					
		Centrality	**Bias**	**Precision**		**Accuracy**	
Variation type	**Similarity level(%)**	**Mean(%)**	**Average relative bias**	**Standard deviation**	**Range**	**RMSE**	**MAE**
Scaling	70	0.696	-0.005	0.047	0.39	0.047	0.032
	80	0.769	-0.039	0.063	0.564	0.07	0.044
	90	**0.901**	0.001	0.062	0.458	0.063	0.038
Transformation	70	0.743	0.104	0.043	0.23	0.084	0.073
	80	**0.792**	-0.010	0.033	0.183	0.034	0.028
	90	0.854	-0.052	0.032	0.277	0.057	0.048
Combination	70	0.734	0.049	0.038	0.284	0.051	0.039
	80	**0.780**	-0.025	0.03	0.287	0.036	0.026
	90	0.877	-0.025	0.035	0.264	0.042	0.025

Table 5.1: Performance metrics for aggregated similarity indices across all parameter combinations.

5.6.2 Optimising the parameters for the first simulation study

For all three parameters separately, the Friedman test yields a p-value less than 0.001. This confirms that the parameters have a statistically significant effect on the accuracy of the spatial similarity test. To visualise this effect, the distribution of the mean similarity deviations between the spatial similarity test's estimates and the true levels of similarity are shown for different combinations of the parameters in Figure 5.2.

For each parameter combination, it can be seen whether that particular combination has a tendency, on average, to underestimate, overestimate or correctly estimate spatial similarity. Additionally, the visualisation provides initial guidance on what effect increasing or decreasing the individual parameters will have on the test's accuracy. Since the effect of the sliding window size, while statistically significant, is less obvious, the effect of parameter optimisation can more clearly be seen when the sliding window is fixed, as shown in Figure 5.3. More examples of the distribution of similarity estimates for varying sliding window sizes and variation types are included in the Appendix.

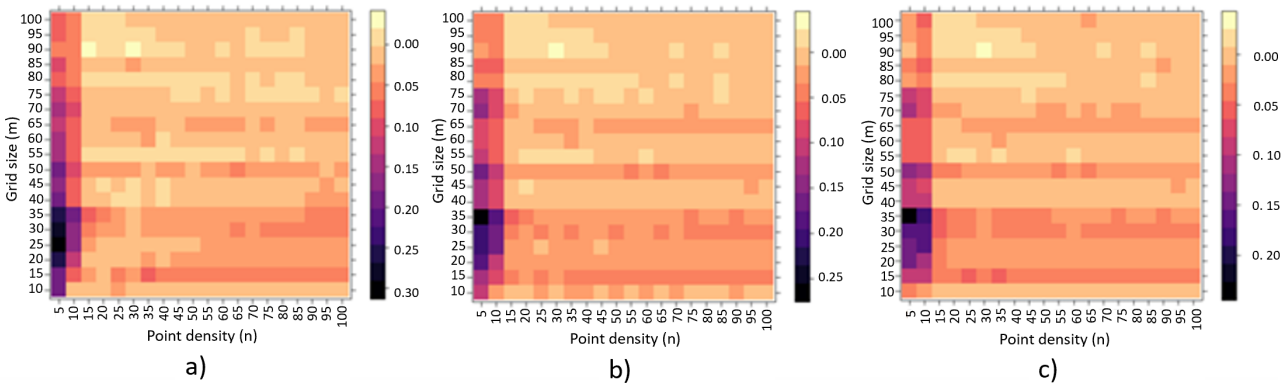


Figure 5.2: The distribution of the mean deviations between global similarity estimates and the true levels of spatial similarity across different point densities n and grid sizes m. Different sliding window sizes were also considered, namely a) $w = 3$, b) $w = 5$, and c) $w = 7$. The simulation scenario used was a combination of edge transformations at a 90% spatial similarity.

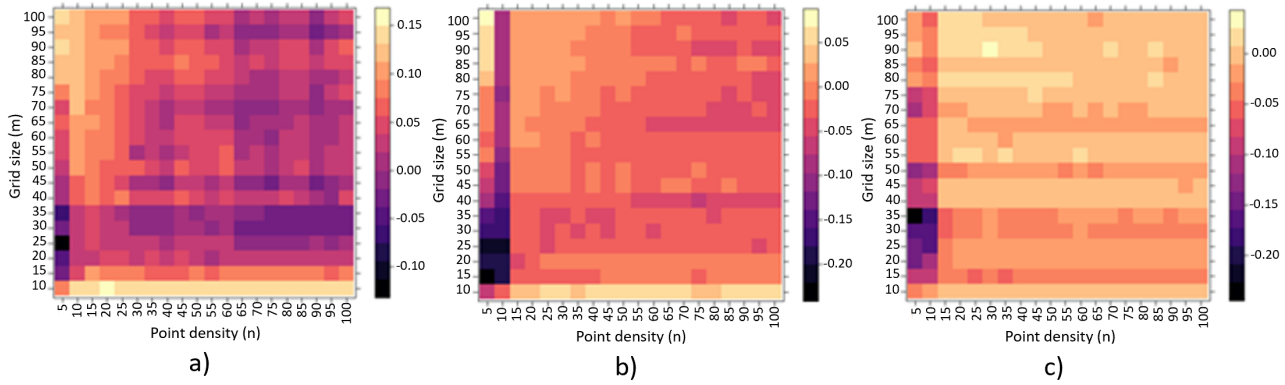


Figure 5.3: The distribution of the mean deviations between global similarity estimates and the true levels of spatial similarity for simulations run using a sliding window of $w = 7$. Permutations of different point density n and grid sizes m are tested for network variations based on a combination of edge transformations. The mean similarity indices are included for a) 70%, b) 80%, and c) 90% similar networks.

Given that the parameters clearly influence the test's accuracy, they need to be optimised. The exhaustive grid search shows that the optimal parameter combination for this data set is $\{w, n, m\} = \{7, 30, 45\}$. The individual simulations scenarios also all have their own specific optimal bands. For example, higher point densities of $35 \leq n \leq 65$ performed better for cases of edge scaling and lower levels of similarity at 70% generally did better with lower grid sizes of $15 \leq n \leq 40$. However, to increase the robustness of the test, the most general parameters are recommended.

Figure 5.4 a) and b) shows a decrease of both RMSE and MAE across all variation types and levels. Specifically, the aggregated relative decrease in these metrics, based on the results in Table 5.2, are 82.8% and 72.9%, respectively. Furthermore, the average deviation between the estimated and true similarity is now only 1.9%. There is therefore a consistent increase in the linear network spatial similarity test accuracy when using the optimal parameters. The improvement of the test is particularly evident at higher levels of similarity at 80% and 90%. The improvements are also more pronounced for the transformation variation type. Optimising parameters is therefore especially beneficial in cases of linear networks with lower levels of similarity primarily driven by the geometric transformation, as is expected in real-world linear networks.

Furthermore, when looking at the bias in Table 5.2, parameter optimisation is seen to reduce rates of under- and over-estimation. With the exception of edge scaling at 70% similarity, the sign of the bias does

		Performance metrics					
		Centrality	Bias	Precision		Accuracy	
Variation type	**Similarity level(%)**	**Mean(%)**	**Average relative bias**	**Standard deviation**	**Range**	**RMSE**	**MAE**
Scaling	70	**0.716**	0.024	0.019	0.026	0.014	0.013
	80	0.779	-0.026	0.010	0.010	0.026	0.025
	90	0.924	0.027	0.010	0.010	0.020	0.019
Transformation	70	0.737	0.052	0.023	0.032	0.055	0.053
	80	**0.797**	-0.004	0.010	0.015	0.008	0.007
	90	0.869	-0.034	0.001	0.031	0.031	0.030
Combination	70	0.729	0.038	0.002	0.003	0.029	0.028
	80	0.788	-0.015	0.002	0.002	0.011	0.010
	90	**0.897**	-0.004	0.004	0.005	0.007	0.006

Table 5.2: Performance metrics for aggregated similarity indices using a sliding window of the optimised parameters $\{w, n, m\} = \{7, 30, 45\}$.

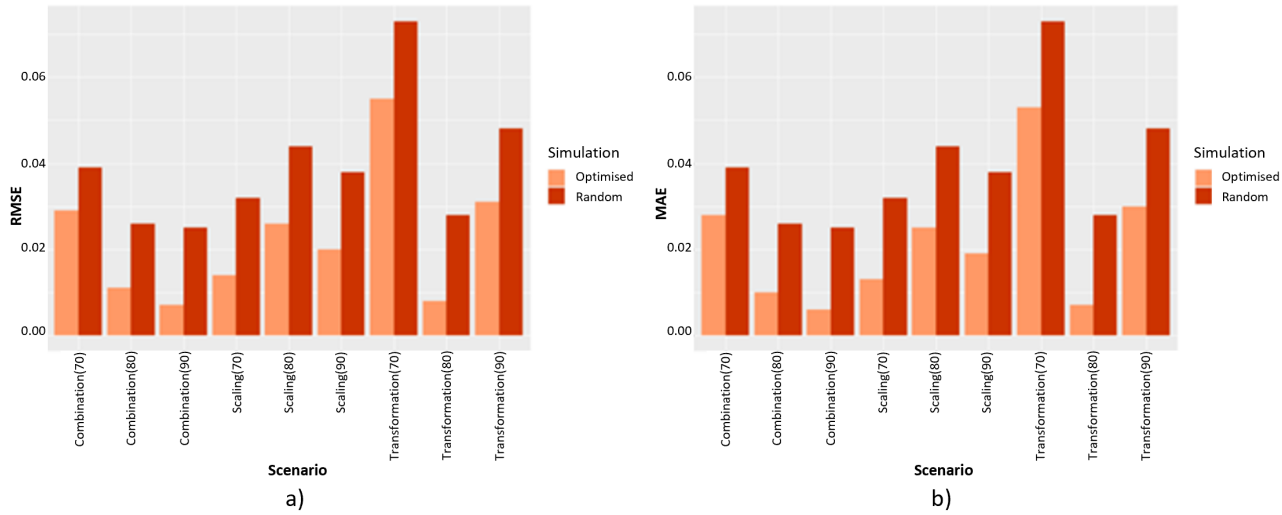


Figure 5.4: Accuracy metrics for optimised simulation parameters, reflecting the change in a) RMSE and b) MAE.

not change with parameter optimisation. Therefore, the test's overall tendency to underestimate similarity at 80% and 90% while overestimating similarity at 70% generally remains. Different parameters would need to be selected if a core requirement of the test's application was to specifically hedge against either under- or overestimation. Similar to Type I and Type II errors where the choice of which error is less costly depends on the research context, the preference for under- or over-estimation will also depend on the application of the linear network spatial similarity test. Generally, a low point density and high grid size will increase the likelihood of overestimation while a high point density and low grid size would increase the likelihood of underestimation.

The optimal parameters themselves can also provide insight into the underlying workings of the spatial similarity test. Firstly, the optimal sliding window size $w = 7$ is larger than hypothesised since the granularity of the smaller window was expected to yield more accurate local similarity estimates. It was found, however, that the smaller sliding window sizes underestimate similarity, especially at the 90% similarity level. This indicates that the small window size is 'too local' and the final results are being skewed by the underlying randomness of the simulations. For real world application, it should therefore be known beforehand that the

spatial similarity test is not entirely robust against local randomness. On the other hand, the largest window size is 'too global' and overestimates similarity.

Secondly, the test performance deteriorates as the point density is decreased. This is shown in Figure 5.3 a), b) and c) within the regions representing simulations conducted using $5 \leq n \leq 10$. When n is too low, the data points become too sparse to accurately approximate the linear network. Fewer data points mean increased space between each point. A slight variation can thus lead to the sparse pixels being completely excluded from the sliding window centred at the reference pixel and thereby bias the similarity estimation. When m is low, there are fewer grids into which the image is divided. Even small changes in a few grids therefore represents a greater proportion of dissimilarity across all the grids when compared to a larger number of grids. Additionally, the simulations are mimicking mobility data and so these geographical data simulations stretch over a very large Cartesian area. This means that lower grid sizes will lose a lot of the granular details and the nuances in variations. As m increases, this is addressed, and underestimation generally decreases. It should be noted, however, that in cases of lower similarity, as shown in Figure 5.3 a), low point density and high grid sizes can actually start to result in slight overestimation. This is again just due to the fact that extremely sparse data does not approximate the underlying structure of the linear network well and so the true variation between networks is misrepresented.

Thirdly, to a lesser degree and for different reasons, performance also starts to decline when n is increased too much beyond the optimal parameter. In the case of higher similarity at 80% and 90%, rates of underestimation start to increase. As n increases, the approximation of the line becomes closer to the literal line segment. More spatial variation may therefore be detected as a result of arbitrary assumptions that simplify, and therefore change, certain parts of the true linear network. On the other hand, for lower similarity at 70%, underestimation generally does not occur. Instead, the general tendency of overestimation in the case of high dissimilarity starts to gradually decrease. The data pixels are no longer spread out too far and the sliding windows more often include the relevant similar pixel.

Fourthly, when m is too low, especially at $m \leq 10$, similarity is grossly overestimated in the case of lower similarity at 70%. The resolution is too low to detect the more subtle variations. When similarity is higher at 80% and 90%, the test seems to perform much better with only minimal deviation. However, it should be noted that this is likely not due to the accuracy of the test but rather due to the fact that there simply are not many variations in the linear networks for the test to oversimplify and therefore overestimate the similarity. When m is increased but still remains below the optimal range of parameter values, a general trend of underestimation can be noted in higher similarity at 80% and 90%. It is assumed that this effect is caused not by the test but by the way in which the linear networks were generated and the skewing effect of the simulations' randomness. Further simulations tests using alternative linear network generation methods could shed more light on this matter.

Finally, when m is too high above the optimal parameter, overestimation can be seen in the case of lower similarity at 70%. The simulated data set is, admittedly, comparatively sparse in relation to very large and complex real-world networks. Therefore, it is possible that the similarity is being overestimated because the

number of network pixels is much smaller than the number of empty space pixels and so the variation in the network pixels is not proportional to the variation in the total number of pixels. At 80% similarity, there does not seem to be a consistent relationship between increasing m and improving test performance. The rate of under- or over-estimation as well as if this rate increases or decreases as m increases varies based on n. This is again attributed to the data structure of the linear networks and the slightly inconsistent levels of similarity that occurred during the artificial generation of the 80% similar class of linear networks. For higher similarity at 90%, however, there is a consistent correlation between increasing m and decreasing the relatively low rate of underestimation, as shown in Figure 5.3 c).

In summary, the linear network spatial similarity test performs best across all variation levels and types when $\{w, n, m\} = \{7, 30, 45\}$. The sliding window size is large enough not to be too heavily biased by underlying randomness while it is still small enough to allow for localised comparison. The point density is high enough to not be too sparse for the relevant sliding windows and is low enough to not approximate the line segment, based on arbitrary assumptions, too stringently. The resolution grid size is also optimised to not be too low, which would cause oversimplification, nor is it too high so as to be biased by the large number of empty space pixels. The accuracy of the test is proven by the means of estimated similarity being close to the true similarities in Table 5.2. The average relative biases are low, as are the measures of variation, indicating that the test is consistently accurate when optimised. In the special case of prior knowledge of the approximate variation level or type, different parameters can be selected to further reduce bias. Similarly, different parameters can also be chosen to influence the likelihood of either under- or over-estimation.

5.6.3 Evaluating the performance of the spatial similarity test for the second simulation study

The Friedman test again results in a p-value less than 0.001 for both the sliding window size and the resolution grid size, which indicates a significant difference in means across both parameters. The exhaustive grid search in this case indicates that the optimal parameters for this data set are $\{w, m\} = \{7, 55\}$. This parameter combinations yields the highest accuracy for the spatial similarity test and is therefore the fairest comparison for the optimised point pattern method. The individual variation types, edge scaling and transformation, also have their own optimal grid sizes $m = 20$ and $m = 75$, respectively. Figures 8.8 a) and 8.8 b), included in the Appendix, show the different variation types have erratic patterns of under- and overestimation at most grid size choices.

The results in Table 5.3 show that the optimised unprocessed image simulations produce presumably adequate results. The average deviation from the true level of spatial similarity is 3.01%. The average relative bias across the different simulation scenarios further shows that this deviation mostly occurs due to underestimation. The test performs best for a combination of variation types. Additionally, the most accurate similarity estimates were made for the 70% similar simulations. Furthermore, the aggregated standard deviation across the different scenarios is 0.005 which indicates that the test is very reliable and consistent in its estimations.

		Performance metrics					
		Centrality	Bias	Precision		Accuracy	
Variation type	**Similarity level(%)**	**Mean(%)**	**Average relative bias**	**Standard deviation**	**Range**	**RMSE**	**MAE**
Scaling	70	**0.684**	-0.024	0.006	0.009	0.017	0.016
	80	0.755	-0.051	0.007	0.011	0.053	0.050
	90	0.945	0.050	0.014	0.020	0.050	0.045
Transformation	70	**0.711**	0.015	0.002	0.002	0.011	0.010
	80	0.756	-0.055	0.001	0.001	0.044	0.044
	90	0.836	-0.071	0.003	0.004	0.064	0.063
Combination	70	**0.697**	-0.004	0.002	0.003	0.004	0.003
	80	0.776	-0.043	0.004	0.005	0.025	0.024
	90	0.891	-0.011	0.006	0.008	0.010	0.009

Table 5.3: Performance metrics for aggregated similarity indices using the parameters $\{w, m\} = \{7, 55\}$ for the unprocessed image method.

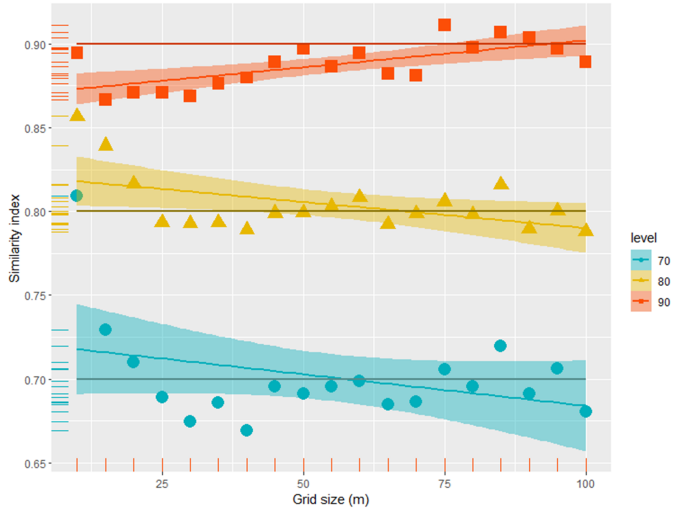


Figure 5.5: Scatter plot graphing the estimated similarity indices against the different grid sizes $10 \leq m \leq 100$. A combination of variation types are considered.

In terms of the sliding window size, it can be assumed that the same theoretical explanations provided for the point pattern method apply here as well. The effect of the resolution grid size, however, varies. This is more clearly visualised in Figure 5.5 by fixing the sliding window size to $w = 7$ and plotting the estimated spatial indices against the increasing grid sizes. The simulation results in Figure 5.5 specifically consider a combination of variation types. Figures 8.8 a) and 8.8 b) in the Appendix additionally prove this point for the specific cases of edge scaling and edge transformation, respectively. From Figure 5.5, it can be seen that increasing m decreases underestimation and improves accuracy for 90% similarity. Low resolution grid sizes, similar to the point pattern method, simply lack too much detail and the distorted structural representation of the linear network's structure results in oversimplification of the spatial similarity estimation. In the case of 80% similarity, deviations and bias remain relatively low once m has been increased beyond a reasonable threshold of $m = 20$. At 70% similarity, increasing m generally reduces the rates of under- and overestimation. As Figure 5.5 shows, however, both under- and overestimation occur relatively randomly. This indicates that this method's performance generally starts to deteriorate at lower levels of spatial similarity. The test is no longer able to reliably distinguish between true structural variations in the linear network and the randomness

of the linear networks' construction caused by the assumptions and data constraints.

5.6.4 Comparing the performance of the point pattern and unprocessed image method

The point pattern method is more accurate for the majority of simulation scenarios across the different variation types and levels, as shown in Figure 5.6. In Figure 5.6, it is shown that the point pattern method simulations have lower RMSE and MAE values for most scenarios with an average relative RMSE and MAE decrease of 30.4% and 27.7%, respectively. This is especially the case for higher levels of similarity at 80% and 90%. As shown in Table 5.3, these scenarios are underestimated. This is similar to the point pattern method, shown in Figure 5.3, where the point density values are high. The suggested reasoning for this is again that the linear network is too stringently dependent on arbitrary assumptions like road centre extraction method, line segment snapping method and data resolution.

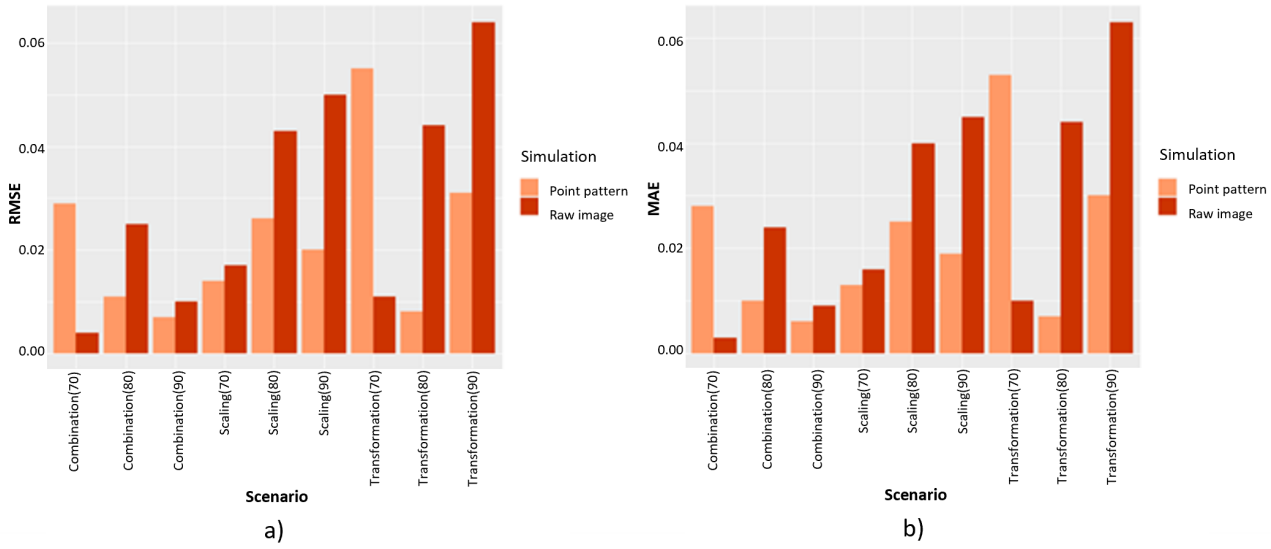


Figure 5.6: Accuracy metrics comparing the performance of the point pattern and unprocessed image methods. The comparison for a) RMSE and b) MAE is shown.

One exception to the improved performance of the point pattern method is the edge transformation scenario at 70% similarity. The unprocessed image method actually has a significant 80% and 81.1% relative decrease in RMSE and MAE, respectively, in this case. The mapping of deviations for the simulation scenario considering edge transformations at a 70% similarity level, shown in Figure 5.7, indicates that overestimation of similarity occurred for almost all the parameter combinations. Despite parameter optimisation, the test will therefore continue to overestimate similarity for linear networks with a higher degree of transformation-driven dissimilarity. Having said that, the rate of overestimation is marginally lower at higher point densities. The overall overestimation is reduced due to the increase of network pixels that counteract the majority of empty space pixels. Given that the continuous line used in the unprocessed pixelated image maximises the number of pixel between to line endpoints, the rate of overestimation will be minimised in this case. Therefore, the unprocessed image method is more accurate for this one scenario.

The results are also mimicked, to a lesser degree, in the combination of variation types which include

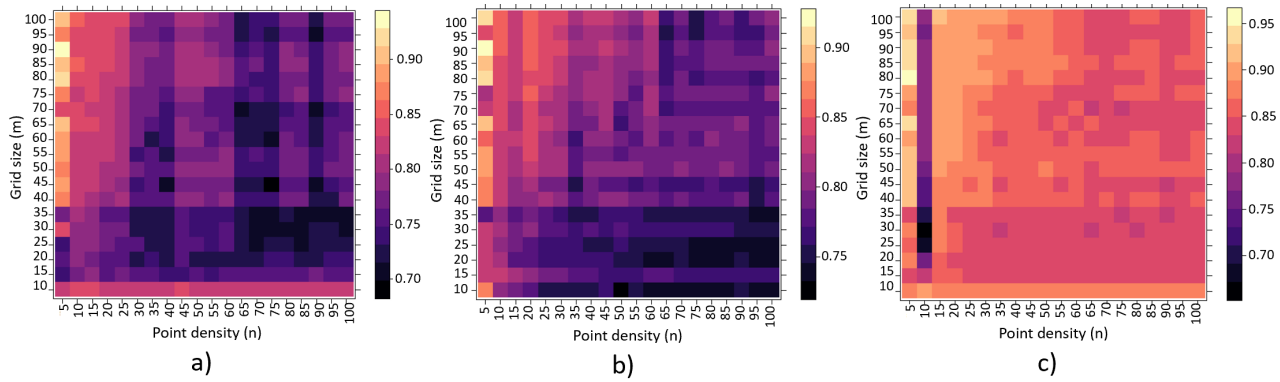


Figure 5.7: The simulation results using a sliding window of $w = 7$ are represented above. Permutations of different point density n and grid sizes m are tested for network variations based on edge transformations. The mean similarity indices are included for a) 70%, b) 80%, and c) 90% similar networks..

instances of edge transformation. This is a limitation of the point pattern method that is mitigated by choosing the optimal parameters, but that remain better addressed with the unprocessed image method. Therefore, the linear network spatial similarity test seems to perform better for higher levels of similarity using the point pattern method and better for lower levels of similarity using the unprocessed image method.

When considering the case of a combination of variation types, it can be seen, however, that the decrease in RMSE and MAE for the unprocessed image method applied to 70% similarity is still less than the overall decrease in RMSE and MAE for the point patterns method applied to 80% and 90%. To make the test accurate in as many scenarios as possible, the point pattern therefore remains the more robust method.

5.7 Conclusion

Chapter 5 focused on two individual simulation studies. The first study proved the reliable and accurate application of the new linear network spatial similarity test on different pairs of linear networks. The study also demonstrated how to optimise the three user-defined parameters used in the configuration of the spatial similarity test. The second study tested the efficacy of the proposed spatial similarity test against an alternative method. Each simulation study was conducted for three different levels of similarity at 70%, 80% and 90% similar linear networks as well as three different variation types including edge scaling, edge transformation and a combination of variation types.

For the first simulation study, the dependence of the test's accuracy on the user-defined parameters, namely the sliding window w, the point density n and the resolution grid size m, was tested. Once it was determined, by means of the Friedman test, that there was a relationship between the parameters and accuracy, grid searches were applied to determine the optimal parameters. The optimal parameters were found to be $w = 7$, $n = 30$ and $m = 45$ which yielded an overall accurate test with only an average deviation of 1.9% deviation from the true level of similarity.

The second simulation study tested the performance of the linear network spatial similarity test in the case where unprocessed images were used. The method was first optimised by determining the parameters in the optimal range, which were $w = 7$ and $m = 55$. The optimised unprocessed image method was then

compared to the optimised point pattern method, which showed that the point pattern method was superior in the case of 80% and 90% similarity, but that the unprocessed image method performed marginally better for 70% similarity.

Overall, the new linear network spatial similarity test performed well. It should be noted, however, that the test simulation only considered a mock mobility data set with seven destinations at a data resolution of 1:2000 m. The size and shape of the spatial domain was also kept constant during the simulations. Both the accuracy and robustness of the test can potentially be improved by considering simulated linear networks with different data generation methods, data resolution and spatial domain size and shape. Including more complex linear networks with many more nodes could possibly shed light on ways to mitigate the adverse effects of the empty space pixels.

In Chapter 6, the new linear spatial similarity test will be applied to two real-world linear networks, namely a social mobility network and an informal road network. The aim is to determine both the degree of global similarity and the distribution of local similarity. Additionally, three different subanalyses will be performed to investigate what effects road condition, seasons and road density have on the spatial similarity between the two networks.

Chapter 6

Application

The optimised linear network spatial similarity test is applied to the informal road network, shown in Figure 6.1 a), and the social mobility network, shown in Figure 6.1 b). The purpose is to understand both the degree and distribution of regional similarity as well as calculate a global similarity estimate for the two linear networks. The first step is to convert the two linear networks to point patterns. In the second step, the point patterns are represented as corresponding pixel images. The third step is to then generate a local similarity map from the two pixel images and, in the fourth step, a global similarity index is calculated based on the local similarity map. Firstly, the test is applied to the two unprocessed networks to estimate overall similarity.

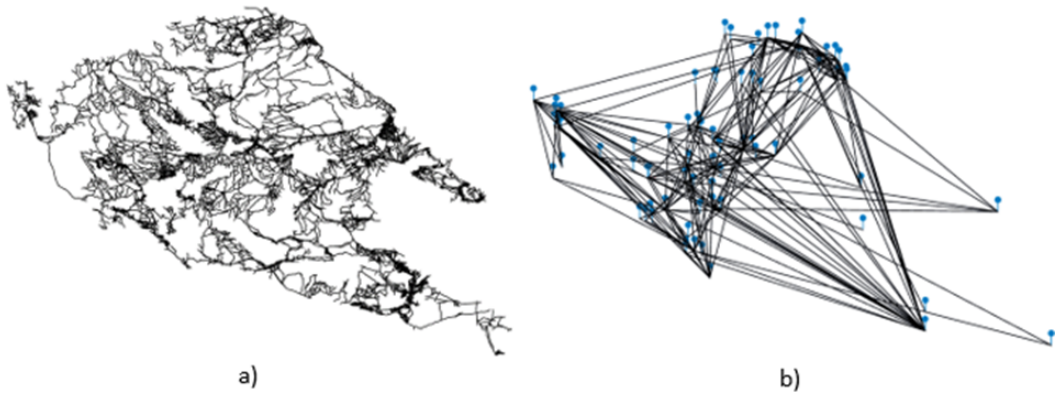


Figure 6.1: Two spatial linear networks being compared with the novel spatial similarity test, namely a) the informal road network, and b) the social mobility network.

Secondly, the informal road network is subdivided based on the different road conditions by means of texture classification. Satellite imagery of the informal roads is pre-processed and partitioned before texture features using grey level co-occurrence matrices and linear binary patterns are extracted. The features are classified using K-means clustering and the optimal number of feature classes, or road conditions, is determined. The spatial similarity test is applied to each of the individual road condition networks to determine if road conditions significantly influence how similar mobility routes and informal roads are.

Thirdly, the social mobility network is subdivided based on the different seasonal routes which are derived from the original mobility study's survey answers. The spatial similarity test is applied to each of the individual

seasonal mobility networks to determine if seasonal changes significantly influence mobility behaviour in relation to informal roads.

Finally, a real-world limitation of the informal road network, namely the biased effect of a disproportionately high density of roads in certain areas, is explained. A threshold-based method for removing subgraphs that are excluded from the scope of this project is proposed. The spatial similarity is applied to the updated informal road network to demonstrate the effect of pre-processing the linear networks on the final global similarity estimate.

6.1 General spatial similarity comparison

The two networks are first converted into unmarked point patterns using the optimised point density $n = 30$. They are represented as two pixel images, as shown in Figure 6.2, using the resolution grid size $m = 45$. The resulting local similarity map, seen in Figure 6.3, is generated based on the sliding window $w = 7$. The final global similarity index is estimated as the average of the local similarity map pixel values and yields a result of 0.377.

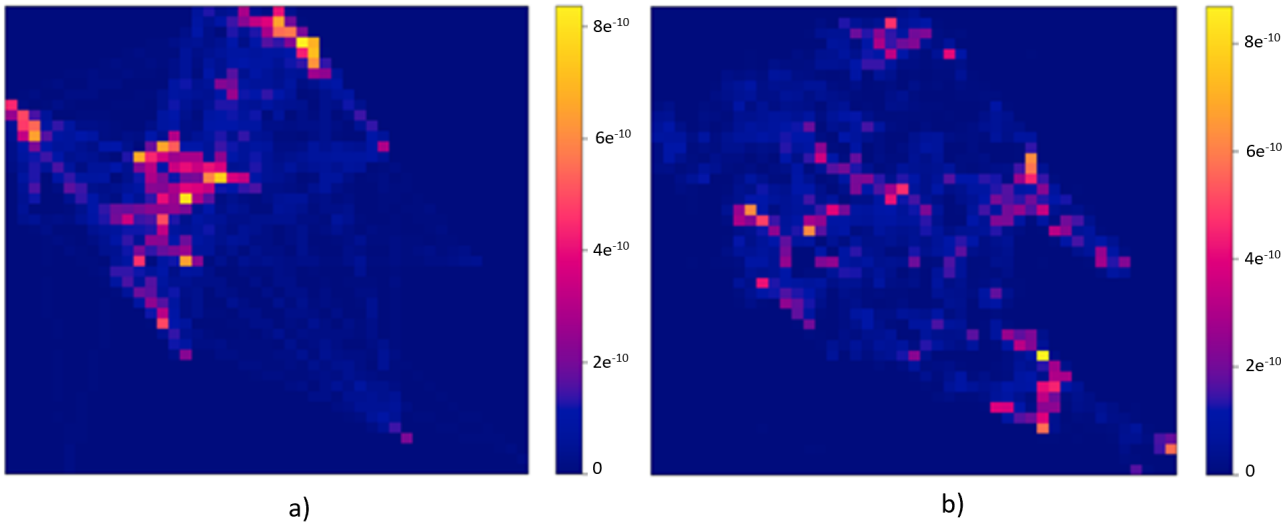


Figure 6.2: Pixel image representation using a point density $\{n, m\} = \{30, 45\}$ for a) the mobility network, and b) the informal road network.

The linear network spatial similarity indicates that the mobility and road networks are approximately 37.7% similar. When looking at the local similarity map in Figure 6.3, we can further see exactly which areas account for the similarity and which areas are blatantly dissimilar. The majority of the local map pixels with a similarity of 1 represent the common empty spaces between or around the linear network edges. As such, these pixels are removed to prioritise the visualisation of the linear network edges.

Interestingly, the areas of lowest similarity, marked in blue, correspond to the most populous villages. The reason for this is that these more developed villages have a much higher density of roads compared to the rest of the network. Rather than having one main road, there are many road edges of varying lengths, orientations and placement. The mobility network, on the other hand, consists of one simplified edge per route with a singular length, orientation and placement. The spatial similarity test detects this difference in both

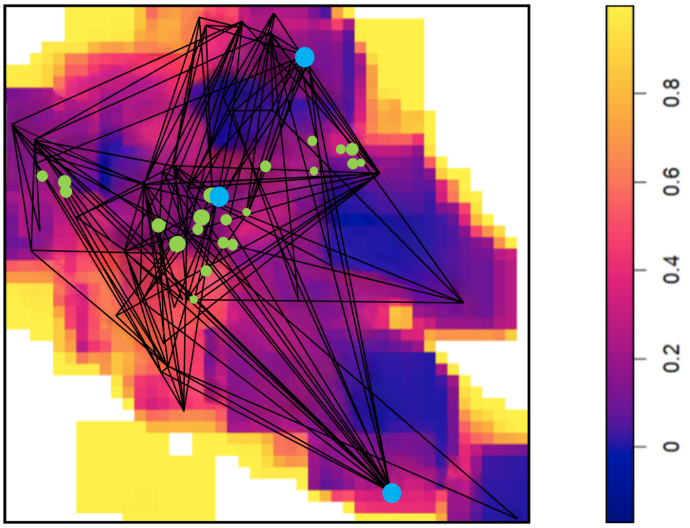


Figure 6.3: The local similarity map comparing the two pixel images of the mobility and road networks using $\{w, n, m\} = \{7, 30, 45\}$. The villages with the highest survey participation density are shown in green. The three key locations with more than 50 visitations each are shown in light blue.

the number and placement of these road edges and thus results in relatively low similarity. The similarity is comparatively higher in the more rural areas, especially on the outskirts of the investigated area, where there is often only one road through the village. The remoteness of the villages makes it more likely fewer people will visit these destinations. The density of informal roads decreases when only a few people visit a village. The spatial similarity of the two linear networks is therefore more likely to increase if the density of the road network is similar to that of the sparser mobility network. The test will then be more like a true 'one-to-one' spatial comparison of the two linear networks' placement as opposed to simply measuring the difference in network density.

The analysis of the three key locations, Opuwo, Okanguati and Epupa, further demonstrates that network density impacts the spatial similarity test's results. As discussed in Section 3.1.1, these key locations are the villages with more than 50 visits each. From Figure 6.3, it can be seen that these locations, indicated by light blue points, also occur in regions of moderately low spatial similarity, indicated in hues of pink and purple. To be specific, Opuwo, Okanguati and Epupa had aggregated global similarity indices of 0.461, 0.382 and 0.374, respectively. Two of the locations therefore have spatial similarity indices above the average while Epupa is very close to the average. This suggests that the increase of the mobility network, caused by the increased visitations and therefore the increased mobility routes, more closely mimics the denser, more complex road network. Therefore spatial similarity within the key locations can be seen to still be lower than the most remote villages, but is higher than the less frequently visited populous villages.

It should be noted, however, that the spatial similarity results may also, to a a degree, be affected by the underlying experimental design of the survey. Figure 6.3 shows that the area of relatively higher dissimilarity has very low survey participation rates, as indicated by the light green points. The original mobility study was specifically designed to focus more on rural villages. The most populous villages were excluded as origin villages. In these areas, dissimilarity is being caused by missing mobility routes as opposed to existing mobility routes that are truly different from the surrounding informal roads. The uneven distribution of data collection

can therefore be seen to slightly bias the spatial similarity test.

The local similarity map is helpful in answering the initial question of how connected the two linear networks are. By looking at the specific areas of similarity and dissimilarity, expert insight in other fields such as anthropology and geography can be used to understand what underlying factors cause the variation in similarity. For example, the cultural differences between villages with varying levels of spatial similarity can be studied to understand their underlying mobility behaviours. Similarly, the geography of areas with varying spatial similarity can be compared to investigate precisely what terrain best sustains travel routes. This could subsequently be used to predict future development of informal infrastructure. The novel linear network spatial similarity test can therefore serve as a useful diagnostic tool for many non-statistical fields in addition to its many statistical applications.

6.2 Seasonal spatial comparison

The social mobility network is classified into seven distinct seasonal patterns according to the mobility study's survey answers. The purpose of the subanalysis is to ascertain what effect seasonal changes have on mobility routes and specifically on the use of informal roads.

As previously mentioned in Chapter 3.1.1, the seasonal routes have different frequencies and average lengths. These differences influence the global similarity indices, which can be seen to vary across seasons. Firstly, Table 6.1 shows that most of the individual seasonal mobility routes yield a global similarity index above that of the overall mobility network. Firstly, the year-round mobility route has the smallest increase in similarity

	Rainy	**Winter**	**Dry**	**Year-round**	**Rainy and winter**	**Winter and dry**	**Dry and rainy**
Contribution to overall mobility volume(%)	9.612	17.871	19.835	45.722	2.173	2.538	2.249
Number of distinct mobility routes	3	96	88	140	12	14	10
Average length of mobility routes (km)	16.327	47.390	56.651	40.873	60.551	59.833	51.076
Global similarity index	0.622	0.515	0.404	0.393	0.496	0.427	0.369

Table 6.1: Metrics for the seven distinct seasonal routes.

at 39.3%. The local similarity map in Figure 6.4 d) shows that the most extreme dissimilarity occurred at the furthest destinations. This concurs with the descriptive statistics in Table 6.1 which show that year-round travel consisted of many short-distance trips. The central, nearby destinations are feasible year-round because adverse weather and seasonal effects are limited along short trips. The highest degree of dissimilarity is therefore around the edges of the informal road network, since people likely aren't able to travel that far all year long.

Secondly, the dry seasonal route has a 40.4% global similarity index. This data set consists of a moderate number of long-distance routes. More long-distance trips cause a difference in the distribution of extreme dissimilarity between mobility routes year-round and in the dry season, as seen in Figures 6.4 c) and d). It

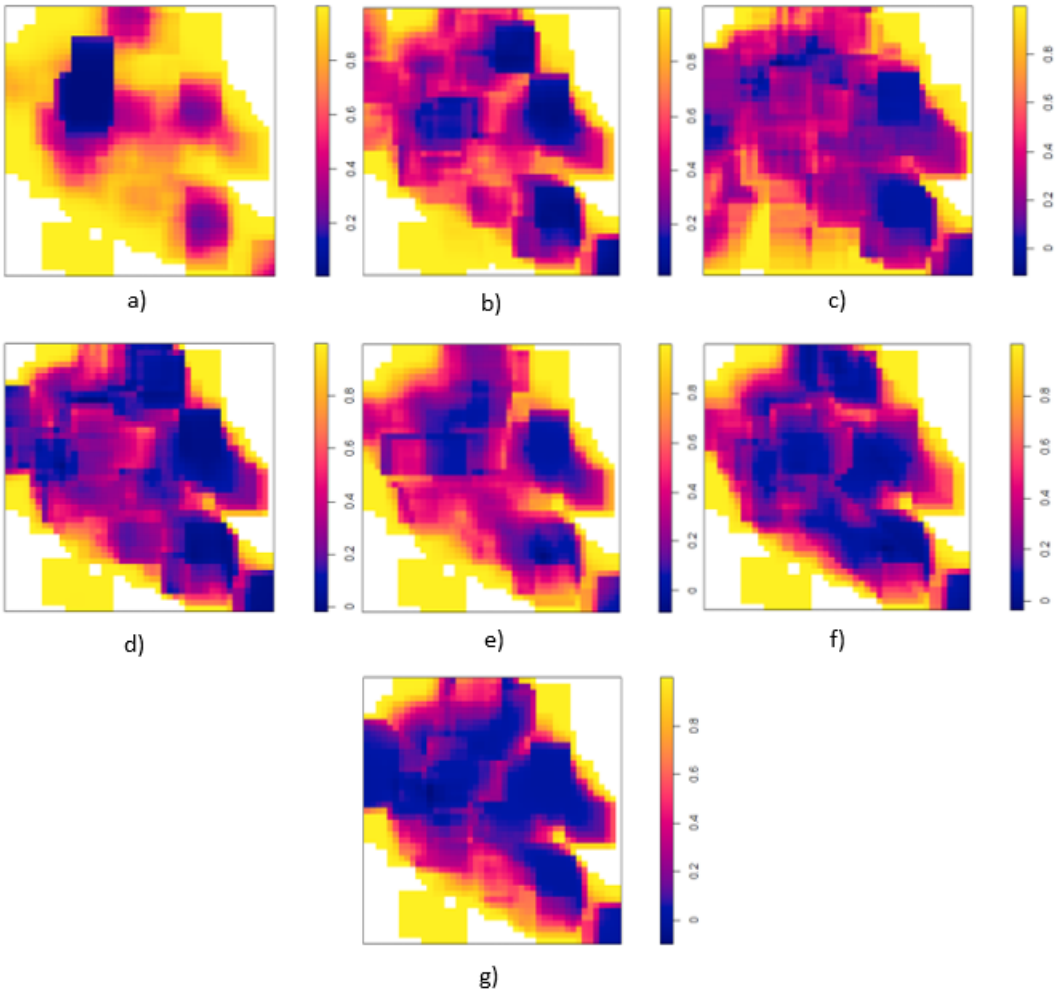


Figure 6.4: Local similarity maps for the different seasons, namely a) rainy, b) winter, c) dry, d) year-round, e) rainy and winter, f) winter and dry, and g) dry and rainy.

is possible that this increase in long-distance trips is due to dried up river crossings which enable seasonal passage. Figure 6.4 d) shows that it is only two of the largest cities where significant spatial similarity occurs. Similar to the overall mobility route compared in Chapter 6.1, this is possibly due to the disproportionately high density of roads in the high traffic villages.

Thirdly, the mobility network for people travelling in both the winter and dry seasons is 42.7% similarity when compared to the surrounding informal road network. This network set consists of a few long-distance routes. The local similarity map in Figure 6.4 f) shows that the dissimilarity occurs mostly in the central regions. It is likely that people choose to only travel during this season to reach destinations otherwise hindered by rivers. The high rates of dissimilarity in the central regions may be explained by the fact there are very few rivers in these areas and therefore people do not have a reason to travel here in this season exclusively.

Fourthly, the dry and rainy seasons have a 49.6% global similarity index. People who travel during these two seasons travel long distances to only a few destinations. From the local similarity map in 6.4 e) it can again be seen that the highest rates of dissimilarity occur at the most populous villages. The effect of the high road density in high-traffic villages is again assumed to therefore influence the similarity index.

Fifthly, the winter seasonal route has the highest overall global similarity index at 51.5%. This mobility consists of many shorter distance routes. The colder weather likely lowers people's tolerance for long-distance

travel, especially since most people either walk or travel by animal-drawn vehicles. These shorter routes can more easily follow along the informal roads and likely explain the higher spatial similarity index. The shorter the distance, the less the chance is of encountering geographical obstacles like a river or mountain which may cause the informal roads to deviate from the linear mobility routes. The local similarity map in Figure 6.4 b) shows that dissimilarity occurs most frequently in the most populous villages. As before, the high road density is likely contributing to the significant dissimilarity.

Sixthly, the dry and rainy seasonal route is the only route that has a global similarity index below the overall estimate at 36.9%. The data set contains a few long-distance routes. The long distances make it more likely to encounter geographical obstacles that cause the informal roads to deviate from the linear mobility routes. Additionally, given the small number of people that only travel during these two very specific seasons, it is likely that the people who travel do so out of necessity and not convenience. The routes are therefore not expected to be deterred by inconvenience such as inaccessible terrain or geographical obstacles. This makes it less likely that the routes will be spatially similar to the informal road networks which, for the sake of practicality, need to circumvent these deterrents. The local similarity map in Figure 6.4 g) shows that this phenomenon occurs across all terrains, since the distribution of dissimilarity is relatively even across the whole spatial domain.

Lastly, Table 6.1 suggests that the rainy season route is 62.2% similar to the informal road network. This result, however, is likely due to how small and scarce the rainy season mobility data set is. The similarity estimate appears to be severely overestimated in the case of comparing two linear networks of significantly different sizes. Future work could investigate possible implementations to mitigate this effect.

In summary, seasonal changes definitely change both the degree and distribution of spatial similarity between mobility routes and informal roads. Different seasons and weather conditions have varying effects on mobility behaviour like destination selection and distance tolerance. Similarity is highest during the winter season since the cold weather shortens routes and lessens the probability of encountering terrain that would deviate the informal roads from the mobility routes. On the other hand, similarity is lowest during the rainy and dry season, since the routes are more driven by need than consideration for terrain. This makes it more likely that informal roads will circumvent the inaccessible terrain and inevitably deviate from the linear mobility route. Additionally, it is suggested that the overestimation effect when comparing linear networks of significantly different sizes is investigated in future work.

6.3 Road condition spatial comparison

The texture analysis methods, as mentioned in Chapter 2.4, are applied to classify the road network according to their different road conditions. The purpose of this is to determine if certain road types are perhaps preferred when travelling and therefore more closely correspond to the mobility network.

The satellite imagery in ArcGIS corresponding to the digitised road polygons is subsetted and exported along with all the corresponding individual road polygons. Using the `R` packages `sp` [157], the full subsetted satellite image, shown in Figure 6.5 a), is further subdivided to only display the satellite image area corresponding

to the respective road polygon, as demonstrated by the blue outline. The central square title, shown in Figure 6.5 b), is then extracted from the image. The central squares have dimensions of 350×350 pixels. The size of the square tile is chosen to strike a balance between retaining detail and maintaining computational simplicity across the large data set. For the few very narrow digitised roads smaller than 350×350 pixels, the largest possible square tile is constructed. Figure 6.5 c) shows how the input tile is converted to a greyscale image to simplify overall computation.

Firstly, the grey level co-occurrence matrix features are extracted from the greyscale image according to the theory in Chapter 2.4.1. This is done in `R` using the `glcm` and `raster` packages [224, 76]. To emphasise the difference in greyscale pixel values, the greyscale input tile is shown in Figure 6.5 d) as a raster image. Five features are derived from the greyscale input tile, namely mean, variance, homogeneity, contrast and entropy. Each feature is calculated for each individual pixel, as shown in Figures 6.5 e) - i). For the purpose of feature extraction, the final value for each feature is calculated as the mean of the feature pixel values across the whole input tile.

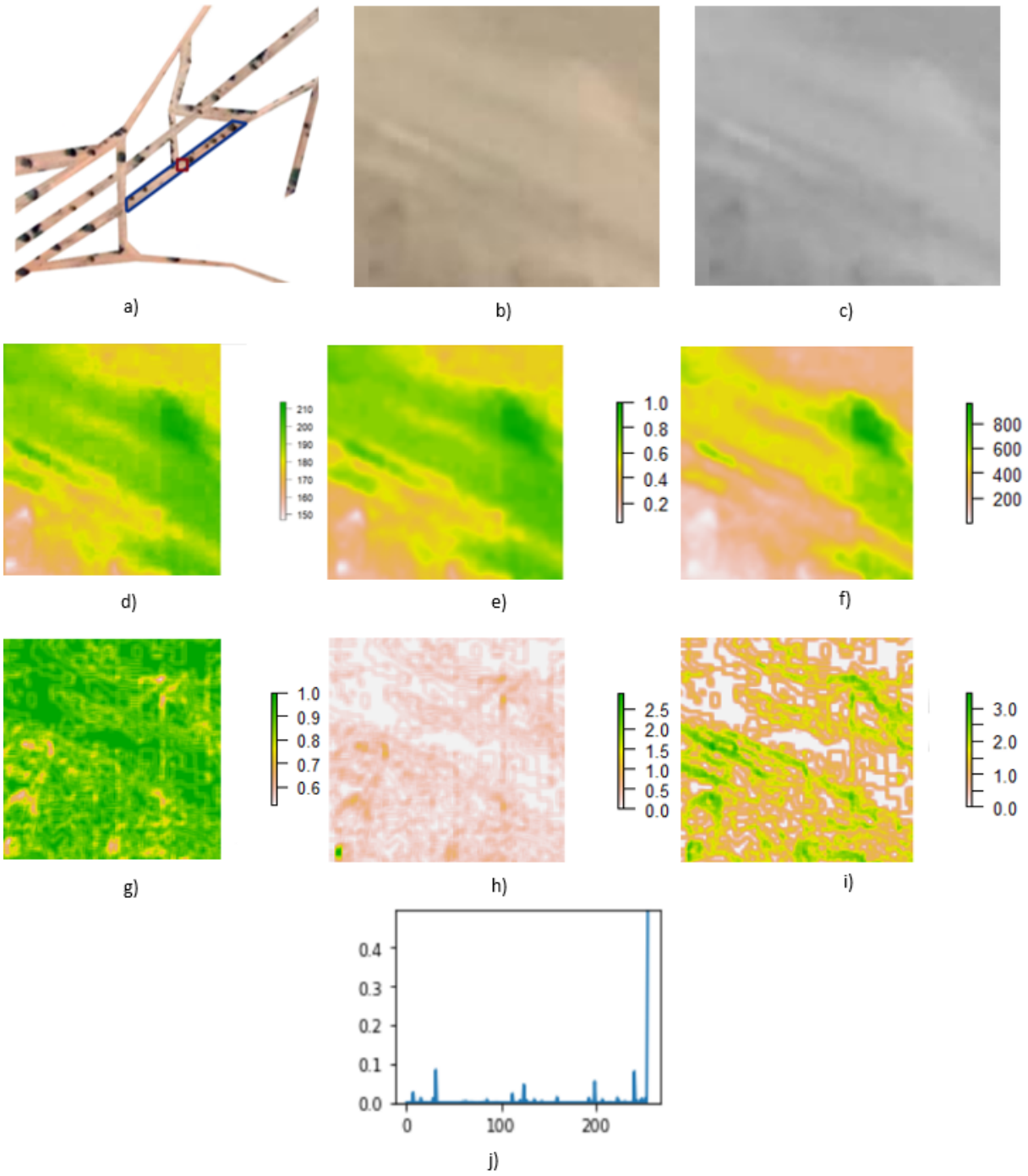


Figure 6.5: A visualisation of the main steps of the texture classification process. The satellite imagery is a) subdivided based on the outline of the respective road polygons, as shown in blue. The centre 350×350 pixel square tile, shown in red, is chosen as b) the input tile, and c) the greyscaled input tile is derived. To emphasise the difference in greyscale pixel values, the greyscale input tile is represented as d) a raster image. From the greyscale input tile, the e) mean, f) variance, g) homogeneity, h) contrast and i) entropy are derived. Additionally, the LBP histogram is calculated based on each pixel's linear binary pattern value.

Secondly, the linear binary values are calculated using the theory in Chapter 2.4.2. The linear binary values are derived in `Python` using the **`numpy`** and **`skimage`** packages [68, 192]. The linear binary values for each input tile are plotted along a histogram as seen in Figure 6.5 j). From the histogram, six additional features are calculated namely mean, standard deviation, median, skewness, kurtosis and entropy.

Each input tile therefore has 11 derived features. This is repeated for all the road polygons. The resulting feature vectors are clustered in $\mathbb{R}^{11}$ using K-means clustering. This is done in `R` with the **`cluster`** package

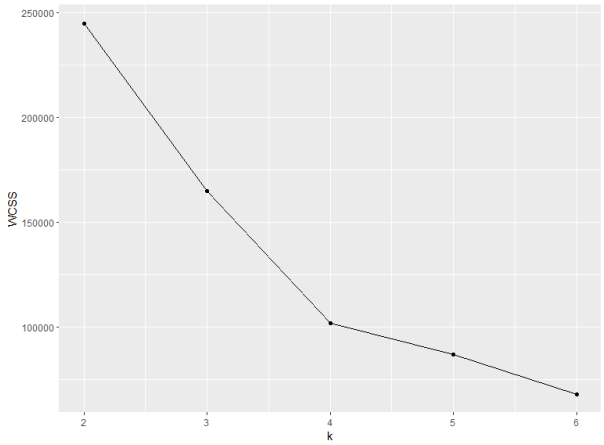


Figure 6.6: An elbow graph for the K-means clustering showing that the optimal number of clusters is $k = 4$.

[119]. The optimal number of texture classes is determined using the elbow method, as shown in Figure 6.6. Based on the elbow graph, the optimal number of road conditions is $K = 4$. The K-means clustering is a form of unsupervised classification, since there are no true labels for the different road conditions. Therefore, the four road conditions identified do not inherently have distinct semantic labels. However, when looking at Figure 6.7, we can see that examples of the four road conditions approximately look like sandy, muddy, rocky and vegetated roads. For the rest of the analysis, the four texture classes are therefore referred to according to these four approximate labels. Once each of the square tiles are classified, their class label is assigned to their

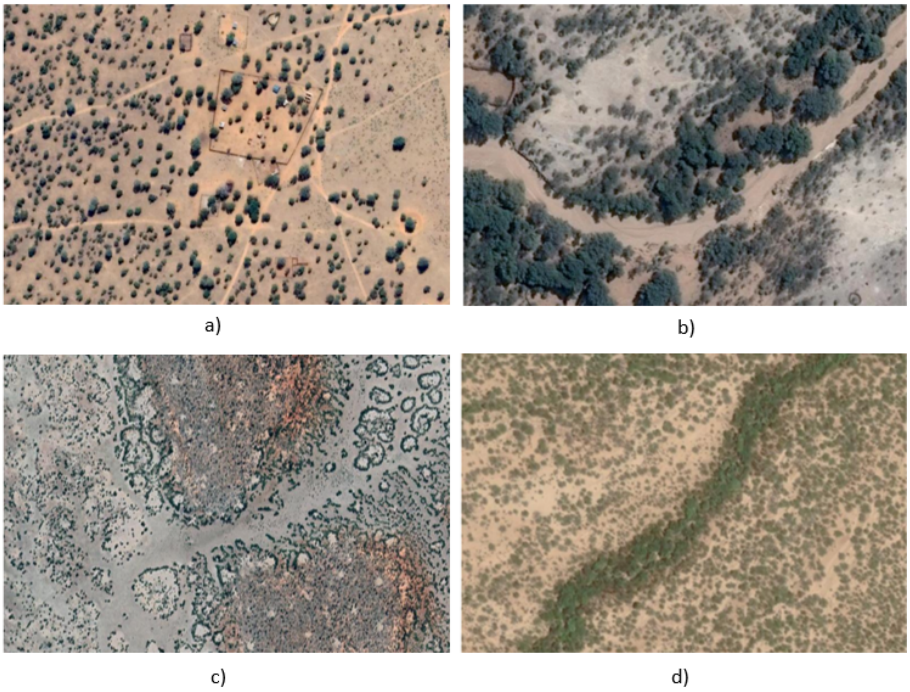


Figure 6.7: Examples of different roads conditions, namely a) sandy, b) muddy, c) rocky, and d) vegetated roads.

original corresponding road polygon. The overall road network is then filtered based on the class attribute and the four respective linear networks are derived, as shown in Figure 6.8.

From Table 6.2 it can be seen that the estimated global similarity indices for the different road conditions clearly differ. An interesting correlation is that the texture class with the lowest total surface area coverage,

vegetated roads, has the highest estimated global similarity indices. Similarly, rocky roads, which have the highest coverage, have the lowest estimated global similarity index. The coverage significantly affects the estimated similarity because of the differences in network density.

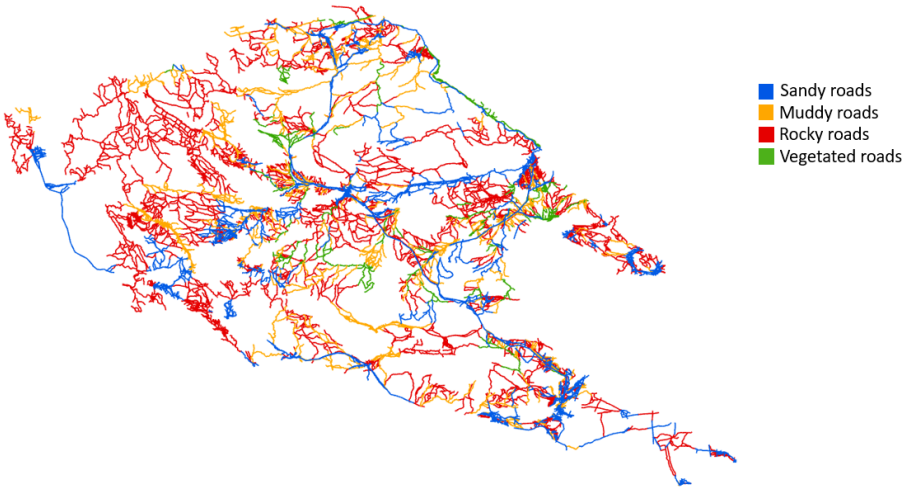


Figure 6.8: The distribution of different road conditions shown across the informal road linear network.

The issue of density is further exaggerated when considering that the road condition distribution in the most populous villages is highly skewed. Most of the roads are either faint, sandy roads caused by frequent walking around the inner village areas, or rocky roads which provide a stable foundation for infrastructure. The muddy and vegetated roads, on the other hand, appear to occur more frequently in the wilder, less densely inhabited areas in between the larger villages. The two former roads are thus more likely to occur at high densities and bias the spatial similarity test.

	Sandy roads	**Muddy roads**	**Rocky roads**	**Vegetated roads**
Coverage (%)	21.7	20.2	51.6	6.4
Global similarity index	0.319	0.315	0.312	0.537

Table 6.2: Global similarity indices for the four different road conditions.

Lastly, occurrences of the respective road conditions also seem to be geographically clustered. This makes sense as road conditions are predominantly shaped by the surrounding terrain and so roads within the same region are likely to be of a similar condition. This clustering causes discontinuity in the overall linear network. The mobility network is inherently more simplified and has more empty space between the network edges. As such, these new discontinuities in the road network are also interpreted as empty space. This explains why texture classes like vegetated roads with the largest discontinuities have the highest estimated similarity.

6.4 Improved density spatial comparison

The above-mentioned effect of the dense road networks within populous villages cannot be improved given that it is an accurate representation of the road networks at the mobility data locations. What can be improved, however, is the disproportionately high road densities in the vast areas between the locations and the effect this has on the spatial similarity test. Figure 6.10 demonstrates that there are a lot of road network edges

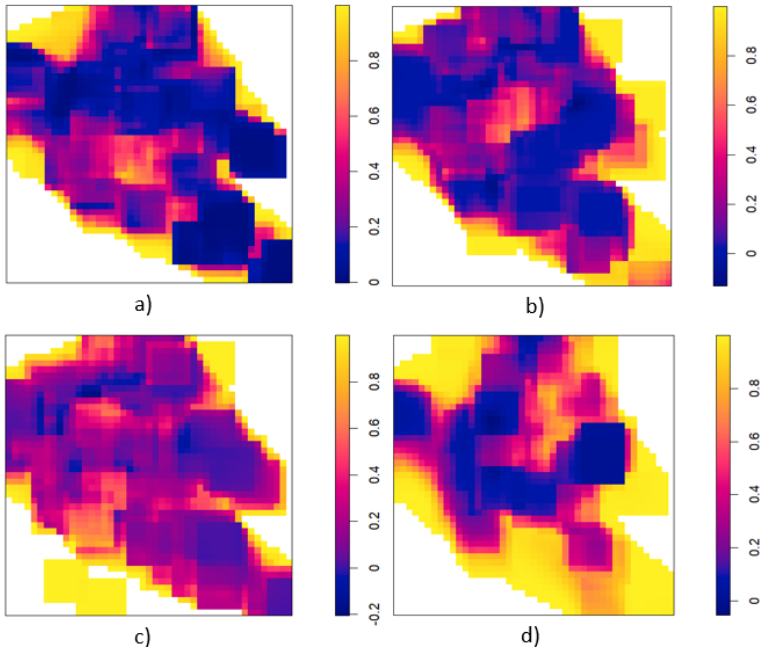


Figure 6.9: Local similarity maps for a) sandy, b) muddy, c) rocky, and d) vegetated roads.

that have been added as part of the digitisation process, but that are too far away from the mobility data to be considered relevant. Some of these roads lead to villages that are not included in the mobility data set and should thus not be compared to the mobility network. The aim of the test, after all, is to see how similar the existing infrastructure is to the mobility data, not how similar mobility data is to all surrounding infrastructure. These unrelated roads lead to a disproportional difference in density between the two linear networks and ultimately affect the spatial similarity estimation.

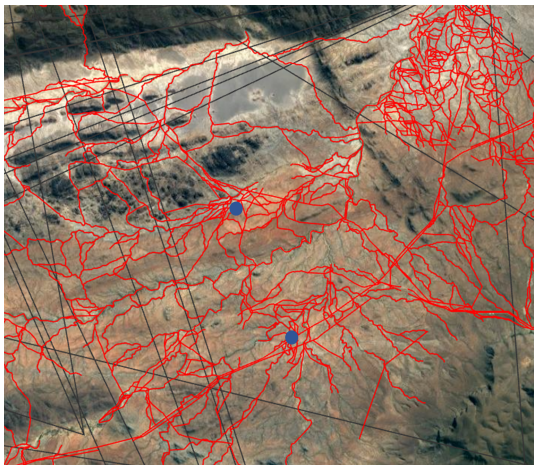


Figure 6.10: An example of two villages, marked in blue, not included in the mobility data set but which have roads leading to them [121].

To determine which roads to include, an inclusion buffer area around the mobility network is created. An edge is included if the midpoint of the edge is within a minimum distance to the closest mobility network edge. To formally express this, we first define the mobility linear network, L_M, and informal road linear network, L_R as

$$L_M = \bigcup_{l=1}^{281} e_{m,l} \tag{6.1}$$

and

$$L_R = \bigcup_{k=1}^{10689} e_{r,k} \tag{6.2}$$

where $e_{m,l}$ are the edges of the mobility linear network and $e_{r,k}$ are the edges of the informal road linear network.

Based on the theory in Chapter 2.1, each edge is defined as the shortest path between two endpoints. Therefore, the informal road edges can be defined as

$$e_{r,k} = \{tv_{k,i} + (1-t)v_{k,j} : 0 \leq t \leq 1\}. \tag{6.3}$$

Additionally, since the linear is spatial and exists within the $\mathbb{R}^2$ Euclidean space, each of the endpoint vertices have a x-coordinate and a y-coordinate. This can be written as

$$v_{k,i} = \{v_{k,i,x}, v_{k,i,y}\} \text{ and } v_{k,j} = \{v_{k,j,x}, v_{k,j,y}\}. \tag{6.4}$$

Similarly, we can express the mobility network edges as

$$e_{m,l} = \{tv_{l,i} + (1-t)v_{l,j} : 0 \leq t \leq 1\} \tag{6.5}$$

with

$$v_{l,i} = \{v_{l,i,x}, v_{l,i,y}\} \text{ and } v_{l,j} = \{v_{l,j,x}, v_{l,j,y}\}. \tag{6.6}$$

Each of the informal road edges $e_{r,k}$ has a midpoint p_k defined as

$$p_k = (p_{k,x}, p_{k,y}) = \left(\frac{v_{k,i,x} + v_{k,j,x}}{2}, \frac{v_{k,i,y} + v_{k,j,y}}{2}\right). \tag{6.7}$$

For each midpoint p_k, the minimum Euclidean distance between the midpoint and all mobility edges is then calculated. The minimum distance q_k is expressed as

$$q_k = \min \frac{\left| \left(v_{l,j,x} - v_{l,i,x}\right)\left(v_{l,i,y} - p_{k,y}\right) - \left(v_{l,i,x} - p_{k,x}\right)\left(v_{l,j,y} - v_{l,i,y}\right) \right|}{\sqrt{\left(v_{l,j,x} - v_{l,i,x}\right)^2 + \left(v_{l,j,y} - v_{l,i,y}\right)^2}}. \tag{6.8}$$

Finally, the updated informal road network L_R^* can be defined as

$$L_R^* = \{e_{r,k} : q_k \leq T\} \tag{6.9}$$

where T is a user-defined threshold.

The linear networks are plotted along an axis with increments of one ten-thousandths of a coordinate degree, as shown in Figure 4.1. The threshold is therefore based on a distance between the plotted coordinates. Alternatively, the threshold could also be defined in terms of metres or kilometres. The threshold would then first be converted to align with the scale of the linear networks.

For this mini-dissertation, a threshold of 2000 is implemented where the distance is measured in one ten-thousandths of a degree. The resulting network is 83.1% of the original coverage. This is shown in Figure 6.11 where the new network is denoted in red and the excluded line segments are in blue. The linear network spatial

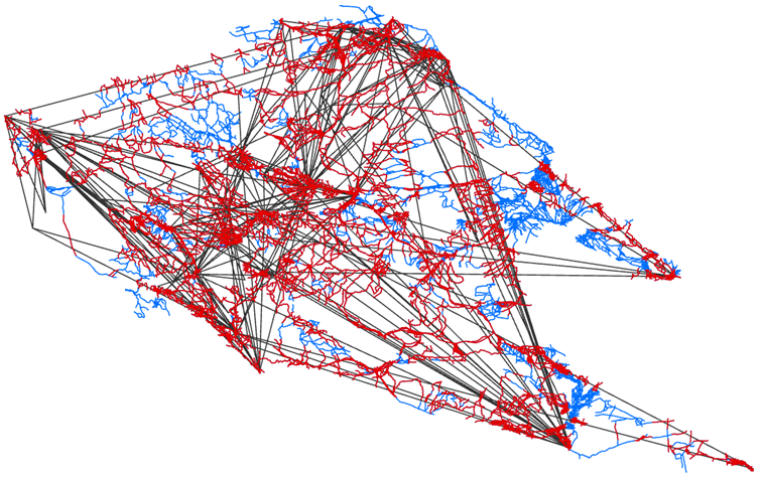


Figure 6.11: The updated informal road network where unrelated roads, shown in blue, are excluded if they lead to villages not considered in the mobility study.

similarity test is applied as before and a global similarity of 0.453 is estimated. This is a 20.2% increase in spatial similarity compared to the original road network. It is more than the 16.9% change in network coverage. When looking at the local similarity map in Figure 6.12, it can be seen that the regions corresponding to the removed network edges have a relatively high similarity. This is because the newly empty spaces between the road network are similar to the sparser mobility network. Spatial similarity therefore increases, but not proportional to the area size since some of the removed areas are dense and had a comparatively higher degree of dissimilarity.

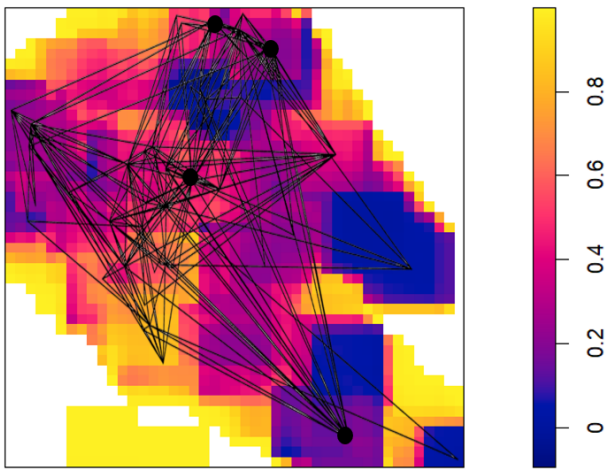


Figure 6.12: A local similarity map when comparing the updated road network with the mobility network.

6.5 Conclusion

Chapter 6 first demonstrated the application of the new linear network spatial similarity test on an informal road network and mobility network covering a rural area in northwestern Namibia. The test was conducted

using the optimised parameters discussed in the simulation studies. The two networks were seen to be 37.7% similar with populous village having the lowest similarity.

The test was then applied to the seven different seasonal routes. Each seasonal route had different global similarity indices, which seemed to correlate with other characteristics like route length and frequency. Similarity was highest for the short-distance routes during the winter season. On the other hand, similarity was lowest for the dry and rainy seasonal route due to the small sample size.

Next, the different road conditions were compared to the social mobility network. These road conditions were derived using texture analysis, specifically grey level co-occurrence matrices and linear binary patterns along with K-means clustering. Sandy, muddy, rocky and vegetated roads were classified. The four different road conditions each yielded a different level of similarity. The sparse vegetated roads had the highest similarity while the denser, more widely covered rocky roads had the lowest similarity.

During the application of the test, the effect of road density was seen to skew the test. As such, a threshold-based method of systematically and mathematically excluding unrelated road edges was applied. This increased the overall similarity between the informal road network and mobility network to 45.3%.

In summary, the linear network spatial test was applied without issue and indicated that there was a slightly below average similarity between informal roads and overall mobility within northwestern Namibia. The two systems are therefore connected in certain regions, especially in the central regions, but generally appear to be unconnected.

Chapter 7

Conclusion

The primary purpose of this mini-dissertation was to develop and demonstrate a novel linear network spatial similarity test. This was motivated by a current lack of robust spatial similarity tests that fully allow for spatio-temporal variation, take into account the spatial context of the data, and accommodate linear networks with varying characteristics.

The first step of the test is to convert the linear networks into point patterns. In the second step, the point patterns are converted into pixel images. The two pixel images are then used to generate a local similarity map using the structural similarity index measure (SSIM) in the third step. The fourth step calculates the global similarity index based on the adapted Andresen's S-index with a non-binary input. The test depends on selecting appropriate point density n, resolution grid size m and sliding window size w parameter values.

The first simulation study applied the linear network spatial similarity test to as many plausible scenarios as possible. This included linear networks of 70%, 80% and 90% similar as well as linear networks with different spatial dissimilarities including edge scaling, edge transformation and a combination of variation types. With the optimised parameters $\{w, n, m\} = \{7, 30, 45\}$, the test yielded good results with an average deviation of only 1.9%. The second simulation study demonstrated the benefits of using point patterns as opposed to simply pixelating unprocessed images of the linear networks. The study concluded that the new proposed method performs better at 80% and 90% levels of similarity while the unprocessed image method did marginally better at 70% when variation was caused by edge transformations.

Finally, the application of the linear network spatial similarity test on the road and mobility data across northwestern Namibia suggested that the two networks were initially 37.7% similar. This was increased to 45.3% when the road network was altered to exclude all unrelated road edges leading to villages not included in the mobility data set. The two linear networks are therefore generally dissimilar, with certain central regions having an above average degree of similarity. Additionally, the effects of road conditions and seasons were both shown to effect spatial similarity. The winter season consisting of many short-distance trips was the most similar to the surrounding informal road network. On the other hand, the vegetated roads with its low coverage, low road density and closely clustered occurrences had the highest similarity across the different road conditions. The degree and distribution of spatial similarity can therefore be further analysed and better

understood with the novel linear network spatial similarity test.

The first suggestion for future work, based on what was observed in the simulations and application, would be to apply the novel spatial similarity test to a more diverse range of linear networks. For example, the simulations were only carried out for linear networks with a resolution of 1:2000 m so as to make any conclusions relevant to the final application. Varying the resolution as well as changing other characteristics such as network complexity, density and overall spatial domain size may result in a more robust test. More data on how the test performs on different types of linear networks may also contribute to mitigating the observed overestimation of similarity when two linear networks of significantly different sizes are compared. Additionally, testing linear networks simulated using different generation algorithms and assumptions would also improve the reliability of the results. It should be noted that the characteristics of networks can rarely be controlled when using real data and so it is important to incorporate expert insight to ensure the varied linear networks remain realistic and practical.

Another suggestion for future work is to conduct a comprehensive sensitivity analysis. Throughout this mini-dissertation, assumptions were made during the test configuration and certain parameters were fixed according to suggestions from previous literature. For example, the weighting of luminance, contrast and structure were assumed to be equal while the sliding window was assumed to always be square. These assumptions, however, are not required. By varying these different parameters, the linear network spatial similarity test's robustness can be further evaluated.

A final suggestion for future work is to extend the linear network spatial similarity test and apply it to linear networks that are located in different spatial domains. This would enable the test to specifically be used to test how structurally similar two linear networks are. Based on the existing generic spatial similarity test, applying various rotations to the linear networks could achieve such a comparison [94]. Additionally, the test can also be applied to linear networks with differently shaped spatial domains including non-rectangular windows.

In summary, this mini-dissertation

1. Outlined the methodology of the novel linear network spatial similarity test by extending the application of an existing spatial similarity test in [94].

2. Performed the first simulation study in which the new spatial similarity test was applied to simulated linear networks to assess the test's overall performance.

3. Performed the second simulation study to prove the efficiency of the new method, especially the use of point patterns, compared to an alternative method.

4. Applied the spatial similarity test to the surveyed mobility data and digitised informal road data networks.

5. Applied the spatial similarity test to the different seasonal routes to determine seasonal variations in spatial similarity.

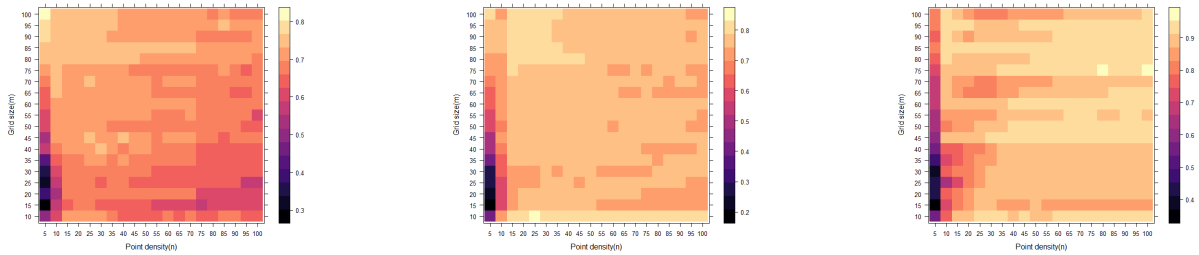
6. Classified the informal road network according to road conditions using texture analysis and applied the test to each respective network to determine the effect of road conditions on spatial similarity.

7. Improved the comparison of the road and mobility networks by systematically removing all unrelated road edges from the network and mitigating the effect of disproportionately high road densities.

In conclusion, this novel work is amazing. Its ease of use, reliable accuracy and insightful interpretability all pave the way for many exciting and versatile application cases to come.

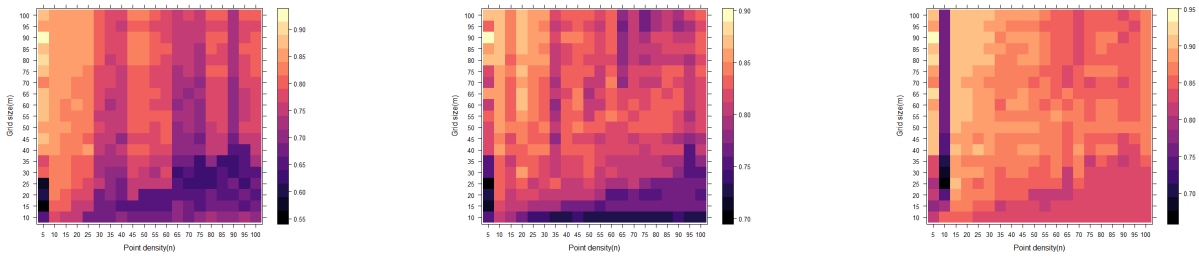
Chapter 8

Appendix



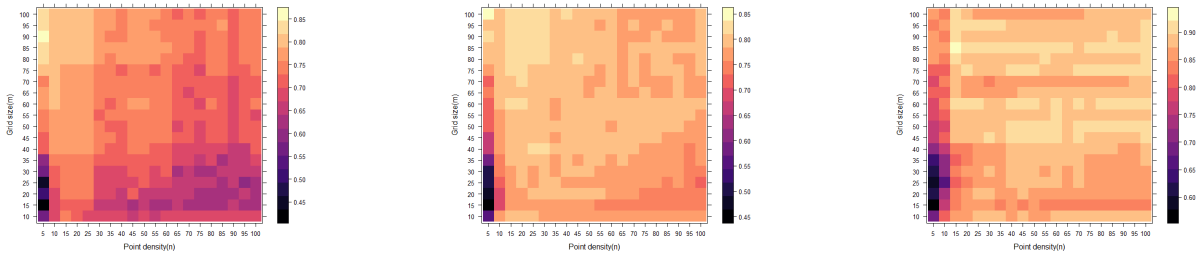
(a) (b) (c)

Figure 8.1: The simulation results using a sliding window of $w = 3$ are represented above. Permutations of different point density n and grid sizes m are tested for network variations based on edge scaling. The mean similarity indices are included for a) 70%, b) 80%, and c) 90% similar networks.



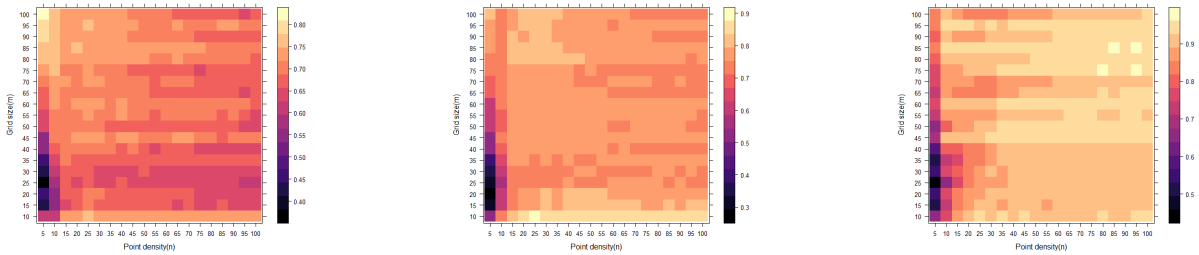
(a) (b) (c)

Figure 8.2: The simulation results using a sliding window of $w = 3$ are represented above. Permutations of different point density n and grid sizes m are tested for network variations based on edge transformations. The mean similarity indices are included for a) 70%, b) 80%, and c) 90% similar networks.



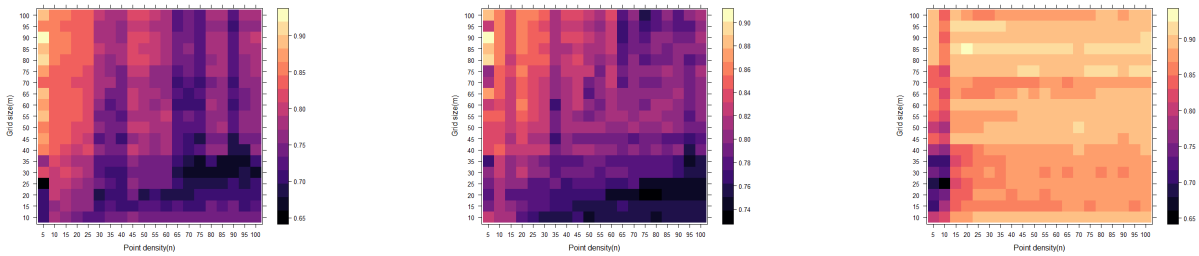
(a) (b) (c)

Figure 8.3: The simulation results using a sliding window of $w = 3$ are represented above. Permutations of different point density n and grid sizes m are tested for network variations based on a combination of edge transformations. The mean similarity indices are included for a) 70%, b) 80%, and c) 90% similar networks.



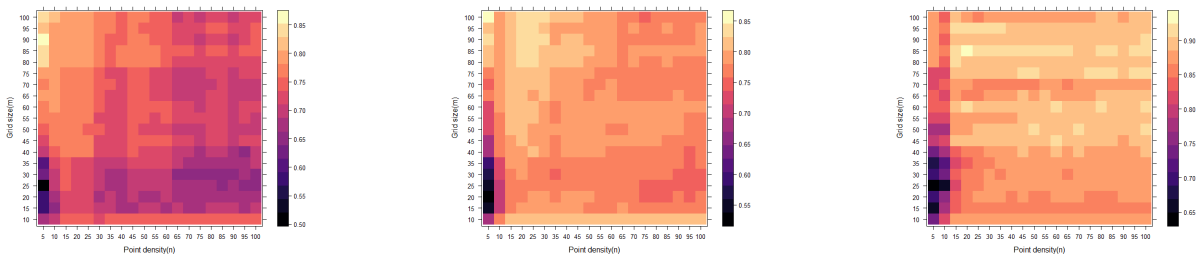
(a) (b) (c)

Figure 8.4: The simulation results using a sliding window of $w = 5$ are represented above. Permutations of different point density n and grid sizes m are tested for network variations based on edge scaling. The mean similarity indices are included for a) 70%, b) 80%, and c) 90% similar networks.



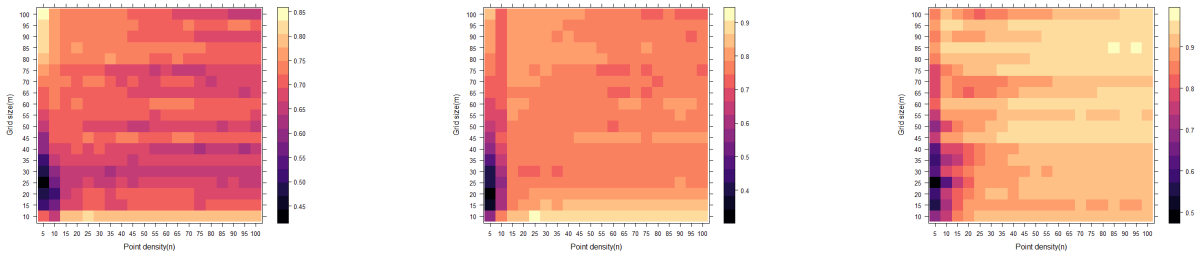
(a) (b) (c)

Figure 8.5: The simulation results using a sliding window of $w = 5$ are represented above. Permutations of different point density n and grid sizes m are tested for network variations based on edge transformations. The mean similarity indices are included for a) 70%, b) 80%, and c) 90% similar networks.



(a) (b) (c)

Figure 8.6: The simulation results using a sliding window of $w = 5$ are represented above. Permutations of different point density n and grid sizes m are tested for network variations based on a combination of edge transformations. The mean similarity indices are included for a) 70%, b) 80%, and c) 90% similar networks.



(a) (b) (c)

Figure 8.7: The simulation results using a sliding window of $w = 7$ are represented above. Permutations of different point density n and grid sizes m are tested for network variations based on edge scaling. The mean similarity indices are included for a) 70%, b) 80%, and c) 90% similar networks.

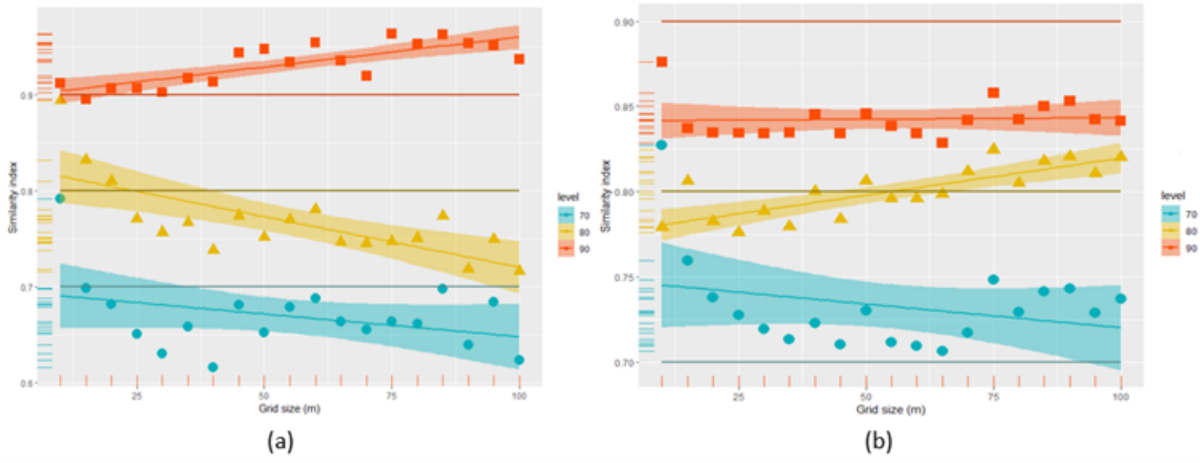


Figure 8.8: Scatter plots graphing the unprocessed image method's estimated similarity indices against the different resolution grid sizes $10 \leq m \leq 100$. The results are shown for a) edge scaling, and b) edge transformation.

Bibliography

[1] Arun Advani and Bansi Malde. Methods to identify linear network models: A review. *Swiss Journal of Economics and Statistics*, 154(1):1–16, 2018.

[2] Mohd Hanafi Ahmad Hijazi, Frans Coenen, and Yalin Zheng. Retinal image classification using a histogram-based approach. In *The 2010 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7, 2010.

[3] Sadia E Ahmed, Carlos M Souza, Júlia Riberio, and Robert M Ewers. Temporal patterns of road network development in the Brazilian Amazon. *Regional Environmental Change*, 13:927–937, 2013.

[4] Ghanim Al-Hasani, Md Asaduzzaman, and Abdel-Hamid Soliman. Geographically weighted Poisson regression models with different kernels: Application to road traffic accident data. *Communications in Statistics: Case Studies, Data Analysis and Applications*, 7(2):166–181, 2021.

[5] Laura Alessandretti, Piotr Sapiezynski, Sune Lehmann, and Andrea Baronchelli. Multi-scale spatio-temporal analysis of human mobility. *PloS One*, 12(2):e0171686, 2017.

[6] Martin A Andresen. Testing for similarity in area-based spatial patterns: A nonparametric Monte Carlo approach. *Applied Geography*, 29(3):333–345, 2009.

[7] Qi Wei Ang, Adrian Baddeley, and Gopalan Nair. Geometrically-corrected second order analysis of events on a linear network, with applications to ecology and criminology. *Scandinavian Journal of Statistics*, 39(4):591–617, 2012.

[8] Wei Ang. Statistical methodologies for events in a linear network. *University of Western Australia*, 2010.

[9] Carlo Arcelli, Gabriella Sanniti Di Baja, and Luca Serino. Distance-driven skeletonization in voxel images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(4):709–720, 2010.

[10] Cagri Aslan, Aykut Erdem, Erkut Erdem, and Sibel Tari. Disconnected skeleton: Shape at its absolute scale. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(12):2188–2203, 2008.

[11] Renato Martins Assunção and Danilo Lourenço Lopes. Testing association between origin-destination spatial locations. In *Advances in Geoinformatics: VIII Brazilian Symposium on GeoInformatics, GEOINFO 2006, Campos do Jordão (SP), Brazil, November 19–22, 2006*, pages 293–304. Springer, 2007.

[12] Sercan Aygün and Ece Olcay Günes. A benchmarking: Feature extraction and classification of agricultural textures using LBP, GLCM, RBO, neural networks, k-NN, and random forest. In *2017 6th International Conference on Agro-Geoinformatics*, pages 1–4, 2017.

[13] Adrian Baddeley, Ege Rubak, and Rolf Turner. *Spatial point patterns: Methodology and applications with R*. CRC press, 2015.

[14] Adrian Baddeley and Rolf Turner. spatstat: An R package for analyzing spatial point patterns. *Journal of Statistical Software*, 12(6):1–42, 2005.

[15] Albert-László Barabási and Eric Bonabeau. Scale-free networks. *Scientific American*, 288(5):60–69, 2003.

[16] Annalisa Barla, Francesca Odone, and Alessandro Verri. Histogram intersection kernel for image classification. In *Proceedings 2003 International Conference on Image Processing (Cat. No. 03CH37429)*, volume 3, pages III–513. IEEE, 2003.

[17] Marc Barthélemy. Spatial networks. *Physics reports*, 499(1-3):1–101, 2011.

[18] Fabio Baselice and Giampaolo Ferraioli. Unsupervised coastal line extraction from SAR images. *IEEE Geoscience and Remote Sensing Letters*, 10(6):1350–1354, 2013.

[19] Michael Batty. A new theory of space syntax. *University of College London*, 2004.

[20] Armando Bazzani, Bruno Giorgini, Sandro Rambaldi, Riccardo Gallotti, and Luca Giovannini. Statistical laws in urban mobility from microscopic GPS data in the area of Florence. *Journal of Statistical Mechanics: Theory and Experiment*, 2010(05):P05001, 2010.

[21] Mariano G Beiró, André Panisson, Michele Tizzoni, and Ciro Cattuto. Predicting human mobility through the assimilation of social media traces into mobility models. *EPJ Data Science*, 5:1–15, 2016.

[22] Yoshua Bengio. Gradient-based optimization of hyperparameters. *Neural computation*, 12(8):1889–1900, 2000.

[23] James Bergstra and Yoshua Bengio. Random search for hyper-parameter optimization. *Journal of machine learning research*, 13(2), 2012.

[24] Robert M Beyer, Jacob Schewe, and Hermann Lotze-Campen. Gravity models do not explain, and cannot predict, international migration dynamics. *Humanities and Social Sciences Communications*, 9(1):1–10, 2022.

[25] Norman Biggs, E Keith Lloyd, and Robin J Wilson. *Graph Theory, 1736-1936*. Oxford University Press, 1986.

[26] Stephen P Borgatti. Centrality and network flow. *Social Networks*, 27(1):55–71, 2005.

[27] MI Borrajo, W González-Manteiga, and MD Martínez-Miranda. Testing for significant differences between two spatial patterns using covariates. *Spatial Statistics*, 40:100379, 2020.

[28] Nikolaos G Bourbakis. A parallel-symmetric thinning algorithm. *Pattern Recognition*, 22(4):387–396, 1989.

[29] AO Brandão Jr and CM Souza Jr. Mapping unofficial roads with Landsat images: A new tool to improve the monitoring of the Brazilian Amazon rainforest. *International Journal of Remote Sensing*, 27(1):177–189, 2006.

[30] Dirk Brockmann, Lars Hufnagel, and Theo Geisel. The scaling laws of human travel. *Nature*, 439(7075):462–465, 2006.

[31] Anna D Broido and Aaron Clauset. Scale-free networks are rare. *Nature Communications*, 10(1):1–10, 2019.

[32] F Buekenhout and PJ Cameron. Projective and affine geometry over division rings. Chapter 2 of the Handbook of Incidence Geometry, 1995.

[33] Gerald AP Carrothers. An historical review of the gravity and potential concepts of human interaction. *Journal of the American Institute of Planners*, 22(2):94–102, 1956.

[34] Ylenia Casali and Hans R. Heinimann. A topological analysis of growth in the Zurich road network. *Computers, Environment and Urban Systems*, 75:244–253, 2019.

[35] Olivier Chapelle, Patrick Haffner, and Vladimir N Vapnik. Support vector machines for histogram-based image classification. *IEEE transactions on Neural Networks*, 10(5):1055–1064, 1999.

[36] Fei Chen, Zengqiang Chen, Xiufeng Wang, and Zhuzhi Yuan. The average path length of scale free networks. *Communications in Nonlinear Science and Numerical Simulation*, 13(7):1405–1410, 2008.

[37] JiaZhou Chen, Qi Lei, YongWei Miao, and QunSheng Peng. Vectorization of line drawing image based on junction analysis. *Science China Information Sciences*, 58(7):1–14, 2015.

[38] Philippa Clarke, Jana A Hirsch, Robert Melendez, Meghan Winters, Joanie Sims Gould, Maureen Ashe, Sarah Furst, and Heather McKay. Snow and rain modify neighbourhood walkability for older adults. *Canadian Journal on Aging/La Revue canadienne du vieillissement*, 36(2):159–169, 2017.

[39] C Comas, S Costafreda-Aumedes, N López, and C Vega-Garcia. On the correlation structure between point patterns and linear networks. *Spatial Statistics*, 29:192–203, 2019.

[40] David R Cox. On the estimation of the intensity function of a stationary point process. *Journal of the Royal Statistical Society: Series B (Methodological)*, 27(2):332–337, 1965.

[41] David Roxbee Cox and Valerie Isham. *Point Processes*, volume 12. CRC Press, 1980.

[42] Ottmar Cronie, Mehdi Moradi, and Jorge Mateu. Inhomogeneous higher-order summary statistics for point processes on linear networks. *Statistics and Computing*, 30(5):1221–1239, 2020.

[43] Paolo Crucitti, Vito Latora, and Sergio Porta. Centrality measures in spatial networks of urban streets. *Physical Review E*, 73(3):036125, 2006.

[44] Balázs Cs Csáji, Arnaud Browet, Vincent A Traag, Jean-Charles Delvenne, Etienne Huens, Paul Van Dooren, Zbigniew Smoreda, and Vincent D Blondel. Exploring the mobility of mobile phone users. *Physica A: Statistical Mechanics and its Applications*, 392(6):1459–1473, 2013.

[45] Pasquale De Meo, Emilio Ferrara, Giacomo Fiumara, and Alessandro Provetti. Generalized Louvain method for community detection in large networks. In *2011 11th International Conference on Intelligent Systems Design and Applications*, pages 88–93. IEEE, 2011.

[46] Yves-Alexandre De Montjoye, César A Hidalgo, Michel Verleysen, and Vincent D Blondel. Unique in the crowd: The privacy bounds of human mobility. *Scientific Reports*, 3(1):1–5, 2013.

[47] Peter Diggle. A kernel method for smoothing point process data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 34(2):138–147, 1985.

[48] Alexander Erath, Michael Löchl, and Kay W Axhausen. Graph-theoretical analysis of the Swiss road and railway networks over time. *Networks and Spatial Economics*, 9(3):379–400, 2009.

[49] James Evans, Jennifer O'Brien, and Beatrice Ch Ng. Towards a geography of informal transport: Mobility, infrastructure and urban sustainability from the back of a motorbike. *Transactions of the Institute of British Geographers*, 43(4):674–688, 2018.

[50] Inger Fabris-Rotelli, Abraham Wannenburg, Gao Maribe, Renata Thiede, Maribe Vogel, Mila Coetzee, Kutloano Sethaelo, Ephent Selahle, Pravesh Debba, and Victoria Rautenbach. An informal road detection neural network for societal impact in developing countries. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 4:267–274, 2022.

[51] Arthur Ahmad Fauzi, Fitri Utaminingrum, and Fatwa Ramdani. Road surface classification based on LBP and GLCM features using k-NN classifier. *Bulletin of Electrical Engineering and Informatics*, 9(4):1446–1453, 2020.

[52] Lars Forslöf and Hans Jones. Roadroid: Continuous road condition monitoring with smart phones. *Journal of Civil Engineering and Architecture*, 9(4):485–496, 2015.

[53] A Stewart Fotheringham and Morton E O'Kelly. *Spatial Interaction Models: Formulations and Applications*. Kluwer Academic Publishers Dordrecht, 1989.

[54] Peter I Frazier. Bayesian optimization. In *Recent advances in optimization and modeling of contemporary problems*, pages 255–278. Informs, 2018.

[55] Riccardo Gallotti, Armando Bazzani, Sandro Rambaldi, and Marc Barthelemy. A stochastic model of randomly accelerated walkers for human mobility. *Nature Communications*, 7(1):1–7, 2016.

[56] Sébastien Gambs, Marc-Olivier Killijian, and Miguel Núñez del Prado Cortez. Show me how you move and I will tell you who you are. In *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Security and Privacy in GIS and LBS*, pages 34–41, 2010.

[57] Andres J Garcia, Deepa K Pindolia, Kenneth K Lopiano, and Andrew J Tatem. Modeling internal migration flows in sub-Saharan Africa using census microdata. *Migration Studies*, 3(1):89–110, 2015.

[58] Michael T Gastner and Mark EJ Newman. Optimal design of spatial distribution networks. *Physical Review E*, 74(1):016117, 2006.

[59] Anthony C Gatrell, Trevor C Bailey, Peter J Diggle, and Barry S Rowlingson. Spatial point pattern analysis and its application in geographical epidemiology. *Transactions of the Institute of British Geographers*, pages 256–274, 1996.

[60] Jean Gibbons. Nonparametric statistical methods. *Technometrics*, 16:477–478, 04 2012.

[61] Muhittin Gökmen and Richard W Hall. Parallel shrinking algorithms using 2-subfields approaches. *Computer Vision, Graphics, and Image Processing*, 52(2):191–209, 1990.

[62] Marta C Gonzalez, Cesar A Hidalgo, and Albert-Laszlo Barabasi. Understanding individual human mobility patterns. *Nature*, 453(7196):779–782, 2008.

[63] Hans Jørgen G Gundersen. Notes on the estimation of the numerical density of arbitrary profiles: The edge effect. *Journal of Microscopy*, 111(2):219–223, 1977.

[64] Diansheng Guo, Xi Zhu, Hai Jin, Peng Gao, and Clio Andris. Discovering spatial patterns in origin-destination mobility data. *Transactions in GIS*, 16(3):411–429, 2012.

[65] David J Hand. Principles of data mining. *Drug Safety*, 30(7):621–622, 2007.

[66] Robert M Haralick. Statistical and structural approaches to texture. *Proceedings of the IEEE*, 67(5):786–804, 1979.

[67] Robert M Haralick, Karthikeyan Shanmugam, and Its' Hak Dinstein. Textural features for image classification. *IEEE Transactions on Systems, Man, and Cybernetics*, (6):610–621, 1973.

[68] Charles R Harris, K Jarrod Millman, Stéfan J van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E Oliphant. Array programming with NumPy. *Nature*, 585(7825):357–362, 2020.

[69] John A Hartigan and Manchek A Wong. Algorithm AS 136: A K-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):100–108, 1979.

[70] Dong-Chen He, Li Wang, and Jean Guibert. Texture feature extraction. *Pattern Recognition Letters*, 6(4):269–273, 1987.

[71] Harold D Head and Marion F Brown. Preoperative vein mapping for coronary artery bypass operations. *The Annals of Thoracic Surgery*, 59(1):144–148, 1995.

[72] Edward Helderop and Tony H Grubesic. Streets, storm surge, and the frailty of urban transport systems: A grid-based approach for identifying informal street network connections to facilitate mobility. *Transportation Research Part D: Transport and Environment*, 77:337–351, 2019.

[73] Norbert Henze and B Zirkler. A class of invariant consistent tests for multivariate normality. *Communications in Statistics-Theory and Methods*, 19(10):3595–3617, 1990.

[74] Elke Hermans, Filip Van den Bossche, and Geert Wets. Combining road safety information in a performance index. *Accident Analysis & Prevention*, 40(4):1337–1344, 2008.

[75] David Hestenes. The design of linear algebra and geometry. *Acta Applicandae Mathematica*, 23:65–93, 1991.

[76] Robert J. Hijmans. *Raster: Geographic Data Analysis and Modeling*, 2023. R package version 3.6-20.

[77] Daniel S Hirschberg. Algorithms for the longest common subsequence problem. *Journal of the ACM*, 24(4):664–675, 1977.

[78] Peter D Hoff, Adrian E Raftery, and Mark S Handcock. Latent space approaches to social network analysis. *Journal of the American Statistical Association*, 97(460):1090–1098, 2002.

[79] Shin-Yi Hsu. Texture-tone analysis for automated land-use mapping. *Photogrammetric Engineering and Remote Sensing*, 44(11):1393–1404, 1978.

[80] Anne Humeau-Heurtier. Texture feature extraction methods: A survey. *Ieee Access*, 7:8975–9000, 2019.

[81] Riccardo Iacobucci, Benjamin McLellan, and Tetsuo Tezuka. Modeling shared autonomous electric vehicles: Potential for transport and power grid integration. *Energy*, 158:148–163, 2018.

[82] Martin Jacobsen and Joseph Gani. Point process theory and applications: Marked point and piecewise deterministic processes. 2006.

[83] Ben K. Jang and Roland T Chin. One-pass parallel thinning: Analysis, properties, and quantitative evaluation. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 14(11):1129–1140, 1992.

[84] Rik D. T. Janssen and Albert M. Vossepoel. Adaptive vectorization of line drawing images. *Computer Vision and Image Understanding*, 65(1):38–56, 1997.

[85] Yanjie Ji, Xinwei Ma, Mingyuan Yang, Yuchuan Jin, and Liangpeng Gao. Exploring spatially varying influences on metro-bikeshare transfer: A geographically weighted Poisson regression approach. *Sustainability*, 10(5), 2018.

[86] Álvaro Barbero Jiménez, Jorge López Lázaro, and José R Dorronsoro. Finding optimal model parameters by discrete grid search. In *Innovations in Hybrid Intelligent Systems*, pages 120–127. Springer, 2008.

[87] Yu Jin, Rui Peng, and Junping Shi. Population dynamics in river networks. *Journal of Nonlinear Science*, 29:2501–2545, 2019.

[88] Bela Julesz. Visual pattern discrimination. *IRE transactions on Information Theory*, 8(2):84–92, 1962.

[89] Ilsuk Kang, Cheolwoo Park, Young Joo Yoon, Changyi Park, Soon-Sun Kwon, and Hosik Choi. Classification of histogram-valued data with support histogram machines. *Journal of Applied Statistics*, 50(3):675–690, 2023.

[90] Shahid Karim, Ye Zhang, Muhammad Rizwan Asif, and Saad Ali. Comparative analysis of feature extraction methods in satellite imagery. *Journal of Applied Remote Sensing*, 11(4):042618, 2017.

[91] Stefan Karpinski, Elizabeth M Belding, Kevin C Almeroth, and John R Gilbert. Linear representation of network traffic: With special application to wireless workload generation. *Mobile Networks and Applications*, 14:368–386, 2009.

[92] Julia E Kelsall and Peter J Diggle. Kernel estimation of relative risk. *Bernoulli*, pages 3–16, 1995.

[93] Audie K Kilfoyle, Robert F Jermain, Manhar R Dhanak, Joseph P Huston, and Richard E Spieler. Effects of EMF emissions from undersea electric cables on coral reef fish. *Bioelectromagnetics*, 39(1):35–52, 2018.

[94] Rene Kirsten and Inger Fabris-Rotelli. A generic test for the similarity of spatial data. *South African Statistical Journal*, 55(1):55–71, 2021.

[95] Takumi Kobayashi. Dirichlet-based histogram feature transform for image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3278–3285, 2014.

[96] Trupti M Kodinariya, Prashant R Makwana, et al. Review on determining number of clusters in K-means clustering. *International Journal*, 1(6):90–95, 2013.

[97] Samuel Kotz, Narayanaswamy Balakrishnan, Campbell B Read, and Brani Vidakovic. *Encyclopedia of Statistical Sciences, Volume 1*. John Wiley & Sons, 2005.

[98] Danai Koutra, Neil Shah, Joshua T Vogelstein, Brian Gallagher, and Christos Faloutsos. Deltacon: Principled massive-graph similarity function with attribution. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 10(3):1–43, 2016.

[99] Vera Kuklina, Irina Bilichenko, Viktor Bogdanov, Dmitrii Kobylkin, Andrey N Petrov, and Nikolay Shiklomanov. Informal road networks and sustainability of Siberian boreal forest landscapes: Case study of the Vershina Khandy taiga. *Environmental Research Letters*, 16(11):115001, 2021.

[100] Vera Kuklina, Andrey N Petrov, Natalia Krasnoshtanova, and Viktor Bogdanov. Mobilizing benefit-sharing through transportation infrastructure: Informal roads, extractive industries and benefit-sharing in the Irkutsk oil and gas region, Russia. *Resources*, 9(3):21, 2020.

[101] Vera Kuklina, Andrey N Petrov, Nikolay I Shiklomanov, Qin Yu, and Victor Bogdanov. Investigating land use changes with development of the informal road networks: Case-study in Siberia. In *American Geophysical Union, Fall Meeting 2019*, 2019.

[102] Tarun Kumar and Karun Verma. A theory based on conversion of RGB image to gray image. *International Journal of Computer Applications*, 7(2):7–10, 2010.

[103] Hirohito Kuse, Akira Endo, and Eiichiro Iwao. Logistics facility, road network and district planning: Establishing comprehensive planning for city logistics. *Procedia-Social and Behavioral Sciences*, 2(3):6251–6263, 2010.

[104] Olli Lahdenoja, Jonne Poikonen, and Mika Laiho. Towards understanding the formation of uniform local binary patterns. *International Scholarly Research Notices*, 2013, 2013.

[105] Minjin Lee, Hugo Barbosa, Hyejin Youn, Petter Holme, and Gourab Ghoshal. Morphology of travel routes and the organization of cities. *Nature Communications*, 8(1):1–10, 2017.

[106] Martin D Levine and Ahmed M Nazif. Dynamic measurement of computer-generated image segmentations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (2):155–164, 1985.

[107] Pengbo Li, Haowen Yan, and Xiaomin Lu. A Siamese neural network for learning the similarity metrics of linear features. *International Journal of Geographical Information Science*, pages 1–28, 2022.

[108] Rui Li and Xiaoyu Zhang. Research on the improvement of EPTA parallel thinning algorithm. In *2018 International Conference on Network, Communication, Computer Engineering (NCCE 2018)*, pages 994–1001. Atlantis Press, 2018.

[109] Wenda Li, Michael Burrow, and Zijun Li. Automatic road condition assessment by using point laser sensor. In *2018 IEEE SENSORS*, pages 1–4. IEEE, 2018.

[110] Xiucheng Li, Kaiqi Zhao, Gao Cong, Christian S Jensen, and Wei Wei. Deep representation learning for trajectory similarity computation. In *2018 IEEE 34th International Conference on Data Engineering*, pages 617–628. IEEE, 2018.

[111] Qun Liu, Zhishan Dong, and En Wang. Cut based method for comparing complex networks. *Scientific Reports*, 8(1):1–11, 2018.

[112] Xiuwen Liu and DeLiang Wang. Texture classification using spectral histograms. *IEEE Transactions on Image Processing*, 12(6):661–670, 2003.

[113] Ze Liu and Shichen Zhao. Characteristics of road network forms in historic districts of Japan. *Frontiers of Architectural Research*, 4, 10 2015.

[114] Ignacio Loor and James Evans. Understanding the value and vulnerability of informal infrastructures: Footpaths in Quito. *Journal of Transport Geography*, 94:103112, 2021.

[115] Danilo Lopes and Renato Assunção. Visualizing marked spatial and origin-destination point patterns with dynamically linked windows. *Journal of Computational and Graphical Statistics*, 21(1):134–154, 2012.

[116] Xin Lu, Linus Bengtsson, and Petter Holme. Predictability of population displacement after the 2010 Haiti earthquake. *Proceedings of the National Academy of Sciences*, 109(29):11576–11581, 2012.

[117] Kang-Rae Ma and David Banister. Urban spatial change and excess commuting. *Environment and Planning A*, 39(3):630–646, 2007.

[118] Christine Mady et al. Exploring Beirut's instability through its informal mobility. *Urbani Izziv*, 32(Supp.):23–36, 2021.

[119] Martin Maechler, Peter Rousseeuw, Anja Struyf, Mia Hubert, and Kurt Hornik. *Cluster: Cluster Analysis Basics and Extensions*, 2022.

[120] Thomas L Magnanti and Prakash Mirchandani. Shortest paths, single origin-destination network design, and associated polyhedra. *Networks*, 23(2):103–121, 1993.

[121] Google Maps. Aerial view of the mobility study area, Kunene region in northwestern Namibia. https://www.google.co.za/maps/place/Kunene+Region,+Namibia/@-19.0571271,11.5764264, 2023.

[122] Stephen Marshall. Route structure analysis. In *Strasbourg: European Transport Conference*, 2003.

[123] Stephen Marshall. Line structure representation for road network analysis. *Journal of Transport and Land Use*, 9(1):29–64, 2016.

[124] Stephen Marshall, Jorge Gil, Karl Kropf, Martin Tomko, and Lucas Figueiredo. Street network studies: From networks to models and their representations. *Networks and Spatial Economics*, 18:735–749, 2018.

[125] Andrzej Materka and Michal Strzelecki. Texture analysis methods: A review. *Technical University of Lodz, Institute of Electronics, COST B11 Report, Brussels*, 10(1.97):4968, 1998.

[126] Jorge Mateu, Mehdi Moradi, and Ottmar Cronie. Spatio-temporal point patterns on linear networks: Pseudo-separable intensity estimation. *Spatial Statistics*, 37:100400, 2020.

[127] Greg McSwiggan, Adrian Baddeley, and Gopalan Nair. Estimation of relative risk for events on a linear network. *Statistics and Computing*, 30(2):469–484, 2020.

[128] Artis Mednis, Atis Elsts, and Leo Selavo. Embedded solution for road condition monitoring using vehicular sensor networks. In *2012 6th International Conference on Application of Information and Communication Technologies (AICT)*, pages 1–5. IEEE, 2012.

[129] Caio César Teodoro Mendes, Vincent Frémont, and Denis Fernando Wolf. Exploiting fully convolutional neural networks for fast road detection. In *2016 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3174–3179. IEEE, 2016.

[130] Hannah R Meredith, John R Giles, Javier Perez-Saez, Théophile Mande, Andrea Rinaldo, Simon Mutembo, Elliot N Kabalo, Kabondo Makungo, Caroline O Buckee, and Andrew J Tatem. Characterizing human mobility patterns in rural settings of sub-Saharan Africa. *Elife*, 10:e68441, 2021.

[131] Harvey J Miller. GIS and geometric representation in facility location problems. *International Journal of Geographical Information Systems*, 10(7):791–816, 1996.

[132] Volodymyr Mnih and Geoffrey E Hinton. Learning to detect roads in high-resolution aerial images. In *European Conference on Computer Vision*, pages 210–223. Springer, 2010.

[133] Jacob Modiba, Inger Fabris-Rotelli, Alfred Stein, and Gregory Breetzke. Linear hotspot detection for a point pattern in the vicinity of a linear network. *Spatial Statistics*, 51:100693, 2022.

[134] Mehdi Mokhtarzade and MJ Valadan Zoej. Road detection from high-resolution satellite images using artificial neural networks. *International Journal of Applied Earth Observation and Geoinformation*, 9(1):32–40, 2007.

[135] Julia S Mollee, Eric FM Araújo, and Michel CA Klein. Exploring parameter tuning for analysis and optimization of a computational model. In *Advances in Artificial Intelligence: From Theory to Practice: 30th International Conference on Industrial Engineering and Other Applications of Applied Intelligent Systems, IEA/AIE 2017, Arras, France, June 27-30, 2017, Proceedings, Part II 30*, pages 341–352. Springer, 2017.

[136] Jesper Møller and Rasmus P Waagepetersen. Modern statistics for spatial point processes. *Scandinavian Journal of Statistics*, 34(4):643–684, 2007.

[137] Sandra Morales, Kjersti Engan, Valery Naranjo, and Adrian Colomer. Retinal disease screening through local binary patterns. *IEEE Journal of Biomedical and Health Informatics*, 21(1):184–192, 2017.

[138] Meinard Müller. Dynamic time warping. *Information Retrieval for Music and Motion*, pages 69–84, 2007.

[139] Elizabeth D Mynatt, Annette Adler, Mizuko Ito, and Vicki L O'Day. Design for network communities. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*, pages 210–217, 1997.

[140] Tomoki Nakaya, Alexander S Fotheringham, Chris Brunsdon, and Martin Charlton. Geographically weighted Poisson regression for disease association mapping. *Statistics in Medicine*, 24(17):2695–2717, 2005.

[141] Mohamed Naouai, Melki Narjess, and Atef Hamouda. Line extraction algorithm based on image vectorization. In *2010 IEEE International Conference on Mechatronics and Automation*, pages 470–476. IEEE, 2010.

[142] Randy J Nelson. Seasonal immune function and sickness responses. *Trends in Immunology*, 25(4):187–192, 2004.

[143] Jaroslav Nešetřil and Svatopluk Poljak. On the complexity of the subgraph problem. *Commentationes Mathematicae Universitatis Carolinae*, 26(2):415–419, 1985.

[144] Mark EJ Newman and Michelle Girvan. Finding and evaluating community structure in networks. *Physical Review E*, 69(2):026113, 2004.

[145] Takao Nishizeki and Norishige Chiba. *Planar graphs: Theory and algorithms*. North-Holland Mathematics Studies, Vol 40, North-Holland Publishing Co, 1988.

[146] RAA Nobrega, CG O'Hara, and JA Quintanilha. Detecting roads in informal settlements surrounding Sao Paulo City by using object-based classification. *Proceedings of the 1st International Conference on Object-based Image Analysis (OBIA 2006), Salzburg, Austria*, pages 4–5, 2006.

[147] Dicky Nofriansyah and Hendriktio Freizello. Python application: Visual approach of Hopfield discrete method for hiragana images recognition. *Bulletin of Electrical Engineering and Informatics*, 7(4):609–614, 2018.

[148] Robert L Ogniewicz and Markus Ilg. Voronoi skeletons: Theory and applications. In *CVPR*, volume 92, pages 63–69, 1992.

[149] Timo Ojala, Matti Pietikäinen, and David Harwood. A comparative study of texture measures with classification based on featured distributions. *Pattern Recognition*, 29(1):51–59, 1996.

[150] Timo Ojala, Matti Pietikainen, and Topi Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(7):971–987, 2002.

[151] Juan De Dios Ortúzar, Jimmy Armoogum, Jean-Loup Madre, and Françoise Potier. Continuous mobility surveys: The state of practice. *Transport Reviews*, 31(3):293–312, 2011.

[152] David O'Sullivan. Spatial network analysis. In *Handbook of Regional Science*, pages 1253–1273. Springer, 2014.

[153] Nobuyuki Otsu. A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man, and Cybernetics*, 9(1):62–66, 1979.

[154] Francesca Pagliara, Filomena Mauriello, and Yin Ping. Analyzing the impact of high-speed rail on tourism with parametric and non-parametric methods: The case study of China. *Sustainability*, 13(6):3416, 2021.

[155] Luca Pappalardo, Filippo Simini, Salvatore Rinzivillo, Dino Pedreschi, Fosca Giannotti, and Albert-László Barabási. Returners and explorers dichotomy in human mobility. *Nature Communications*, 6(1):1–8, 2015.

[156] George Paschos and Maria Petrou. Histogram ratio features for color texture classification. *Pattern Recognition Letters*, 24(1-3):309–314, 2003.

[157] Edzer J. Pebesma and Roger S. Bivand. Classes and methods for spatial data in R. *R News*, 5(2):9–13, November 2005.

[158] Pietari Peltonen. Impacts of traffic environment, weather, road conditions and maintenance on walking and cycling travel. Master's thesis, Aalto University School of Engineering, Otaniemi, Finland, 2018.

[159] Luiz Pessoa. Understanding brain networks and brain organization. *Physics of Life Reviews*, 11(3):400–435, 2014.

[160] Shijin P.S and Dharun Vs. Extraction of texture features using GLCM and shape features using connected regions. *International Journal of Engineering and Technology*, 8:2926–2930, 12 2016.

[161] John Pucher and John L Renne. Urban-rural differences in mobility and mode choice: Evidence from the 2001 NHTS. *Bloustein School of Planning and Public Policy, Rutgers University*, pages 1–22, 2004.

[162] Xinyu Que, Fabio Checconi, Fabrizio Petrini, and John A Gunnels. Scalable community detection with the Louvain algorithm. In *2015 IEEE International Parallel and Distributed Processing Symposium*, pages 28–37. IEEE, 2015.

[163] Jagdish Lal Raheja, Sunil Kumar, and Ankit Chaudhary. Fabric defect detection based on GLCM and Gabor filter: A comparison. *Optik*, 124(23):6469–6474, 2013.

[164] Jean-Yves Ramel, Nicole Vincent, and Hubert Emptoz. A structural representation for understanding line-drawing images. *International Journal on Document Analysis and Recognition*, 3:58–66, 2000.

[165] Ayushman Ramola, Amit Kumar Shakya, and Dai Van Pham. Study of statistical methods for texture analysis and their modern evolutions. *Engineering Reports*, 2(4):e12149, 2020.

[166] Chiara Renso, Stefano Spaccapietra, and Esteban Zimnyi. Mobility data: Modeling, management, and understanding. In *Mobility Data*, 2013.

[167] Erdos P Renyi A. On random graph. *Publicationes Mathematicate*, 6:290–297, 1959.

[168] Fred S Roberts. *Graph Theory and its Applications to Problems of Society*. SIAM, 1978.

[169] Thomas K Rudel and Samuel Richards. Urbanization, roads, and rural population change in the Ecuadorian Andes. *Studies in Comparative International Development*, 25(3):73–89, 1990.

[170] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, 1986.

[171] John C Russ, James R Matey, A John Mallinckrodt, and Susan McKay. The image processing handbook. *Computers in Physics*, 8(2):177–178, 1994.

[172] Gert Sabidussi. The centrality index of a graph. *Psychometrika*, 31(4):581–603, 1966.

[173] Punam K Saha, Gunilla Borgefors, and Gabriella Sanniti de Baja. *Skeletonization: Theory, Methods and Applications*. Academic Press, 2017.

[174] Punam K Saha, Gunilla Borgefors, and Gabriella Sanniti di Baja. A survey on skeletonization algorithms and their applications. *Pattern Recognition Letters*, 76:3–12, 2016.

[175] Shunta Saito and Yoshimitsu Aoki. Building and road detection from large aerial imagery. In *Image Processing: Machine Vision Applications VIII*, volume 9405. International Society for Optics and Photonics, 2015.

[176] Oliver G Selfridge. Pattern recognition and modern computers. In *Proceedings of the March 1-3, 1955, Western Joint Computer Conference*, pages 91–93, 1955.

[177] Claude Elwood Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423, 1948.

[178] Poonam Sharma, Saloni Sharma, and Ayush Goyal. An MSE (mean square error) based analysis of deconvolution techniques used for deblurring/restoration of MRI and CT images. In *Proceedings of the Second International Conference on Information and Communication Technology for Competitive Strategies*, pages 1–5, 2016.

[179] Vladimir Shepelev, Sergei Aliukov, Kseniya Nikolskaya, and Salavat Shabiev. The capacity of the road network: Data collection and statistical analysis of traffic characteristics. *Energies*, 13(7):1765, 2020.

[180] Shinichiro Shirota, Alan E Gelfand, and Jorge Mateu. Analyzing car thefts and recoveries with connections to modeling origin–destination point patterns. *Spatial Statistics*, 38:100440, 2020.

[181] Minna Sikiö, Kirsi K Holli-Helenius, Pertti Ryymin, Prasun Dastidar, Hannu Eskola, Minna Sikiö, Lara Harrison, and Hannu Eskola. The effect of region of interest size on textural parameters. In *2015 9th International Symposium on Image and Signal Processing and Analysis (ISPA)*, pages 149–153, 2015.

[182] Stephen M Stahl. *Essential psychopharmacology: Neuroscientific basis and practical applications*. Cambridge University press, 2000.

[183] Dietrich Stoyan, Wilfrid S Kendall, Sung Nok Chiu, and Joseph Mecke. *Stochastic geometry and its applications*. John Wiley & Sons, 2013.

[184] Zhonghua Sun and Kebin Jia. Road surface condition classification based on color and texture information. In *2013 Ninth International Conference on Intelligent Information Hiding and Multimedia Signal Processing*, pages 137–140. IEEE, 2013.

[185] Edward James Taaffe. *Geography of Transportation*. Morton O'kelly, 1996.

[186] Hideyuki Tamura, Shunji Mori, and Takashi Yamawaki. Textural features corresponding to visual perception. *IEEE Transactions on Systems, Man, and Cybernetics*, 8(6):460–473, 1978.

[187] Mattia Tantardini, Francesca Ieva, Lucia Tajoli, and Carlo Piccardi. Comparing methods for comparing networks. *Scientific Reports*, 9(1):1–19, 2019.

[188] Peter Tarabek. A robust parallel thinning algorithm for pattern recognition. In *2012 7th IEEE International Symposium on Applied Computational Intelligence and Informatics (SACI)*, pages 75–79. IEEE, 2012.

[189] Renate Nicole Thiede, Inger Nicolette Fabris-Rotelli, Alfred Stein, Pravesh Debba, and M Li. Uncertainty quantification for the extraction of informal roads from remote sensing images of South Africa. *South African Geographical Journal*, 102(2):249–272, 2020.

[190] Kevin Toohey and Matt Duckham. Trajectory similarity measures. *Sigspatial Special*, 7(1):43–50, 2015.

[191] William Thomas Tutte and William Thomas Tutte. *Graph theory*, volume 21. Cambridge University Press, 2001.

[192] Stéfan van der Walt, Johannes L. Schönberger, Juan Nunez-Iglesias, François Boulogne, Joshua D. Warner, Neil Yager, Emmanuelle Gouillart, Tony Yu, and the scikit-image contributors. Scikit-image: Image processing in Python. *PeerJ*, 2:e453, 6 2014.

[193] JMJ Van Leeuwen, J Groeneveld, and J De Boer. New method for the calculation of the pair correlation function. *Physica*, 25(7-12):792–808, 1959.

[194] Marie-Colette NM van Lieshout. On estimation of the intensity function of a point process. *Methodology and Computing in Applied Probability*, 14:567–578, 2012.

[195] Peter R. van Nieuwenhuizen, Olaf Kiewiet, and Willem F. Bronsvoort. An integrated line tracking and vectorization algorithm. *Computer Graphics Forum*, 13(3):349–359.

[196] Jay M Ver Hoef. Kriging models for linear networks and non-Euclidean distances: Cautions and solutions. *Methods in Ecology and Evolution*, 9(6):1600–1613, 2018.

[197] Steven Verstockt, Viktor Slavkovikj, Pieterjan De Potter, Jürgen Slowack, and Rik Van de Walle. Multi-modal bike sensing for automatic geo-annotation of road/terrain type by participatory bike-sensing. In *2013 International Conference on Signal Processing and Multimedia Applications (SIGMAP)*, pages 39–49. IEEE, 2013.

[198] Matheus P Viana, Emanuele Strano, Patricia Bordin, and Marc Barthelemy. The simplicity of planar networks. *Scientific reports*, 3(1):3495, 2013.

[199] T Vigneshl and KK Thyagharajan. Local binary pattern texture feature for satellite imagery classification. In *2014 International Conference on Science Engineering and Management Research (ICSEMR)*, pages 1–6. IEEE, 2014.

[200] David Vlahov and Sandro Galea. Urbanization, urbanicity, and health. *Journal of Urban Health*, 79:S1–S12, 2002.

[201] Olivier J Walther, Lawali Dambo, Moustapha Koné, Michiel van Eupen, et al. Mapping travel time to assess accessibility in West Africa: The role of borders, checkpoints and road conditions. *Journal of Transport Geography*, 82(102590):10–1016, 2020.

[202] Jinzhong Wang, Xiangjie Kong, Feng Xia, and Lijun Sun. Urban human mobility: Data-driven modeling and prediction. *ACM SIGKDD Explorations Newsletter*, 21(1):1â19, may 2019.

[203] QJ Wang. Using genetic algorithms to optimise model parameters. *Environmental Modelling & Software*, 12(1):27–34, 1997.

[204] Richard Y. Wang and Diane M. Strong. Beyond accuracy: What data quality means to data consumers. *Journal of Management Information Systems*, 12(4), 1996.

[205] Weixing Wang, Nan Yang, Yi Zhang, Fengping Wang, Ting Cao, and Patrik Eklund. A review of road extraction from remote sensing images. *Journal of Traffic and Transportation Engineering*, 3(3):271–282, 2016.

[206] Wenjun Wang, Lin Pan, Ning Yuan, Sen Zhang, and Dong Liu. A comparative analysis of intra-city human mobility by taxi. *Physica A: Statistical Mechanics and its Applications*, 420:134–147, 2015.

[207] Zhou Wang and Alan C Bovik. A universal image quality index. *IEEE signal processing letters*, 9(3):81–84, 2002.

[208] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE transactions on Image Processing*, 13(4):600–612, 2004.

[209] Duncan J Watts and Steven H Strogatz. Collective dynamics of small-world networks. *Nature*, 393(6684):440–442, 1998.

[210] Stanisław Węglarczyk. Kernel density estimation and its application. In *ITM Web of Conferences*, volume 23, page 00037. EDP Sciences, 2018.

[211] Joan S Weszka, Charles R Dyer, and Azriel Rosenfeld. A comparative study of texture measures for terrain classification. *IEEE transactions on Systems, Man, and Cybernetics*, (4):269–285, 1976.

[212] Harrison C White. Chains of opportunity. In *Chains of Opportunity*. Harvard University Press, 2013.

[213] Alan Geoffrey Wilson. The use of entropy maximising models, in the theory of trip distribution, mode split and route split. *Journal of Transport Economics and Policy*, pages 108–126, 1969.

[214] Feng Xia, Li Liu, Behrouz Jedari, and Sajal K Das. PIS: A multi-dimensional routing protocol for socially-aware networking. *IEEE Transactions on Mobile Computing*, 15(11):2825–2836, 2016.

[215] Feng Xie and David Levinson. Measuring the structure of road networks. *Geographical Analysis*, 39(3):336–356, 2007.

[216] Weixiang Xu and Rongxin Gao. Prediction of road conditions ahead based on travel plans. In *2019 IEEE 5th International Conference on Computer and Communications (ICCC)*, pages 262–266. IEEE, 2019.

[217] Chunsun Zhang et al. An UAV-based photogrammetric mapping system for road condition assessment. *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci*, 37:627–632, 2008.

[218] Kaiwen Zhang, Shuozhong Wang, and Xinpeng Zhang. New metric for quality assessment of digital images based on weighted mean square error. In *Second International Conference on Image and Graphics*, volume 4875, pages 491–497. SPIE, 2002.

[219] Lei Zhang. *Search, Information, Learning, and Knowledge in Travel Decision-making: A Positive Approach for Travel Behavior and Demand Analysis.* PhD thesis, University of Minnesota, 2006.

[220] Shichao Zhang. Cost-sensitive k-NN classification. *Neurocomputing*, 391:234–242, 2020.

[221] Tongjie Y Zhang and Ching Y. Suen. A fast parallel algorithm for thinning digital patterns. *Communications of the ACM*, 27(3):236–239, 1984.

[222] George Kingsley Zipf. The P_1P_2/D hypothesis: On the intercity movement of persons. *American Sociological Review*, 11(6):677–686, 1946.

[223] Pedro J Zufiria, David Pastor-Escuredo, Luis Úbeda-Medina, Miguel A Hernandez-Medina, Iker Barriales-Valbuena, Alfredo J Morales, Damien C Jacques, Wilfred Nkwambi, M Bamba Diop, John Quinn, et al. Identifying seasonal mobility profiles from anonymized and aggregated mobile phone data. application in food security. *PloS one*, 13(4):e0195714, 2018.

[224] Alex Zvoleff. *GLCM: Calculate Textures from Grey-Level Co-Occurrence Matrices (GLCMs)*, 2020. R package version 1.6.5.