MSC. GIS AND REMOTE SENSING RESEARCH PROJECT

Bicycle path planning in Johannesburg: Aggregating user-defined spatial criteria to create efficient routes for bicycle infrastructure.

University of the Witwatersrand
Student Number: 1396945
Author: Spencer M. Johnson
Supervisor: Dr. Stefania Merlo

Spencer M. Johnson

A Master's research project submitted to the Faculty of Science, University of the Witwatersrand, in partial fulfilment of the requirements for the degree of Master of Science in GIS and Remote Sensing.

Johannesburg, 2017

DECLARATION

I declare that this Research Report is my own, unaided work. It is being submitted for the Degree of Master of Science at the University of the Witwatersrand, Johannesburg. It has not been submitted before for any degree or examination at any other University.

______________________________ **(Signature of candidate)**

20th Day of August, 2017

ABSTRACT

Recent initiatives by the city of Johannesburg to increase non-motorised transport through the installation of bicycle infrastructure were conducted without consulting the cycling preferences of the public. This study distributed a cycling preference survey, achieving fair reliability using the weighted Kappa statistic, in which potential users indicated the most important spatial factors for ideal cycling routes through Likert-scale answers. Importance rankings derived by Likert sums were combined with variability-explaining rankings derived by modified principal component analysis using polychoric correlation coefficients to produce a final list of retained spatial variables. These variables were quantified using secondary spatial data sets which were dichotomized into Boolean operators for network attributes in ArcGIS Network Analyst. The solved routes using the spatial factors derived by survey respondents were significantly different from the simple shortest-path routes between pre-defined origin and destination nodes. Shortcomings in the directness of the solved routes qualify their use as an initial step for non-motorised transport planning rather than a strict, unmodifiable route for bicycle lanes. Further experimentation with higher quality spatial data, custom routing algorithms, and a larger survey population may yield improved results in the future. The incorporation of local cyclists and future cyclists are a key factor in bicycle route design that should be included in non-motorised transport planning.

List of Acronyms

BCI – Bicycle Compatibility Index, developed by Harkey *et al.*, 1998
ESRI – Environmental Systems Research Institute, makers of ArcGIS for Desktop
EXCEL – Microsoft Excel Software
GCRO – Gauteng City-Region Observatory
GIS – Geographic Information System
GPDRT – Gauteng Province Department of Roads and Transport
JMPD – Johannesburg Municipal Police Department
JRA – Johannesburg Roads Agency
JUCA – Johannesburg Urban Cyclists Association
NGI – National Geospatial Information (of South Africa)
R – R Statistical Software

Figures

Tables

Equation

1. Introduction

Over the past seven years, the City of Johannesburg (Joburg) has allocated significant resources for developing and designing bicycle infrastructure (City of Joburg, 2009; *Making Joburg a Cycle Friendly City*, 2014). Dedicated bicycle lanes were installed on streets in neighbourhoods including Braamfontein, Soweto, and Brixton, and numerous cycling promotion campaigns were implemented by the City of Joburg, Johannesburg Urban Cyclist Association (JUCA), and Cycle Wits. In recent iterations of bicycle infrastructure development, the Johannesburg Roads Agency (JRA) hired an engineering consulting firm and provided them with a transportation corridor or even a proposed route between two or more neighbourhoods. The firm reviewed local and national transport policies, collected detailed data in the route area (e.g. traffic counts), refined the routes based on various factors, and completed a detailed design of where and how the cycle routes would be implemented (GIBB, 2015).

The factors used in these feasibility and design reports were identified and prioritized primarily by the direction and objectives set out by the JRA, coupled with the discretion and knowledge of the consulting firm, who has experience in designing cycle infrastructure and routes in other cities such as Cape Town (GIBB, 2015). One potential drawback of this method is that the actual users and future users of this cycle infrastructure were not consulted about what factors should or should not be included, and which are the most important.

City and neighbourhood route environments, as well as locals' perceptions of cycling, may significantly differ from place to place (Wahlgren and Schantz, 2011). A bicycle route model employed in one city may not necessarily fit in another. Johannesburg has unique features for a large metropolitan area. Heavy topographic relief, significant crime, and a sprawling, disconnected design shaped by decades of Apartheid rule all may impact the choices of potential cyclists (Todes, 2012; Stones, 2013; *Making Joburg a Cycle Friendly City*, 2014). Casual observation of constructed cycle paths in the city implies low use by cyclists, as well as common misuse in the form of pedestrians, parked automobiles, collective taxis, and waste trucks (Boye, 2015). A complimentary approach to cycle infrastructure design is to implement a survey which asks local users for input on cycling route criteria, a popular method present in cycle planning studies and literature (Winters *et al.*, 2010).

1.1 Research Problem

Recent efforts by the City of Johannesburg to increase the modal share of bicycling for commutes experienced mixed results, with positive PR campaigns tempered by a lack of actual cyclists using newly-constructed routes (Boye, 2015; Huchzermeyer, 2016). By surveying current and potential users, spatial criteria for cycle infrastructure route planning can be ranked and aggregated to produce more effective bicycle route planning results. These criteria can be inserted into a multi-objective shortest path algorithm to identify ideal routes on which cycling infrastructure can be installed, ultimately guiding an urban planner in designing popular, usable routes in the City of Johannesburg.

1.2 Aim and Objectives

This study specifically seeks to find the best routes for bicycle infrastructure, as opposed to identifying the best routes for cyclists to ride on the road, a project already completed by JUCA. Priority in the study is given to current and potential future cyclists in terms of identifying criteria. In practical applications, cost and political factors inevitably play an important role, but this study focuses on a theoretical framework to identify routes solely with the cyclists' interests in mind. The study area of the Johannesburg road network was selected and further refined using survey results as described in Section 3.2.

The overall aim of this study is to plan routes for cycle infrastructure based on user-defined factors and locations. The objectives to complete this aim are as follows:

1. Identify and rank spatial criteria and origin nodes that dictate route choice for local cyclists and potential cyclists based on survey results.
2. Combine spatial criteria to create multiple distinct factors or objectives that will be used in conjunction with the simple shortest distance path to identify efficient bicycle infrastructure routes.
3. Identify efficient bicycle path routes between origin and destination nodes that reflect where the survey respondents live and work.
4. Assess whether the planned routes differ significantly from the shortest distance path.

1.3 Key Terms and Concepts

This study focuses on pathfinding for *bicycle route infrastructure*; that is identifying a path or route where bicycle infrastructure should be installed. For the purposes of this study, *bicycle infrastructure* is defined as a constructed path running parallel to the street, or simply a painted lane marked at the edges of the roadway, such as a bicycle lane. *Bicycle* and *cycle* and the corresponding gerund verb forms are used interchangeably.

The term *road network* defines the network of streets, roads, alleyways, and informal paths (i.e. any linear area that bike infrastructure could be installed on) within the study area. *Intersections* are the nodes or points connecting each individual link or segment within the road network, and can take the form of a simple unmarked junction, a traffic circle, an intersection controlled by traffic lights, or many others.

The terms *factor*, *criteria, and variable* are used interchangeably in the literature and this study. These words both refer to the various aspects which can affect a cyclist's choice of a perceived ideal route from one point to another. Distance, weather, air pollution, and crime are examples, and the word *spatial* may precede these words to further qualify their geographic nature.

When referring to participants of the survey, it is vital to distinguish between *current* and *potential* cyclists. These are two distinct groups of people that may or may not have different priorities when it comes to route choice factors. As explained in Section 3.3, while *potential* users may not currently cycle often, if at all, the study hypothesizes that the installation of cycle infrastructure on routes they prefer would increase the frequency of cycling for this group, as suggested in the literature (Goodman *et al*, 2013; Rissel *et al*, 2015; Sahlqvist *et al*, 2015). Therefore, both *potential* and *current* cyclists – those who already cycle regularly – are equally represented in the resulting ranked criteria list from the survey.

2. Literature Review

A total of 60 published academic journal papers, along with some text books, news articles, and government and consulting firm presentations and reports were reviewed and the salient points are considered below.

2.1 Historical Background

Bicycle route planning has been studied and implemented as early as the 1950s in the Netherlands (McClintock, 2002), and the 1970s in the United Kingdom (Hudson, 1978) and the U.S.A (Pucher *et al*., 1999). These early studies were generally at odds with urban planning dogmas dominated by the passenger car until a relative boom in cycle infrastructure planning and implementation in the mid-1990s (McClintock, 2002). These planning patterns differed geographically, but studies concerning cycle infrastructure and planning have generally increased since the nineties, and there is a vast catalogue of academic studies to consider even from the past 5 years alone (Buehler and Dill, 2016).

While Bil *et al.* (2012) tackled a GIS database of cycle infrastructure for the whole of the Czech Republic, studies generally focused on the neighbourhood or city scale (Klobucar and Fricker, 2007; Larsen *et al.*, 2013), and have covered cities across the globe such as Cardiff, U.K. (Sahlqvist *et al*., 2015), Berkeley, U.S.A. (Huang, 1995), and Auckland, New Zealand (Wang *et al*., 2014). There is, however, a significant lack of cycle planning research on cities in South Africa and the African continent.

2.2 Importance of Cycle Infrastructure Planning

Researchers consistently point out the improved sustainability of cycling as a mode of transport over the automobile (Wang *et al*., 2014; Buehler and Dill, 2016), and the opportunity for regular physical activity in an increasingly unhealthy and sedentary society as reasons for the importance of cycling (Winters *et al*., 2010; Wahlgren and Shantz, 2011; Wong *et al*., 2011; Duthie and Unnikrishnan, 2014; Krenn *et al*., 2014; Dalton *et al*., 2015). Kamargianni (2015) noted the missed opportunities in many cities where short trips that would take less than 20 minutes by bike were still being completed by motorized vehicles. Tying in to sustainability mentioned above, Huang and Ye (1995) cited over two decades ago the challenges of heavy automobile traffic and poor air quality in cities; two major problems faced by Johannesburg today (City of Joburg, 2009).

If it is accepted that "cycling is one of the most energy efficient and healthy transport modes" (Ehrgott *et al.*, 2012, p. 652), then it is in the interest of local governments to provide efficient, well-connected cycling routes and infrastructure to promote cycling as a form of regular transport. Indeed, the City of Johannesburg has recognized the benefits of cycling mentioned above and has recently declared that "the city is committed to making walking, cycling and public transport the modes of choice for city residents" (*Making Joburg a Cycle Friendly City*, 2014, p. 9).

2.3 Determining Route Choice Criteria

A basic problem in determining cycling routes is figuring out what factors or criteria that affect where a cyclist would want to travel (Kamargianni, 2015). To identify the most important factors, numerous studies distribute surveys. Krenn *et al.* (2014) asked questions to current cyclists, while Akar and Clifton (2009), Caulfield *et al.* (2012) and Bawa (2015) included potential cyclists in their surveys, which is important when considering building infrastructure to encourage non-cyclists to become regular users. In some instances, studies combined all forms of non-motorized transport and interviewed pedestrians as well (Wigan *et al.*, 1998). Goodman *et al.* (2013) reduced sampling bias by drawing results from a general transportation survey that was open to the entire population of the study area. Surveys may have limitations with biased sampling techniques. By surveying only university students and staff, Wang *et al.* (2014) generated a sample the researchers admitted did not reflect the entire population of the study area.

Another method allows current cyclists to track their trips using the GPS antenna embedded in their smart phones (Hood and Charlton, 2011). While this method has privacy loss implications, Sainio *et al.* (2015) proposed a technique to protect user privacy by using aggregated heat maps of user routes. The resulting paths can be analysed to deduce what factors are being prioritized over others. Krenn *et al.* (2014) combined results from a transportation questionnaire, and then selected some respondents out of the survey to get fitted with GPS units to compare survey answers to actual routes taken.

2.4 Analysing survey responses and Likert-scale data

In deploying a survey, it is important to be able to assess the reliability of the responses (Sampat *et al.*, 2006). For the Likert scale "the notion of reproducibility is exact" and "data are not subject to random measurement errors" (Gwet, 2008, p.7). Therefore, it makes sense that Green (2007, p. 4), who says "If agreement is high, we feel more confident the ratings reflect the actual circumstance," uses the kappa statistic to measure intra-rater reliability because it simultaneously controls for chance agreements in Likert ratings. In this study, a modified Bland-Altman plot is used to visually assess survey reliability using methods from Rankin and Stokes (1998), while the kappa statistic is also calculated for a numerical measure.

Many different publications offered advice and general rules for analysing ordinal, Likert scale data, which applies to the focus of the cycling survey. Gob *et al.* (2007) listed analysis of empirical sums, means, variances, and correlation coefficients as good starting points for organizing Likert scale data, as well as using typical descriptive statistics. Flora and Curran (2004) asserted that the underlying test assumptions for confirmatory factor analysis and other parametric tests do not hold when applied to non-continuous data, but Norman (2010) disagreed, saying that just the means of the ordinal data needed to be normally distributed, not the data itself. Heeren and D'Agostino (1987) echoes this line of thinking, acknowledging that applying a t-test to ordinal data breaks the assumption of normality, yet through testing it is found that applying a t-test to ordinal data, even with a small sample size (n=10), is adequately robust. Gob (2007) mentions the lack of a common standard for analysing Likert and ordinal data within the statistics community, and advises that clear methods need to be chosen and justified.

Other more complex methods of analysing ordinal and Likert-scale data are reviewed in Refaat (2007), including using the Pearson chi-squared statistic in multi-category contingency tables to measure agreement between two sample sub-groups. Finally, Baglin (2014) explores using exploratory factor analysis on ordinal data with specialized software, and offers insight on how to interpret eigenvalues and scree plots when the results are not necessarily clear.

2.5 What data is used to define factors?

Consider a simple example: a vector shapefile of a street network is used to define distance, an obvious factor in cyclist route choice (Su *et al.*, 2010). But how can a more nebulous factor, such as bicyclist safety, be quantified? Yiannakoulias *et al.* (2012) mapped all traffic collisions on a road network, while Chandra (2014) used automobile speed limits and traffic density as proxies for safety because traffic accident data was not available. Interestingly, Zolnik and Cromley (2007) found that total descent (negative slope) of a route had the greatest effect on frequency of collisions and therefore safety. The services, amenities, and even aesthetic aspects along a route are often identified as important factors (Winters *et al.*, 2010; Buehler and Dill, 2016), and can be measured by using a route buffer and analysing what land use types or points of interest fall inside that buffer (Dalton *et al.* 2015).

Other studies combine many more 'objective' factors, such as road width, on-street parking, and cycle facilities (Ehrgott *et al.*, 2012) and aggregate them into a single suitability factor or Bicycle Compatibility Index (BCI) (Harkey *et al.*, 1998). This method has the advantage of combining and manipulating objective data in to quantitative indices (Ehrgott *et al.*, 2012). The factor of slope, or gradient, over any given distance is also common in bicycle route modelling, due to the extra energy the user must expend to cycle up steep hills (Buehler and Dill, 2016). Quantifying slope is often as easy as subtracting the elevation of two linking nodes and dividing by the distance of the link (Huang and Ye, 1995), but Scarf and Grehan (2005) developed a method to be more precise by assigning modified travel times for each meter of elevation gained.

Air pollution and health factors are driven by automobile traffic exposure, and one study aimed to "provide new information to enable bicycle network analysis with consideration of exposure risks" (Bigazzi and Figliozzi, 2015, p. 14). This study collected on-road concentration of pollution-indicator compounds (e.g., toluene), and then extrapolated those measures over the entire network to quantify cyclist exposure to air pollution. MacNaughton *et al.* (2014) used portable air monitoring equipment and rode the full routes of the study area to measure levels of carbon and nitrous dioxide along routes and confirmed their hypothesis that bike lanes had higher concentrations than separated bike paths. It is clear from the literature that the systematic quantifying of more objective data is one of the most complex and important parts of the route planning process.

Defining areas of relatively high crime is an exercise in geostatistical interpolation based on an input dataset of points (Eck *et al.*, 2005). When crime data is not available, proxies can be used, such as census data for neighbourhood quality, but this type of proxy may miss crucial street-scale variations for application to cycling routes (Cervero and Duncan, 2003). There are numerous methods and different types of data sets that can be used to define any given factor, based on the accessibility, accuracy, and scope of the data used.

2.6 Pathfinding Models and Algorithms

Dijkstra's shortest-path algorithm, which finds the shortest path from an origin node to a destination node in a network of links (Mamoulis, 2012), was originally designed in 1959. This algorithm can be modified and built upon based on the application (Möhring *et al.*, 2005), and can become more complex when additional objectives are added to the equation (Perederieieva *et al.*, 2013). Different studies create unique versions of bi- or multi-objective algorithms, often with the goal of minimizing computation time (Xiong and Wang, 2014; Duthie and Unnikrishnan, 2014), accommodating more objectives (Li *et al.*, 2013), or incorporating the effect of automobile traffic (Raith *et al.,* 2013).

Specific to bicycle route planning, Ehrgott *et al.* (2012) employed a bi-objective shortest path algorithm using travel time and suitability, while Zolnik and Cromley (2007) separately mapped out two objectives and compared them. Consulting for the City of Johannesburg, GIBB (2015) summed weighted values of fourteen factors in a matrix, including commuter preference, parking restraints, and cost implications, to determine cycling routes connecting Rosebank with other nearby neighbourhoods. The *Network Analysis* tool in ArcGIS was, employed in this study, makes use of Dijkstra's algorithm on the Johannesburg road network to determine the least-cost path one junction at a time between the selected origin and destination node. The costs are determined by the results and rankings of the survey, detailed in Section 4.

With practical studies conducted as early as the 1950's, bicycle and non-motorised route planning is an important facet of regional urban planning in many developed cities today. Route choice criteria, often determined by surveying local users, can be used to assess which routes bicyclists would prefer to take. Routes are then mapped using algorithms for least-cost paths.

3. Methodology

3.1 Research Design

This study follows a sequence of steps to complete the central aim. By collecting and analysing user-defined spatial factors, ideal routes for bicycling infrastructure are generated using the shortest-path algorithm built in to Network Analyst of ArcGIS 10.3. The research design is an opinion-based experimental method based on the quantified results of the distributed survey.

The user-defined spatial variables in the study are objective (e.g., route distance, traffic counts). Non-Objective factors (e.g., is the route aesthetically pleasing?) are not included as they are different based on perceptions and biases from each user, making it much more difficult to quantify and input into a route model (Harkey *et al.*, 1998). These non-objective types of factors were not included as criteria in this study to ensure factors could be quantified as accurately as possible with the data. Aggregating and quantifying the objective factors into comparable, relatable indices or weights to be used in the pathfinding algorithm is the central challenge of the project, as described in detail in Sections 3.5 and 3.6.

After origin and destination nodes are selected, the routes are created using the Network Analyst toolbox and the finalized criteria. The routes are then compared to that of a simple shortest path to assess the advantages and shortcomings of the final solved routes.

Software used in the data analysis and spatial visualizations of this research include Microsoft Excel 2016, SAS Enterprise Guide 7.1 and SAS 9.4 (SAS Institute, 2013), R 3.3.1 using the RStudio interface (R Core Team, 2016), and ArcGIS 10.3.1 (ESRI, 2015). For the principal component analysis in Section 4.5., the data was generated using the Real Statistics Resource Pack software (Release 4.3, Copyright (2013 – 2015) Charles Zaiontz), an add-in to Excel.

3.2 Study Area

The study area incorporates a road and path network that encompasses a specific area of Johannesburg, extending to certain neighbourhoods based on answers to the survey and the availability of pertinent crime data (Sections 3.3, 3.6.2). The final study area (Figure 1) aligns with the geographical distribution of survey respondents to provide routes that reflect the views of the users.

Survey respondents were asked for their location and answered with either a full street address, or more commonly, a four-digit postal code. These locations were inputted into a composite address locator and geocoded using ArcGIS to create points reflecting each respondent. Locations reflecting postal codes resulted in a point at the centroid of east postal district, while full street addresses generated more precise locations.

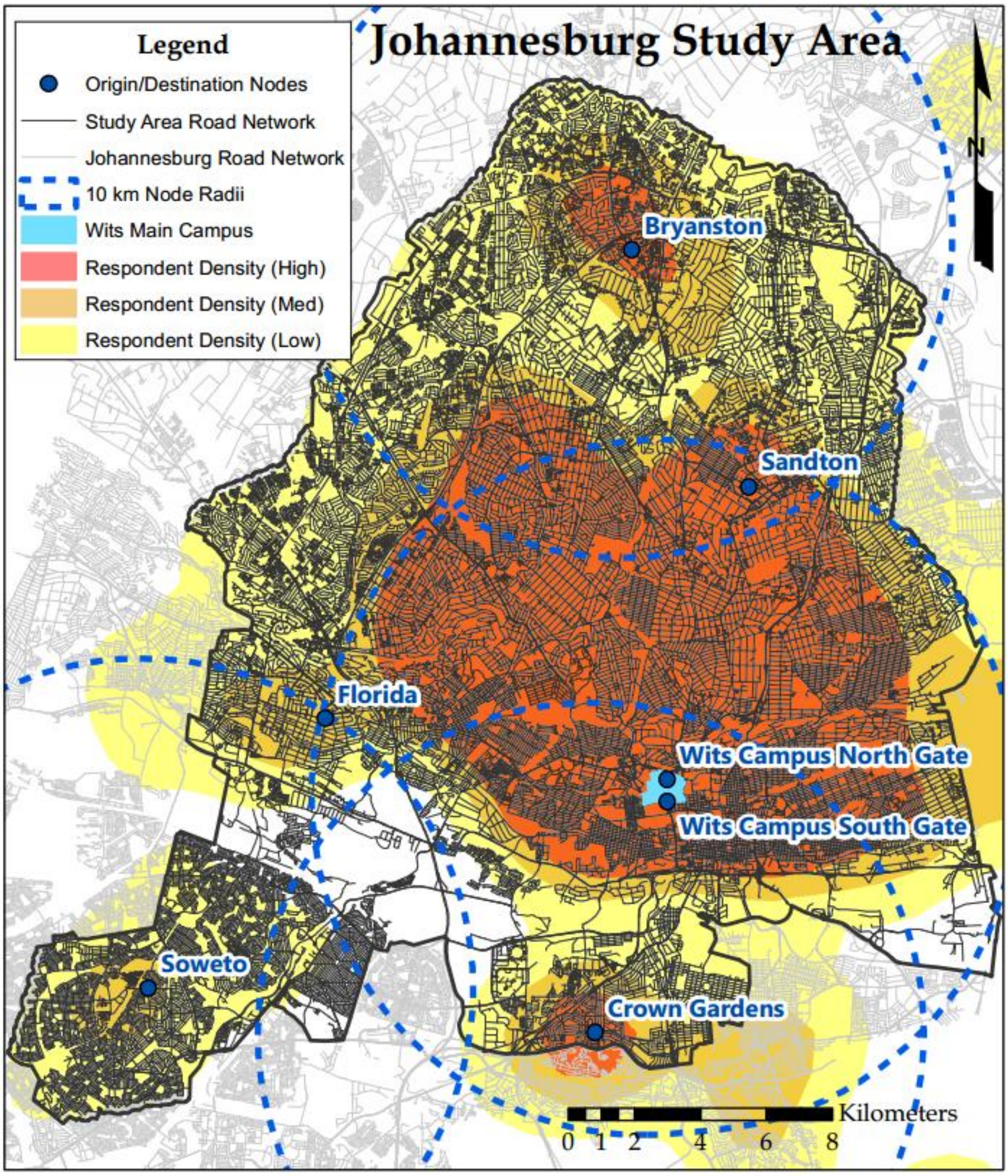


Figure 1 - Final study area in Johannesburg based on the availability of data, location of survey respondents, and comparison to the City of Johannesburg report and feasibility study (City of Joburg, 2009; Mohammed, 2009).

Using the *spatstat* package in R (Baddeley *et al.*, 2015), a point density field was calculated using the kernel density method and a bandwidth of $h = 0.2$. Lower h values (< 0.1) produced density contours with too much noise while higher h values (> 1) produced contours that were oversmoothed. The bandwidth of 0.1 was selected by visually assessing the capture of spatial differences on a neighbourhood scale within the Johannesburg study area. Then, four contours were drawn representing the boundaries between five percentiles of relative density for the survey respondent locations. These contours were extracted as polygons at the three highest levels of density (80^{th}-100th percentile, 60^{th}-80^{th} percentile, and 40^{th}-60^{th} percentile) using the 'gPolygonize' function within the *rgeos* package (Bivand and Rundel, 2016) to visualize the high-density areas seen in Figure 1. The centre of these high- and medium-density areas in orange became potential origin and destination nodes, along with the main Wits campus, for the final routing output using Network Analyst. Radii of 10 kilometres are included to serve as a guide for the approximate reasonable cycling distance between origin and destination nodes to be used for route finding (Harkey *et al.*, 1998). Edges of the lower density contour polygons (yellow) were used to clip the road network data to a smaller area, shrinking the study area to better reflect where survey respondents live.

The spatial limitations of data availability further focused the study area. The road network data provided by ESRI, including traffic, point of interest (POI) and land use data, was confined to the municipal boundaries of Johannesburg. This eliminated an area of high respondent density in the Edenvale/Sebenza neighbourhood to the east. Another limiting data set was the crime incidents from 2014. Five precincts where crime data was sourced for this study were eliminated from the study area for failing to reach a minimum geocoding match percentage of 50%. These precincts are highlighted in Figure 3 and Section 3.6.2 provides more discussion on the geocoding and normalization of raw crime incident data collected from Johannesburg precincts. Derivation of the final origin and destination nodes seen in Figure 1 is further explained in Section 3.7.

3.3 Cyclist Survey

The main objective of implementing a survey was to identify a finite list of factors which cyclists believe affect cycling route choices. The survey was designed to allow for a combination of current cyclists and potential cyclists to answer the questions. This is important for two reasons: it produces a larger sample size, as the population of current cyclists at Wits is comparatively small (Bawa, 2015), and it ensures that views of the 'potential' group are represented, assuming a significant number of potential users will be inspired to cycle more if adequate cycling infrastructure is installed.

After securing ethical clearance from the university, the cycling survey was implemented using an online survey service called Typeform (typeform.com), allowing access via PC, tablets, and mobile devices. The survey was distributed using the active Wits email catalogue for students, as well as a very limited number of staff members. The limited distribution to staff members was a result of university policy and could not be avoided. This resulted in a disproportionate amount of student to staff respondents (700:13). Additionally, the survey was distributed to members of the Johannesburg Urban Cyclists Association (JUCA) via their Facebook page. The goal of the survey is to gather data from a sample of people who live and work in Johannesburg, especially ones who live with a reasonable daily commute to the Wits main campus.

The survey asked respondents demographic questions of race, age, and gender before asking cycling-specific questions regarding bicycle access and ownership, frequency of cycling, and motivators and deterrents of cycling. The survey finished with nine questions, measured on a 5-point importance-based Likert Scale with the following levels: *1 – Not important at all, 2 – Slightly Important, 3 – Somewhat important, 4 – Important, 5 – Very Important*. Each of the nine Likert-scale questions asked how important it was to consider a specific spatial variable when choosing a route for cycling. The latent variable these nine factors are describing is the optimal route or path through a road network in Johannesburg for cycling paths to be installed on.

The survey was available to participate in online for a total of one month. A written transcript of the survey questions and possible answers is included in this report as Appendix A. The survey was distributed solely in English and online, did not offer any incentives to participants, and targeted two main groups: Wits students and JUCA members. For this reason, there is the obvious issue of sample bias towards English-speaking university students, who may be more open to change and more active than the general population of Johannesburg, those with access to email and the internet, and JUCA members, who will cycle much more often than the average citizen. With this targeted sample population, however, the survey responses crucially reflect the preferences of people living and working in and around the study area. Responses of survey participants who did not live in the greater Johannesburg area, which represented 95 total responses out of 805 in the Wits sample, were not included in the data analysis stage to ensure the final route output would best serve the needs of locals.

3.3.1 Ethics

The ethics considerations for this study were centred primarily in the survey. A main component of the survey asked where respondents live, along with other demographic data such as ethnicity. It was made clear that these questions were not mandatory, respondents would never be asked for their names, and that sensitive demographic information for a single respondent would never be published in the study. Various home address locations were mapped and aggregated to select averaged origin nodes or regions, as opposed to publishing exact addresses that correspond to single survey participants. Furthermore, survey participants were not asked to provide their street address, but rather their neighbourhood or postal code to ensure anonymity. Participants were not offered any incentives for participating in the survey. The formal ethics clearance process required by the faculty was adhered to, and the survey was implemented after the ethics clearance was complete.

3.4 Data Sources, Collection, and Validity

This study used open source and proprietary secondary data to complete the aim of bicycle infrastructure route design. While the survey ultimately informed the study on what factors and therefore data sources were used, the following list of data in Table 1 are consistent with similar previous studies in the literature, and were pursued prior to the finalization of survey results.

Secondary data were collected from sources in accordance with the Spatial Data Infrastructure Act of 2003 (SDI Act), including formal request letters on Wits letterhead describing the purpose of the study and details of how data sets would be used and published. ESRI South Africa maintains access to the HERE dataset for the country that includes location, land use, road network, and traffic data for the entire country. This data is updated quarterly and the second quarter 2016 data for the Johannesburg area was obtained free of charge from ESRI South Africa with the understanding it would be used solely for research purposes. This limitation on the HERE data set was followed and the data will be deleted at a future date.

Elevation contour data was sourced from NGI, with the most recent update of December 2006. Although there may be other more up to date elevation datasets, this date was free to access and readily available, and provided more than adequate spatial and temporal accuracy for the application. Location data described above was also available within the NGI dataset, however the HERE dataset was used instead as it was more up to date.

Table 1 - Spatial data types and sources. POI = Points of Interest

Data Set	Data Type	Source(s)
Road Network	Vector: polyline	ESRI South Africa
Elevation	Vector: polyline contours	NGI
Traffic Volume	Vector: point or line	ESRI South Africa
Land use	Vector: polygons	ESRI South Africa
Public Transportation stops	Vector: polyline	ReaVaya, Gautrain, City of Joburg
Crime	Vector: point or polygon	JMPD stations
Population Density	Vector: polygon	SA Census
POI – Shopping Areas	Vector: multipoint	ESRI South Africa
POI – Commercial Hubs	Vector: multipoint	ESRI South Africa

Public transportation stops were incorporated using data from the ReaVaya, Gautrain, and City of Joburg websites, as well as cross-checking these with data from ESRI South Africa. Crime instance data spanning the entire calendar year of 2014 was collected by Sulaiman Salau (Statistics Department, Wits University) and used in this study with permission. These data were collected directly from a contact at the Johannesburg Metro Police Department in database format. It was collated from each police precinct within the potential study area into a master table with addresses. These addresses were geo-coded to create points for each crime instance in a process described in detail in Section 3.6.2.

With numerous data sets originating from multiple government and private sector sources, any existing metadata was consulted to ensure the scale, accuracy, and validity of the data. Any qualifiers of the data due to unavailable metadata, unknown collection methods, or other problems are properly identified and communicated in sections below and summarized in Section 6.

3.5 Data Analysis and Interpretation

Data analysis in this study is broken down into three major tasks: analysing data from the survey, quantifying the factors into network attributes, and employing the attributes in Network Analyst to create the solved bicycle routes.

3.5.1 Survey Validity

To properly test the validity of the Likert-scale answers in the survey, an identical survey was distributed on two separate occasions to a group of the same respondents. This was achieved after the deployment of the original survey by sending the identical survey to a separate, smaller group of respondents who would be able to be contacted multiple times. The respondents in this survey were identified by the final three digits of their mobile phone number rather than their names. This was asked to facilitate matching their first and second responses while still maintaining anonymity. The survey was sent to a separate control group of 35 people, 26 of whom responded. In the initial survey email, the group was advised that a "second portion" of the survey would be sent in two weeks' time, but it was never revealed that the second part of the survey was identical to the first. After deployment of the second identical survey, there were a total of 22 respondents whose Likert answers were correlated to the first survey, and these were used in intra-rater agreement testing.

3.5.2 Descriptive Statistics and General Observations

General descriptive statistics were calculated in both Excel and SAS to have a better understanding of general attitudes towards the nine variables asked about in the Likert scale questions. As suggested by Refaat (2007), the modes of each variable were assessed for these questions, as well as the frequency at which each rating (1-5) was chosen. Bar charts were used to visually display the most and least popular variables, as well as compare against the different sampling groups.

3.5.3 Comparison Tests

Using SAS software, comparison tests were carried out to compare different strata of the survey responses. The data generated from the survey was ordinal in scale. Normal t-tests for comparison were not used because the assumption of normal distribution and the requirement for continuous data could not be met. Instead, tests for categorical data were used in the place of t-tests, specifically the two-sample proportion test using the 'Table Analysis' feature in SAS.

3.5.4 Ranking Spatial Factors

The main aim of statistical analysis of the survey data is to rank the nine variables based on their importance in affecting bicycle lane route choice. Once they are ranked, it can become more clear which variables should be included in the final algorithm and how heavily they should be weighted. Additionally, it can be decided which variables, if any, should be completely removed from the model. One way to achieve an importance ranking is to use total sums of the Likert scale answers, with the highest sum ranked as the most important (Gob *et al.*, 2007). In the Likert scale used in this survey, however, there are four 'positive' levels and just one 'negative' level (Not important at all). For this reason, these empirical ranked sum results are also compared to sums based on an adjusted scale: (-1) for not important at all, and then 1-4 based on the four levels of importance. The latter method applies more negative weight to answers when a respondent feels a variable has no importance in deciding cycling routes.

In addition to ranking based on empirical sums of Likert answers, the mode and frequency percentages of each answer to the Likert-scale answers are compared to identify the most and least important spatial variables which the sample consider important in determining bicycle lane routes. The results of these rankings are displayed in Table 11 and the final selection for which variables to use in the shortest-path algorithm is discussed in Section 4.5.

3.5.5 Principal Component Analysis for Dimension Reduction

In ranking the spatial variables, the variability explained by the Likert scale ratings of each of the nine factors is not taken into account.

Using principal component analysis (PCA) in conjunction with the importance rankings affords a stronger picture of which variables should be included and which should be left behind. It is widely agreed that performing PCA or other types of factor analysis is specifically reserved for continuous data (Basto and Pereira, 2012). However, many sources found that using a different type of correlation matrix with PCA, called a polychoric correlation matrix, is an appropriate approach when dealing with ordinal data on a Likert scale (Flora and Curran, 2004; Holgado-Tello *et al.*, 2010). The polychoric correlation coefficient assumes bivariate normal distribution of the latent variable (optimal bikeable route), which allows for the statistical assumptions of PCA to be fulfilled without using continuous data (Ledesma and Valero-Mora, 2007).

3.6 Quantifying Spatial Variables within Network Analyst

The Network Analyst geoprocessing toolbox in ArcGIS 10.3 was used to quantify each variable and create optimal cycling routes once these variables were included in the network dataset. The network dataset includes the edited road network for the study area, with each edge of the network able to be related to geodatabase tables containing information on measured data for automobile traffic and free-flowing traffic speed. The HERE dataset also included Points of Interest (POI) such as gas stations, business facilities, banks, and grocery stores (Appendix F) and land use data used for quantifying commercial hubs, shopping centres, and public transportation stops. With these inputs, tools embedded in the Network Analyst processing environment in ArcGIS enable creation of solved routes, navigation, location-allocation, and multiple-stop delivery routes.

Network attributes are properties of the network dataset that are used to control navigation and routing. Cost attributes act as a summable impedances over each link, and usually default to distance or travel time. Distance was used as the cost attribute to ensure the optimal bicycle routes are following the shortest path from origin to destination nodes before considering the seven retained variables. Restriction-based attributes are traditionally used in automobile routing to avoid roads that are closed or turns that are illegal at certain intersections. In this study, restriction attributes are used to dictate links to be preferred and avoided based on their association with the seven spatial variables. The methods and expressions used to create these Boolean evaluators for each restriction attribute as they relate to the chosen variables are described in detail in Sections 3.6.1 through 3.6.7.

3.6.1 Avoidance of roads with high speed differentials

The HERE dataset includes data for each network edge, in both directions, on the average free-flowing traffic speed that is derived from measured traffic data, automobile speed data, and speed limit information throughout the city. To create the network attribute for the avoidance of roads with high speed differentials between cyclists and automobiles, the following steps were followed:

1. Add field to the street network attribute table to be used as the Boolean Yes/No indicator of whether a street has a high or low speed differential.
2. Dichotomize the free flow automobile traffic speed data points by choosing a 30 km/hour speed cut-off.
3. Populate the new field with 'Y' if the free-flowing automobile traffic speed is > 30 km/hr and 'N' if < 30 km/hr.
4. Repeat steps 1-3 for the 'FROM' direction for each link.

5. Add new attribute under network dataset properties, with a 'restriction' usage type and 'unknown' units, 'Boolean' for data type, and 'Avoid' for restriction usage.
6. Select the 'Evaluators' box, under 'Source Values' tab, select 'Field' for the 'Type' column and then click the 'Evaluator Properties' button.
7. Create a simple expression using VB Script (faster performance than Python within Network Analyst) to select all attributes where the created Boolean field equals 'Y'.
8. Repeat step 7 for the opposite direction.

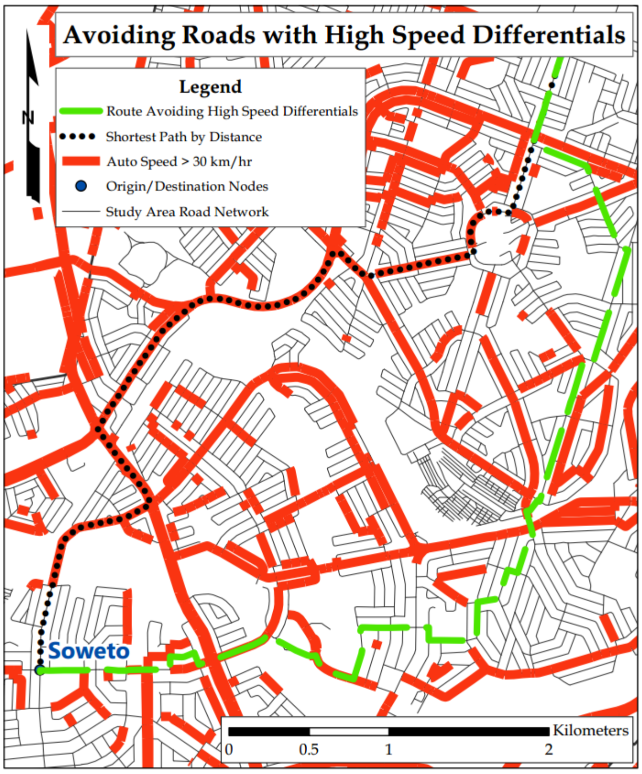


Figure 2 - Testing for route solving using the 'Avoid roads with high speed differentials' attribute. Roads with an average free-flowing automobile traffic speed greater than 30 km, derived from the HERE dataset.

Steps 1-4 quantify what represents a high-speed differential (automobile free-flow speeds of greater than 30 km/hr) by dichotomizing the free-flow speed variable and populating Boolean fields to indicate which side of 30 km/hr the speeds fall on. General commuting cycling speeds in a city environment are assumed to average 24 km/hr (Huang and Ye, 1995), and this factor attempts to single out roads with significantly higher automobile speeds. Steps 5-8 then use the properties of the network dataset for the study area to create a new network restriction attribute which is based on this data by executing the expression in Step 7. Creating a new route with this added restriction to avoid edges with high speed differentials now routes bicycle lanes differently compared to a simple shortest distance cost path using Network Analyst. In Soweto, for example, the shortest path towards Florida follows a main road with fast moving vehicle traffic, while the path avoiding high speed differentials begins east and cuts north along roads with slower automobile traffic (Figure 2).

3.6.2 Avoidance of areas with high crime

Raw crime data sourced from the JMPD was used to create a density-based quantification of crime incidents as they relate spatially to the road network within the study area. The study area defined in Section 3.2 includes fifty-two police precincts where crime incident data is collected including the location and type of crime. The data used in this study included fifteen types of crime (listed in Appendix F) and covers all recorded crime incidents in 2014, the most recent complete year data was available. The dataset represents 24207 separate instances of crime within these precincts which occurred at 13,333 unique locations.

The location data for these crime incidents relies, and is limited by, the data input by police officers reporting the crimes. The precision of locations varies wildly in the data set, from a simple description of nearby stores to a precise street address including house number and postal code. The inconsistent style of addressing in Johannesburg, especially in cases of informal settlements and townships, necessitated the need for a strong composite locator to correctly geocode as many crime points as possible.

The data sourced from ESRI included a complete composite locator for Johannesburg, meaning there would be a hierarchical nature to matching location addresses from the raw crime data to points on a map. When using the batch geocoding tool with the composite locator, the tool first tries to match an address to an exact street address location. If that match is unsuccessful, it will try to match to a single street (no property number), followed by points of interest (POI), a census place, a post code, and finally just a neighbourhood. This method increases the match percentage, but matching point locations to neighbourhoods is far less precise and potentially more inaccurate than matching with a specific house number.

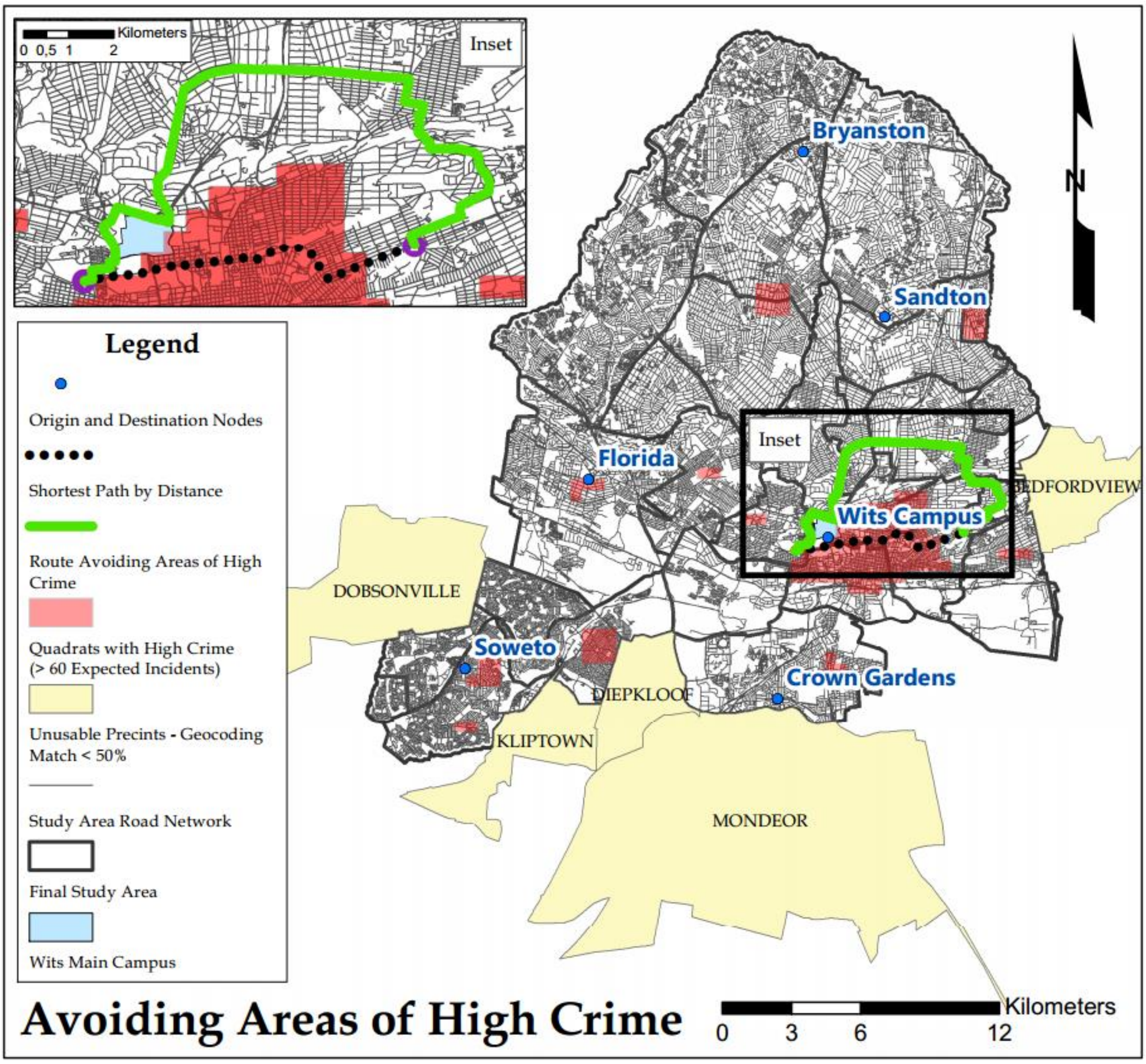


Figure 3 - Analysis and example of routing to avoid areas of high crime.

This method, combining geocoding processes using multiple fields for the address and a single line, resulted in a 79% overall match rate. Each precinct was analysed for individual match rates to examine which precincts suffered from poor location addresses. The five precincts in yellow (Figure 3) were eliminated from the study area due to match rates below 50% that represented an intolerable error rate for mapping overall crime in the study area.

The remaining 47 precincts exhibited geocoding match rates varying from 56 to 99 percent due to varying levels of quality of primary data recorded by officers. Some areas would incorrectly report more crime simply due to the existence of more accurate and comprehensive address data for that precinct. To fix this issue, crime incident data from the remaining precincts was normalized against the remaining precinct with the lowest match rate (Alexandra, 56%). Deleting crime points to complete the normalization of the data was accomplished by randomly assigning each point a number in Excel, ranking those numbers, and deleting the *n* lowest numbers for each precinct based on the ratio of match percentage to the lowest matched precinct. The resulting multipoint shapefile was used as the input for the following point density analysis.

Point density analysis was chosen after a series of tests using various kernel density methods in ArcGIS including varying quadrat sizes and neighbourhood weight matrices (queen's versus rook's case). The final density mapping was chosen using 500 square-meter quadrats over the study area, calculating the expected counts for point density in each quadrat using a rectangular 3x3 neighbourhood matrix (queen's case). A cut-off of 60 crime incidents (normalized) per quadrat for high crime was chosen, visualized in Figure 3. This cut-off highlighted problem areas of heavy crime which were to be avoided during bicycle routing. To create the network attribute, the density quadrats were polygonised and any network edges which intersected the quadrats were categorized in a new Boolean field as 'Y', indicating high crime. All other network links were categorized as low crime and are preferred during bicycle network routing. The inset in Figure 3 shows the solved route, which avoids areas of high crime, run north around the quadrats with crime incidents > 60, as opposed to the shortest path which runs directly through the CBD area of high crime.

3.6.3 Avoidance of roads with high automobile traffic

The ESRI dataset includes traffic pattern data presented as ratios of speed to the free-flow speed for each edge in a geodatabase table. Different values for each hour of the day and each day of the week are measured. When live or historical traffic data is visualized, the colours are coded as seen in Table 2. The hour of 07:00 to 08:00 on standard Wednesday morning was chosen to simulate a time of relatively heavy morning commute traffic in Johannesburg. The field containing ratios for this time and day was joined to the road network attribute table, and new Boolean fields were created for each direction as described in Section 3.6.1.

Table 2 - Categories for traffic speed based on HERE dataset (ArcGIS Telecom Team, 2015).

Colour	Traffic Description	Ratio to Free-flowing traffic
Green	Free-flowing	> 0.85
Yellow	Moderate	0.65 to 0.85
Orange	Slow	0.45 to 0.65
Red	Stop and Go	0 to 0.45

Selecting the cut-off for what constitutes 'heavy traffic' was done by trial and error to see which cut-off points would represent an appropriate number of network edges as having high traffic. Ultimately, the traffic ratio of 0.5 was selected to separate high and low, which means roads would tend to be avoided if their traffic speed was stop and go or on the slowest end of the 'slow' traffic type in Table 2. The Boolean fields were populated and the network attribute for avoiding roads with high traffic was added to the network dataset using the same steps listed in Section 3.6.1.

3.6.4 Connection to population centres

According to survey respondents, connecting to population centres is the most important variable for cycle routes to prefer, as the other variables higher on the list of importance are all avoidance variables. Instead of selecting 'Avoid' for restriction usage, 'Prefer' is selected so the links near population centres are preferred during route selection instead of avoided. Using South Africa Census data from the South African National Census in 2011, population densities for each census place falling within the study area were calculated with the field calculator in the attribute table.

Those census places with population densities in the 80th percentile or higher compared to others in the study area were defined as 'High density.' Any other area below this cut-off was considered 'Low density.' The dichotomization of the variable in this way enable integration of this variable into Network Analyst as a network attribute. Roads intersecting these census places with high population density were assigned the 'Y' label in a newly-created Boolean operator field. This was accomplished in the same fashion as described in Section 3.6.1.

3.6.5 Connection to commercial hubs

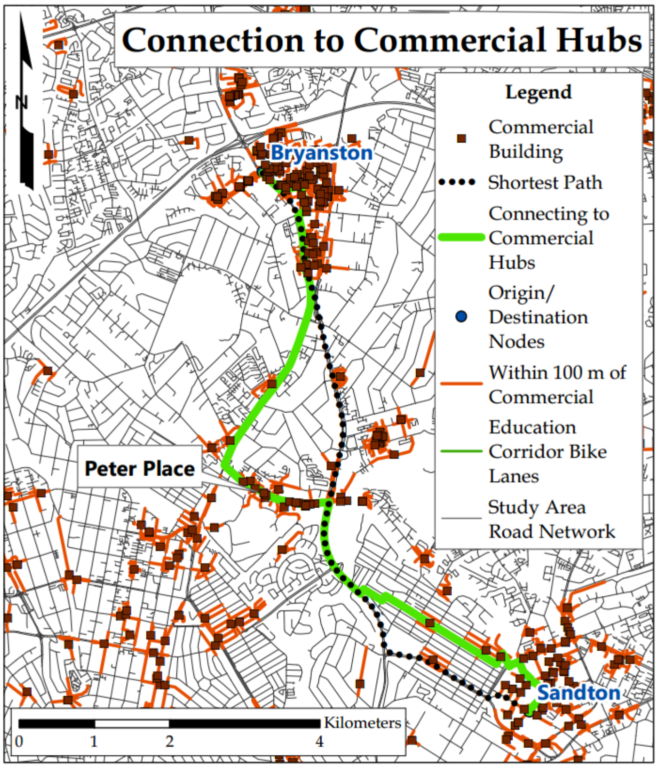


Figure 4 - Example of using Network Analyst to solve a route to prefer roads within 100 meters of a commercial building. Major deviations from the shortest path by distance or made to connect the user to clusters of commercial buildings.

The HERE data set also includes POIs, with the attribute table containing 71 different categories for possible points of interests. To quantify which roads were connected to commercial hubs, first the 'select by attributes' tool was used to select POIs under the following categories: Business Facility, City Hall, Civic/ Community Centre, Convention/ Exhibition Centre, County Council, Court House, Embassy, and Government Office (Appendix F). With a new multipoint dataset created representing these commercial-related POIs, a simple 'select by location query' was performed on the road network, only selecting network links that fell within 100 metres of a commercial building. When the route was solved to connect to commercial hubs, it changed significantly from the shortest path by distance to connect to multiple commercial buildings along Peter Place (Figure 4).

To incorporate this variable into the network dataset, a new text field was created to hold the Boolean 'Y' and 'N' operators, which were populated using the active selection of links that were within 100 metres of the commercial POIs. A new network attribute was created in a similar fashion to steps 5-8 listed in Section 3.6.1, except in this instance, 'Prefer' was selected under the 'Restriction Usage' menu. Any link with the 'Y' operator would be preferred over links with 'N' for the Commercial Hub connection field during routing.

3.6.6 Connection to shopping centres

Quantifying connection to shopping centres was identical in approach to quantifying commercial hubs (Section 3.6.5), with different POI types selected to represent shopping centres. Seventeen different POI types were selected (Appendix F) and the select by location tool was again used to select nearby network links. With this data, there were more shopping-related POIs than commercial one, so the 100 metre distance resulted in too many links being selected to be used in the network attribute. After a series of trial and error, 50 metres was chosen as the search radius to select network links, and these links were assigned the 'Y' operator in a new Boolean field for this variable. As with the commercial hub connection (Section 3.6.5), these links were selected to be preferred rather than avoided to reflect the connection variable type.

3.6.7 Connection to public transportation

This variable was quantified and added as a network attribute using the same methods as Sections 3.6.5 and 3.6.6. POI types for the points were Train Stations (including Gautrain), Park and Ride, Commuter Rail Stations, Bus Stations, and Taxi Stands. Additional ReaVaya and Johannesburg City bus stop point location sourced from the JRA were also added into this multipoint shapefile. The search radius used to select network links was 300 metres, representing a reasonable distance a user may cycle away from built routes to connect with public transportation. This radius was also much larger than those used for commercial hubs and shopping areas because there were significantly fewer public transportation points than the former POI types.

3.6.8 Weighing variables based on importance

Table 3 - Network Attribute relative weighting scheme for the seven used factors.

Network Attribute	**Type**	**Scale/Weight**
Avoid roads with high speed differential	Avoid	High
Avoid roads with high crime	Avoid	High
Avoid roads with heavy traffic	Avoid	Medium
Connect to population centres	Prefer	Medium
Connect to commercial hubs	Prefer	Medium
Connect to shopping centres	Prefer	Low
Connect to public transportation	Prefer	Low

With all seven variables quantified and incorporated into the network dataset as network attributes, the next step was to give different weights to the variables to reflect their order of importance listed in Table 14. This was accomplished by adjusting the Avoid or Prefer priorities in the network dataset settings based on the final rankings computed in Table 14. For these restriction-based attributes, they could either be avoided completely, avoided on three scales (High, Medium, Low), or preferred on those same three scales. The three most important variables were avoidance variables followed by four preference variables. While unable to assign specific weight ratios, the scales were assigned to each variable to give the most important ones more relative strength over the less important (Table 3). Using these relative weights, the shortest paths were solved for potential bicycle infrastructure in a way that reflected the surveyed users' answers.

3.7 Origin and Destination Nodes

Origin and destination nodes were selected based on a combination of locations of high survey response density, locations of population centres based on census data, and locations identified by the City of Johannesburg in their initial cycling infrastructure study (City of Joburg, 2009). The Wits main campus was a natural choice for a node due to its central location and connection to the clear majority of survey respondents. Sandton City was chosen due to the high density of commercial buildings, approximate 10 km distance from Wits, and inclusion in the City of Johannesburg study as an NMT focus area (Mohammed, 2009). Florida and Soweto's survey respondents were fewer than the northern suburbs, but they represented strategic nodes consistent with NMT focus areas (Mohammed, 2009) and in Soweto's case, extremely high population densities. Bryanston and Crown Gardens were chosen as nodes for their combinations of high survey respondent density, inclusion as an NMT focus area (Mohammed, 2009), and numerous commercial buildings nearby.

4. Results

This section describes the results and interpretations made from the statistical analysis of the Likert-scale answers in the cycling survey, as well as the final calculated routes using network analyst in ArcGIS.

4.1 Survey Validity

Intra-rater agreement testing was carried out to create a visual and nominal measure of survey reliability. Replicating techniques used in Rankin and Stokes (1998), a Bland-Altman plot and a weighted Kappa coefficient were created in Excel to measure agreement. The Kappa coefficient was also generated in SAS with accompanying confidence limits to test the value at the 5% significance level. Due to the ordinal nature of the data, the Bland-Altman plot was modified to include a z-value of frequency to clearly illustrate the spread of occurrences of agreements and disagreements.

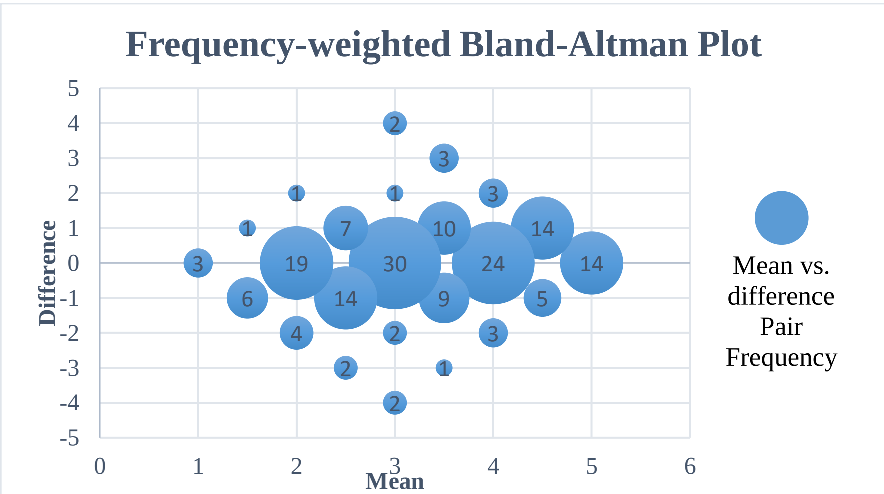


Figure 5 - Frequency-weighted Bland-Altman plot for analysing survey reliability.

To create a Bland-Altman plot, the mean and difference for each pair of corresponding Likert-scale answers from the two identical surveys was calculated. Plotting this data on a traditional Bland-Altman plot, however, would provide little clarity on agreement because of the nature of the ordinal Likert data falling on each integer 1 through 5. To solve this problem, the frequency of each mean-difference pair was also calculated for all responses. By including this z-value of frequency in the Bland-Altman plot as varying circle sizes, a much clearer image of intra-rater agreement is displayed.

Perfect intra-rater agreement would produce a plot with all features along the horizontal 0 difference axis of Figure 5. This Bland-Altman plot, is 'Frequency- weighted' for better visual interpretation, shows that most data seems to fall on or close to the zero-difference axis. This suggests good survey agreement. When survey respondents disagreed with themselves, the most common disagreements were changing from 4 to 5 (Important to Very important) and 3 to 2 (Somewhat important to Slightly important), both having fourteen instances. Extreme disagreements are much more uncommon, such as changing from 5 to 1 or 1 to 5, which both had only two instances. Because the successive identical surveys were implemented just two weeks apart, these rare extreme changes in importance ratings or more likely due to respondent error or laziness than a complete change in ideals (Cohen, 1968).

While the Bland-Altman plot provides a helpful visual map of intra-rater agreement, it does not provide an exact measurement of agreement like the Cohen's Kappa statistic is capable of. The weighted kappa statistic is used to measure agreement because it considers the degree of error when a respondent disagrees with themselves (Green, 1997). This is especially relevant when testing agreement of ordinal data because of the amount of close disagreements shown in Figure 5.

Table 4 - SAS Output of Kappa statistics and 95% confidence limits for intra-rater agreement (SAS Institute Inc., 2013).

Kappa Statistics				
Statistic	**Value**	**ASE**	**95% Confidence Limits**	
Simple Kappa	0.3509	0.0474	0.2580	0.4439
Weighted Kappa	0.4499	0.0489	0.3542	0.5457

Observation and weight contingency tables were populated in Excel to create the expectations table (Appendix B) which was used to calculate a Weighted Kappa statistic of $\kappa = 0.45$. To confirm this statistic at the 5% significance level, this procedure was recreated in SAS using PROC FREQ code in the 'Table Analysis' menu. In testing intra-rater agreement, the null hypothesis states that the ratings are independent, while the alternative hypothesis is that the ratings of the second survey are dependent on the first (Cohen, 1968). Cohen (1968), Green (1997), Rankin *et al.* (1998), and Sampat *et al.* (2006) agree that κ values greater than 0.75 reflect high agreement, values between 0.40 and 0.75 reflect "fair to good agreement beyond chance alone" (Green, 1997, p. 5), and any value below 0.40 exhibits poor agreement. In this test, the Kappa statistic of 0.45 classifies the intra-rater agreement as 'fair' and can be treated as legitimate beyond random chance. The significance of this statistic is confirmed by the 95% confidence limits of 0.35 and 0.55, within which zero does not fall, so the null hypothesis can be rejected (Table 4). Notice the weighted Kappa statistic is significantly higher than the simple Kappa because of the partial credit given to close disagreements.

4.2 Descriptive Statistics and General Observations

Of the 713 complete responses to the Wits survey, 55.3% were female and 42.78% were male. There were 33 complete responses to the JUCA survey, 27 of which came from regular cyclists. Females constituted only 42.4% of this group to 54.6% males. As expected for a cycling association, a majority (81.8%) of the JUCA respondents were regular cyclists (meaning they cycled at least 1-3 times per month) as opposed to just 12.2% of the Wits population. 30.9% of Wits respondents owned or had regular access to a bicycle while only 21.4% cycled at least 1 -3 times per year, meaning nearly 10% of Wits respondents had a bicycle, yet they chose not to ride it. A majority 56.4% of Wits respondents lived less than 10 km from campus, a reasonable distance to commute cycle if there is appropriate infrastructure (Sahlqvist *et al.*, 2015). Still, only 4.8% of that same sample commuted to campus via bicycle.

Table 5 shows that among active cyclists, the most common reason for riding a bicycle is to increase physical activity or for a fun leisure activity. Increasing the proportion of cyclists who ride bikes for commuting purposes is a key goal in increasing the usage of bicycle lanes (City of Joburg, 2009). But what are the main constraints for Johannesburg citizens that influence them to not ride a bicycle (or ride a bicycle less frequently)? In both Wits and JUCA populations, automobile traffic is the biggest deterrent, followed by the lack of cycling lanes and paths, and the safety concern of street crime. The complete tables of cycling motivators and deterrents are in Appendix C.

Table 5 - Motivators and reasoning for cycling in Johannesburg, Wits and JUCA responses.

	To commute	Physical Activity	Leisure Activity	Can't afford car	Save fuel	Competition
JUCA	45%	85%	73%	0%	24%	3%
Wits	22%	65%	81%	3%	13%	<1%

The nine factors presented in the survey were an accumulation of common spatial factors used in published bicycle routing research papers (Akar and Clifton, 2009; Winters *et al.*, 2010; Duthie and Unnikrishnan, 2014; Sahlqvist *et al.*, 2015). Survey respondents were also invited to include other factors or variables that they would like to be used in bicycle route planning, or that would be particularly pertinent to the Johannesburg study area. From both surveys, 200 unique responses to this question were tallied and some patterns emerged. The most popular response was 38 instances of people calling for laws and/or strict enforcement disallowing motor vehicles to use bicycle lanes, particularly taxi vans, which are prevalent and often drive erratically on Johannesburg roads. Other popular responses tended to be infrastructure-related rather than spatial in nature. This included the need for high quality, wide, and well-protected bicycle lanes (31 responses), the installation of adequate bicycle parking and locking infrastructure (24), regular bicycle infrastructure maintenance (14), physical separation of bicycle lanes from automobile traffic (12), and the need for better bicycling PR, instruction, and education for an overall change in culture related to cycling (10).

Spatial factors that attracted multiple mentions included the need for connecting to learning institutions (7), which was already a priority of the City (City of Joburg, 2009; Mohammed, 2009), better continuity and completeness (7), and the avoidance of intersections as much as possible (4). Continuity and completeness certainly is the goal, but can take a very long time to achieve. Intersections are nearly impossible to avoid, but avoiding large, complicated ones as they relate to automobile traffic could produce large benefits for bikeability (Pucher *et al.*, 1999).

4.3 Comparison Tests

Table 6 - Categorization of survey responses into binary indicators to facilitate proportion comparison testing.

Survey Response	**Categorical Value**	**Interpretation**
1-3 times per week, 1-3 times per month, 1-3 times per year	1	Regular Cyclist
Never, question left blank	0	Non-cyclist
Other	0 or 1	Judged based on written-in "other" response

Using the findings of the general descriptive statistics, various comparison tests were used to determine if the differences between populations were statistically significant. Using the table analysis tool in SAS, the cycling frequencies of males and females were compared. In the survey, there were four listed answers to the question "How often do you ride a bicycle", an 'other' option, and finally the possibility of the respondent leaving the question blank. To reduce these six possible responses to three categorical ones, categorical or binary values were applied to the responses (Table 6).

Equation 1 – Null and alternative hypotheses for comparison of male and female survey respondents.

$$H0: \pi Females = \pi Males; HA: \pi Females \neq \pi Males$$

Equation 1 states the null (*H*0) and alternative (*H*A) hypotheses for testing whether the proportion of the population of females who are regular cyclists is equal to or not equal to the population of males who are regular cyclists. The crucial SAS outputs from this table analysis includes the contingency table and the risk estimates table (Table 7). In the frequency table, the columns represent non-cyclists (0) and cyclists (1) based on the two genders. It is clear in this table that there is a higher proportion of non-cyclists for females (59.7%) than males (40.3%) among the sample.

Table 7 - Proportion comparison test between males and females and cycling habits, derived from SAS Output (SAS Institute Inc., 2013).

Gender and Cycling Frequency				
	Cycling Frequency			
Gender		0	1	Total
F	Frequency	329	79	408
	Percent	59.7	43.7	
M	Frequency	222	102	324
	Percent	40.3	56.3	

Column 1 Risk Estimates			
	Risk	Asymptotic 95% Confidence Limits	Exact 95% Confidence Limits
Row 1	0.806	0.768 - 0.845	0.765 - 0.844
Row 2	0.685	0.635 - 0.736	0.632 - 0.735
Total	0.753	0.722 - 0.784	0.720 - 0.784
Difference	0.121	0.058 - 0.185	

Values in the risk estimates table can test whether this a statistically significant difference that would reflect in the total populations can be tested using the values in the risk estimates table. This shows that the highlighted confidence interval is 0.058 to 0.185 for the asymptotic test. The null hypothesis is rejected because the confidence interval does not include the value of zero, and the following conclusion can be made: At the 5% significance level, there is a statistically significant difference between the population proportions of male and female cyclists. Specifically, males are more likely to be regular cyclists than females.

The proportion tests can also be used to test whether there is a statistically significant difference between Likert answers of Wits and JUCA respondents. This is crucial to better understand how perceptions of cycling differ between a group of primarily students and a group of primarily professionals who cycle regularly. Using proportions also adequately solves the problem of differing sample scales, as the respondents to the Wits survey vastly outnumbered those for the JUCA survey. Proportions are created from the Likert answers by comparing the amount who answered 1-*Not important at all* to those who answered 2-5, which all reflect varying levels of importance. The results are listed in Table 8, with corresponding SAS outputs available in Appendix D.

These conclusions are consistent with the findings in Section 4.4, where Likert-scale responses from Wits and JUCA respondents are ranked and compared. JUCA respondents felt avoiding roads with high crime was second-most important, compared to Wits respondents who believe crime avoidance was the most important variable in the location of cycling routes. Both JUCA and Wits respondents felt avoiding roads with steep slopes was the least important variable out of the nine offered in the survey. While each group valued this variable the least, JUCA respondents felt it was much less important overall than Wits respondents.

Table 8 - Results for comparison testing of various subgroups derived from the survey results.

How important is it to you that bicycle path routes are coordinated with public transportation stops?					
90.27	9.73	73.53	26.47	0.0175, 0.3173	The proportion of Wits respondents who believe coordinating bicycle paths to public transportation stops is important is greater than that of JUCA respondents.
How important is it to you that bicycle paths avoid areas with high crime rates?					
97.46	2.54	85.29	14.71	0.0021, 0.2413	The proportion of Wits respondents who believe bicycle paths should avoid areas of high crime is greater than that of JUCA respondents.
How important is it to you that bicycle paths avoid roads with steep slopes?					
85.63	14.37	67.65	32.35	0.0205, 0.3392	The proportion of Wits respondents who believe it is important that bicycle paths should avoid steep slopes is greater than that of JUCA respondents.
Important	**Not Important**	**Important**	**Not Important**	**Confidence Interval = 0.05**	**Conclusion made at the 5% significance level based on the inclusion or omission of 0 in the confidence interval.**
WITS		**JUCA**			

Both Wits and JUCA groups represent local citizens and residents of Johannesburg, but their perceptions differ significantly in what variables are important in selecting bicycle path routes. Even when there is seemingly agreement on the surface (both groups rating avoidance of slopes as relatively unimportant), the comparison tests showed there was still a significant statistical difference in proportions of 'Not important at all' ratings. This validates the need to consider the views of both groups when selecting the final set of criteria for bicycle routes. In the future, developing and distributing a survey to account for the entire Johannesburg population is essential for getting results reflective of the entire community.

4.4 Ranking Spatial Factors

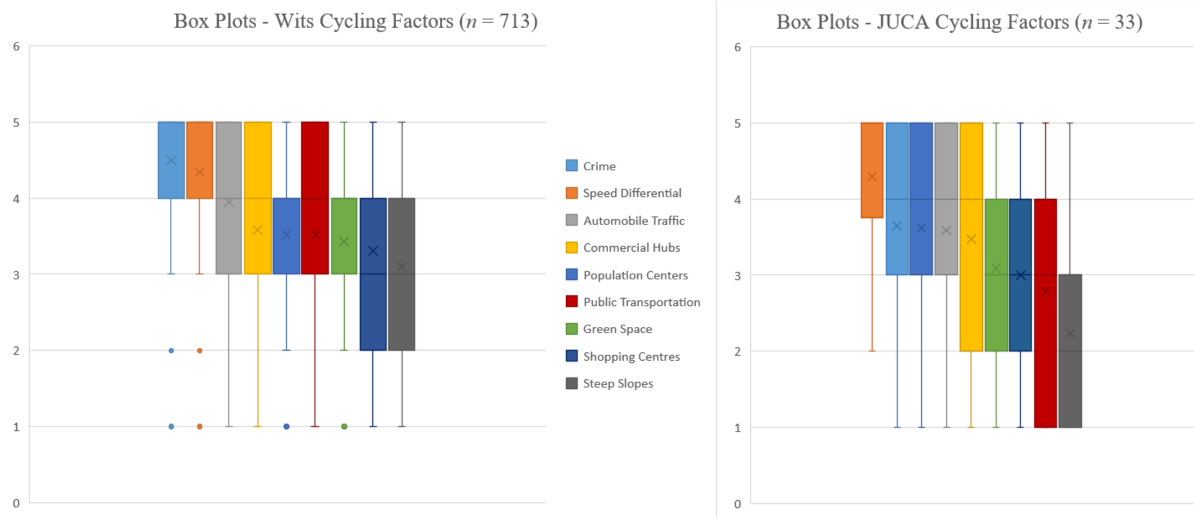


Figure 6 – Ordered box plots of Wits and JUCA responses to Likert-scale spatial variable questions. 1 = Not important at all, 5 = Very Important. Boxes represent the interquartile range; whisker ends represent minimum and maximum values.

To visualize the spread of answers in the Likert scale questions, two histograms reflect the results of both the JUCA and Wits surveys (Figure 6). All nine variables are represented, including four 'avoidance' variables and five 'connection' variables. Areas of crime, roads with high speed differentials between automobiles and cycles, roads with high automobile traffic, and roads with steep slopes are variables which cyclists may want to avoid. Commercial hubs, population centres, public transportation stops, green space, and shopping centres are all spatial factors which cyclists may want to connect to. In the survey, respondents were asked to apply an importance ranking to each variable based on the Likert scale discussed in Section 3.3.

In the Wits survey, avoiding areas of crime and roads with high speed differentials are clearly the two most important variables, with the ranges extending only to *3-somewhat important* save some outlier responses of 1 and 2 (Figure 6). Crime and speed differential also exhibited low variability, the only two variables which have the standard deviations below one ranking level (0.92, 0.99, respectively). The seven other variables (automobile traffic, commercial hubs, population centres, public transportation, green space, shopping centres, and steep slopes) have higher variability (standard deviations ranging from 1.12 to 1.27), but mean values between *3-somewhat important* and *4-very important*. The moderate nature of the responses for these seven variables do not make an obvious case for any variable to be removed from the final route criteria list.

Examining JUCA responses, speed differential is the top variable of importance and the only variable with a standard deviation of less than one (0.97). Unlike the Wits sample, the avoidance of steep slopes stands out as a much less important variable than the rest with a mean value of 2.24. As suggested in Gob *et al.* (2007), a more robust method of comparing Likert scale results is ranking using empirical sums and analysing the mode instead of the mean values of ratings. Because the ranking levels 1 through 5 are discrete, ordinal values and not continuous ones, the mode of each variable rating applies better to the data type (Norman, 2010).

Table 9 - Rankings for Wits responses based on empirical sums of Likert-scale questions. The Alt. sum represents the alternative method of assigning a score of 1 (Not important at all) a -1.

Wits Respondents **Variable**	Crime	Speed Differential	Automobile Traffic	Commercial Hubs	Population Centres	Public Transportation	Green Space	Shopping Centres	Steep Slopes
Mode	5	5	5	4	4	5	3	3	3
Sum	3190	3066	2763	2536	2495	2491	2423	2345	2199
Rank	1	2	3	4	5	6	7	8	9
Ratio (%)	100	96.11	86.61	79.49	78.21	78.09	75.95	73.51	68.93
Alt. Sum	2462	2338	2029	1778	1732	1713	1654	1570	1387
Rank	1	2	3	4	5	6	7	8	9
Ratio (%)	100	94.96	82.41	72.22	70.35	69.58	67.18	63.77	56.34

Referring to Tables 9 and 10, the modes for each variable represent the most popular answer from respondents for each question. Avoidance of crime, streets with high speed differentials, automobile traffic, and linkage to public transportation were all considered very important variables to Wits respondents more than any other level (mode = 5). Connections to shopping centres, green space and avoidance of steep slopes were most often considered somewhat important (mode=3) despite being relatively the least important in rank. For JUCA respondents, connection to commercial hubs (office parks, etc.) has a mode of 5, and public transportation is rated *1-not important at all* more often than any other level. This disparity in ratings from JUCA and Wits respondents is likely due to the older sample of JUCA members (average age of 39 compared to 25 in Wits respondents), and the fact that Wits respondents most likely commute to a campus, not an office park.

The sums were used in Tables 9 and 10 to rank the relative importance of each variable, with the corresponding table ratios showing the percent of an inferior variable divided over the top-ranked variable. These percentage values are useful to achieve a better understanding of relative importance of each variable. The alternate sum is designed to place a negative penalty for any ratings of 1-*Not important at all* as the other four ratings are varying levels of at least some importance. Using this sum does not necessarily change the relative rankings, but it does make it clearer which variables may be too far separated from the important ones to be included in the study at all. Using a cut-off level of a 70% sum ratio, for example, eliminates the last four variables in both Wits and JUCA surveys if the alternate sum is used, whereas the same percentage cut-off would retain 8 of 9 variables in Wits responses and 6 of 9 for JUCA.

Table 10 - Rankings for JUCA responses based on empirical sums of Likert-scale questions.

JUCA Respondents **Variable**	Speed Differential	Crime	Population Centres	Automobile Traffic	Commercial Hubs	Green Space	Shopping Centres	Public Transportation	Steep Slopes
Mode	5	5	4	5	5	2	3	1	1
Sum	146	124	123	122	118	105	102	95	76
Rank	1	2	3	4	5	6	7	8	9
Ratio (%)	100	84.93	84.25	83.56	80.82	71.92	69.86	65.07	52.05
Alt. Sum	112	85	85	85	80	68	62	52	31
Rank	1	2	2	2	5	6	7	8	9
Ratio (%)	100	75.89	75.89	75.89	71.43	60.71	55.36	46.43	27.68

The most notable difference between the two groups is the importance of public transportation, valued highly among Wits respondents and relatively low for JUCA respondents. The goal of the study is to include local residents' opinions on cycling route variables, not just Wits students and staff, so each group's opinions are weighted equally in this case despite the JUCA group constituting a smaller sample and population size. Ideally, further study would facilitate a survey with a wider reach of Johannesburg residents for a more representative target population.

Table 11 - Combined importance rankings of Wits and JUCA response to Likert-scale spatial variable questions.

Combined JUCA and Wits Responses **Variables**	Speed Differential	Crime	Automobile Traffic	Population Centres	Commercial Hubs	Green Space	Shopping Centres	Public Transportation	Steep Slopes
Ratio	97.68%	87.95%	79.74%	73.12%	71.98%	64.14%	59.65%	58.05%	42.01%
Rank	1	2	3	4	5	6	7	8	9

The mean of each corresponding variable importance ratio for the alternative sums in Tables 9 and 10 is calculated to achieve a final ranking reflecting both JUCA and Wits respondents equally. The results, displayed in Table 11, use a 70% cut-off point to eliminate four variables which rank as the least important based on the combined responses of Wits students and staff and JUCA members. Overall, avoiding roads with a high automobile-to-bicycle speed differential is considered the most important variable in deciding routes for bicycle lanes and paths. This is followed by avoiding crime, avoiding heavy automobile traffic, and connecting to population centres and commercial hubs.

4.5 Principal Component Analysis for Dimension Reduction

PCA was performed using a combination of the PROC CORR function in SAS, and the specialized Excel add-in Realstats (Zaiontz, 2015). Realstats uses a custom array function in Excel to compute eigenvalues and eigenvectors from the polychoric correlation coefficients generated in SAS. From this data, analysis of principal component loadings and retention was completed using the Kaiser criterion and examining scree plots, also generated in Excel. Consistent with the importance ranking data, this exercise was completed separately for both the Wits and JUCA response data.

Table 12 - Polychoric correlation coefficient matrix for Wits responses. Coefficients generated in SAS and displayed using Excel.

	Polychoric Correlation Matrix – Wits Responses								
	Crime	Spd_ diff	Auto_ traf	Comm	Pop	Pub_ trans	Green	Shop	Slopes
Crime	1.0000	-0.0270	-0.0595	-0.0148	0.2060	0.1422	0.0226	0.0591	0.1911
Speed Differential	-0.0270	1.0000	0.1820	0.1156	-0.0109	0.0453	0.0142	0.0273	0.0282
Automobile Traffic	-0.0595	0.1820	1.0000	-0.0068	-0.0139	0.0414	0.0068	0.1084	0.0036
Commercial Hubs	-0.0148	0.1156	-0.0068	1.0000	0.1438	0.1082	0.0298	0.1522	0.1314
Population Centres	0.2060	-0.0109	-0.0139	0.1438	1.0000	0.2095	0.0081	0.0825	0.0689
Public Transport	0.1422	0.0453	0.0414	0.1082	0.2095	1.0000	0.0795	0.1997	0.1548
Green Space	0.0226	0.0142	0.0068	0.0298	-0.0081	0.0795	1.0000	0.0661	0.0494
Shopping Centres	0.0591	0.0273	-0.1084	0.1522	0.0825	0.1997	0.0661	1.0000	0.1062
Steep Slopes	0.1911	0.0282	0.0036	0.1314	0.0689	0.1548	0.0494	0.1062	1.0000

In accordance with Flora and Curran (2004), polychoric correlation coefficients were generated in SAS using the PROC CORR and POLYCHORIC functions for each pair of the nine variables. These correlation coefficients are "maximum likelihood estimates of the Pearson's correlations for those underlying normally distributed variables" (Basto and Pereira, 2012, p. 4). The theoretical underlying variables are assumed to have a bivariate normal distribution, which satisfies the assumptions for normal PCA (Refaat, 2007). If PCA was conducted normally with a covariance matrix or standard matrix of correlation coefficients, the results would be highly skewed due to the difference in scales between continuous and ordered ordinal data (Holgado-Tello *et al.*, 2010). SAS is unable to generate a matrix of the polychoric correlation coefficients, so a simple list of coefficients for each pair of variables was created for both the Wits and JUCA responses. These coefficients are displayed in Appendices E1 and E2.

Using Excel, the coefficients generated by SAS were organized into a square matrix seen in Table 12 for the Wits responses (Appendix E3 for JUCA). The polychoric coefficients in this matrix are all generally low, with the highest correlation coming between connection to population centres and connection to public transportation stops (highlighted). This correlation coefficient is still quite low at 0.21, suggesting that each of the nine variables in the survey were somewhat independent and unique spatial factors for determining optimal cycling routes.

Table 13 - Principal Component Analysis matrix with eigenvalues and eigenvectors for all nine principal components. Using a cut-off of 0.4, variables with high loadings to certain principle components are highlighted in yellow.

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9
Eigenvalues	**1.6723**	**1.2371**	**1.0923**	**1.0237**	**0.9313**	**0.8530**	**0.8270**	**0.7121**	**0.6511**
Crime	0.3744	-0.2923	0.4135	-0.2759	-0.1802	-0.3885	-0.2528	0.2385	-0.4724
Speed Differential	0.0905	0.6592	0.0469	0.0604	-0.0878	-0.3391	-0.5758	-0.2718	0.1530
Automobile Traffic	-0.0508	0.6143	0.3475	-0.2708	-0.0023	0.3147	0.2462	0.5079	-0.0912
Commercial Hubs	0.3455	0.2773	-0.2805	0.4260	0.2858	-0.3389	0.4134	0.0544	-0.4161
Population Centres	0.4264	-0.0997	0.2030	-0.1532	0.6782	-0.0731	-0.0114	0.0899	0.5197
Public Transportation	0.4866	0.0824	-0.0277	-0.2485	0.0095	0.5487	0.0059	-0.5584	-0.2844
Green Space	0.0915	0.0841	-0.5078	-0.6778	-0.2152	-0.3346	0.2769	-0.0022	0.1799
Shopping Centres	0.3901	-0.0446	-0.4848	0.1773	-0.1848	0.3153	-0.3927	0.5310	0.0982
Steep Slopes	0.3880	-0.0070	0.3041	0.2989	-0.5811	-0.0489	0.3769	-0.0805	0.4231

Next, the eVECTORS function in Excel from Realstats (Zaiontz, 2015) was used to create the eigenvalues of each principal component, along with the eigenvectors displayed in Table 13. This matrix displays which variables load highly onto each principal component. These are listed in order of descending value on the top row. Using a cut-off point of ±0.4, principal component loadings are highlighted to gain a visual understanding of which variables correspond to the principal components. For the Wits respondents, connection to population centres and public transportation stops both load highly on the first principal component, which in turn described the most variability in the data (18.6%) compared to the other principal components. Avoiding steep slopes, on the other hand, does not load highly on the first four principal components, only PC5 and PC9. The Kaiser criterion, choosing principal components to retain based on eigenvalues greater than 1, is often used in concert with analysis of the elbow position of a scree plot (Baglin, 2014). In this case, the Kaiser criterion would eliminate the final 5 principal components and retain the first four, which are correlated highly to all variables except for avoiding steep slopes.

The Scree plot (Figure 7) generated from this data gives another visual representation of how much variability each principal component describes. The elbow in the plot, agreed by many as the indicator or cut-off point for retaining principal components (Basto and Pereira, 2012; Flora and Curran, 2004; Zaiontz, 2015) occurs at PC2. After this point, the variability explaining power of each principal component decreases steadily from PC 2 to PC 9. In this case, the Scree plot identifies just two principal components to retain, which cumulatively explain only 32.3% of the variability in the data. The Kaiser criterion, on the other hand, identified four principal components which explain a total of 55.8% of the variability in the data. To ensure better explaining power and to not remove any variables that may be important to determining optimal cycling routes, the Kaiser criterion method is used to retain the first four principal components.

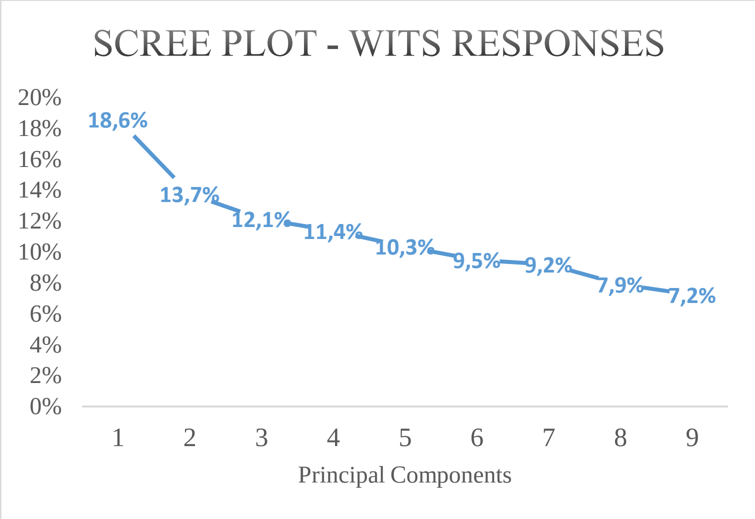


Figure 7 - Scree plot of proportional eigenvalues for Wits Responses. Y-axis represents percent variability explaining power.

As seen in Table 13, the only variable that does not load highly on principal components one through four is the avoidance of steep slopes. This variable is also ranked as the least important by Wits respondents, so it may be helpful to eliminate this variable from the route-finding study. The same principal component analysis was completed for the JUCA response data, and the corresponding SAS outputs, matrices, and scree plot are in Appendix E3. Using the Kaiser criterion, only the first three principal components are retained in this case, while the Scree plot is generally inconclusive, with variability explanation percentages decreases uniformly. In retaining the first three principal components, only steep slope avoidance and connection to green spaces do not load highly on at least one of the three PCs when using a cut-off point of ±4. Public transportation and population centres are again loaded highly to the first principal component, indicating these are both important factors in explaining the variability of the data.

Table 14 - Final seven retained spatial variables, ordered by importance based on combined importance rankings.

	Final Retained Spatial Variables, In Order of Importance						
Variable	Speed Differential	High Crime	Auto Traffic	Population Centres	Commercial Hubs	Shopping Centres	Public Transportation
Type	Avoid	Avoid	Avoid	Connect	Connect	Connect	Connect
Rank	1	2	3	4	5	6	7

Based on the combined available evidence, the variable of avoiding steep slopes can safely be eliminated from the final shortest-path algorithm, as it is ranked last in importance overall, and also does not load highly on the retained principal components for both the Wits and JUCA response data. While connection to shopping centres and public transportation stops rank poorly in importance (7th and 8th, respectively) in the overall rankings, these variables load highly on the retained principal components for both response sets, explaining significant variability in the data that should be retained for the final model. Connection to green spaces can be considered on the bubble. It ranks moderately (6th) in overall importance, yet loads highly only on principal components 3 and 4 of the Wits data, while absent from the retained principal components of the JUCA data. In the name of streamlining future spatial analyses for the routing task, this variable was eliminated as well.

Table 14 lists the final set of variables that were retained for input into the shortest-path algorithm. By applying the relative weights generated by the importance rankings in Table 11, each of these variables can be scaled relative to their importance as rated by Johannesburg locals.

4.6 Solved Routes and Shortest Path Comparisons

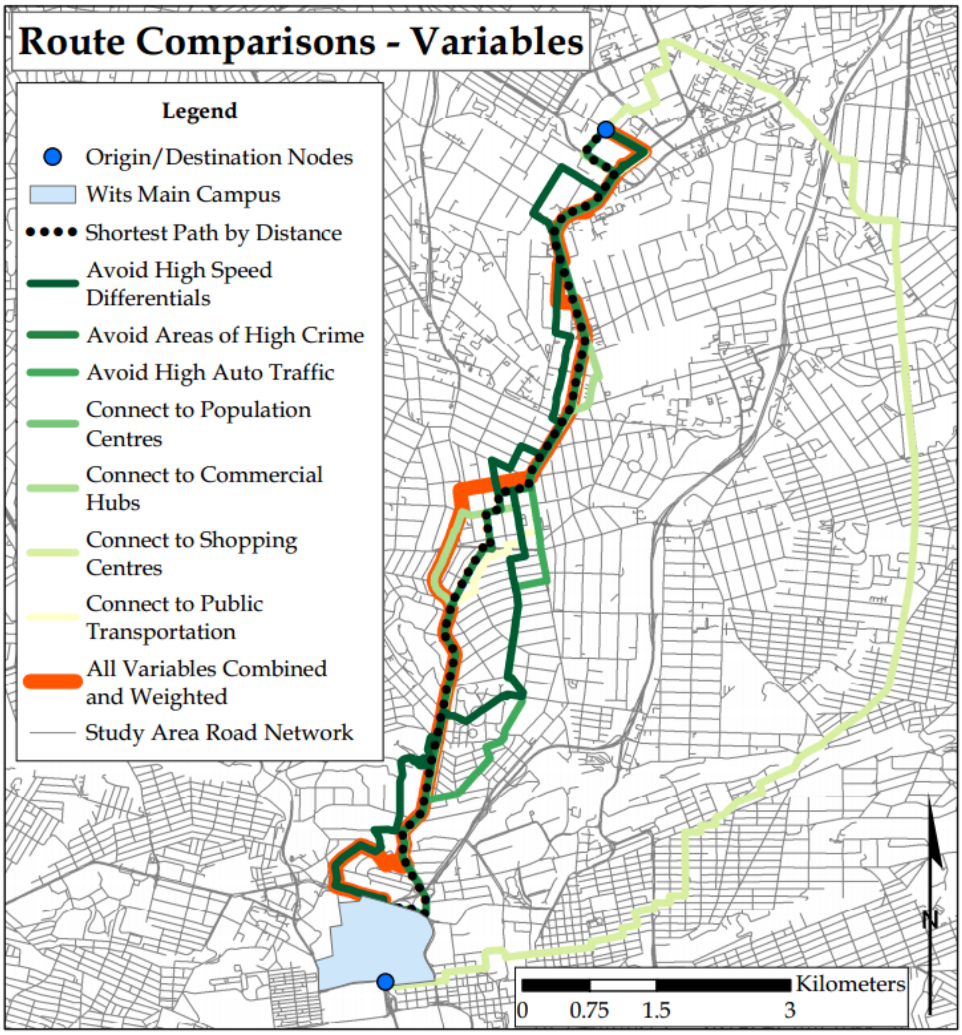


Figure 8 - Route comparison of each separate factor, all combined, and the shortest path by distance.

With the seven additional network attributes created to enforce routing restrictions based on the variables, and the relative weighting of the variables in place, a route was created using the *Solve* function in Network Analyst. Each network attribute could be activated or kept deactivated before solving the route, allowing for comparison of routes in Figure 8 based on which variables are used as route restrictions. As an example, the origin node of Central Sandton to the destination node of the North Main Gate at the Wits main campus is shown.

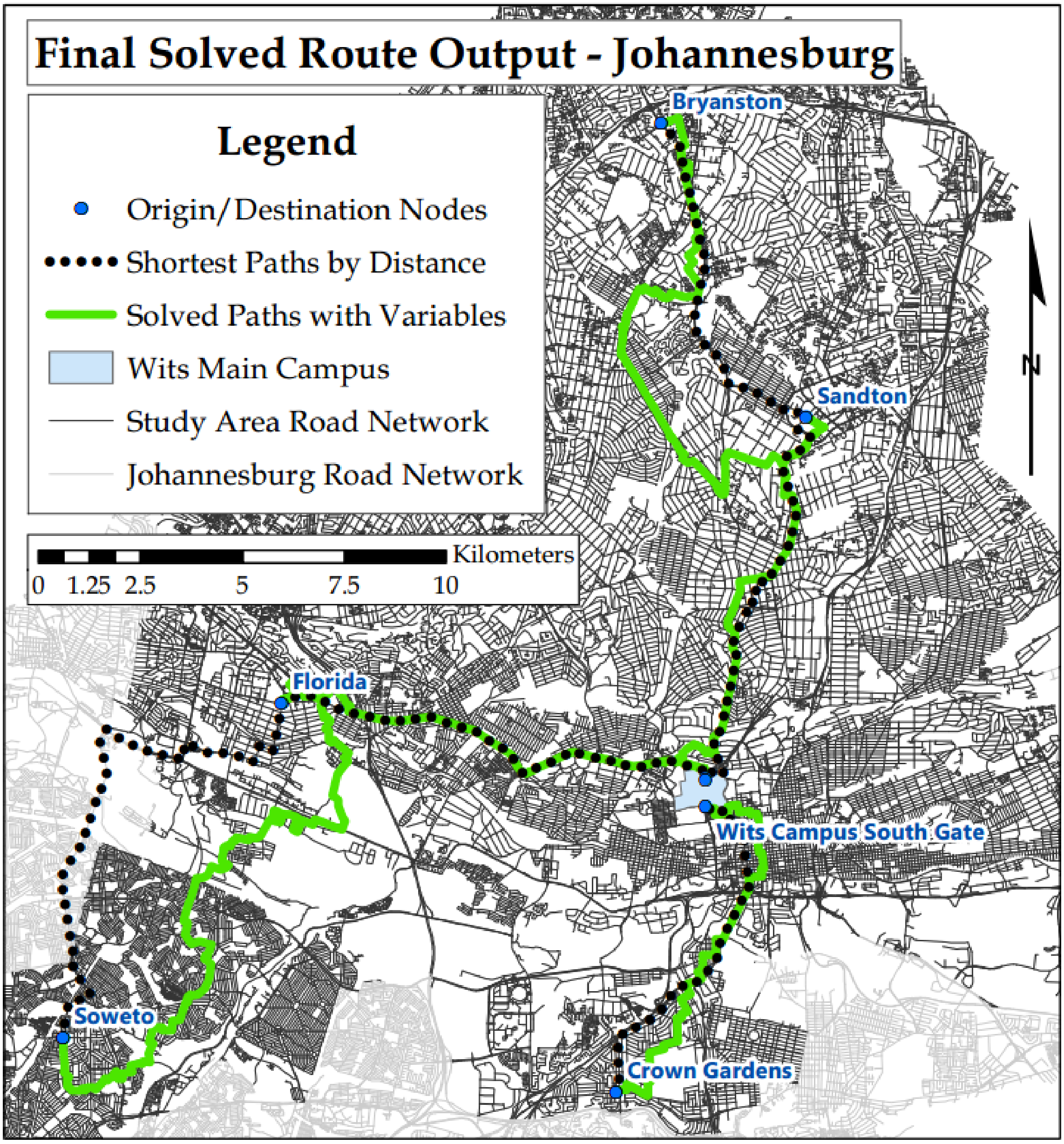


Figure 9 - Final solved routing output for the full theoretical bicycle lane network in Johannesburg.

The shortest possible path along the Johannesburg road network shown on Figure 8 covers 11.1 km while, while the other paths prioritizing one variable each range between 11.1 and 19.0 km. When routing only relies on one survey-defined variable, it may result in a much longer route that goes out of the way to fulfil the variable (connect to shopping centres), or a route that winds its way through the neighbourhood making many unnecessary turns (avoid high speed differentials). When these variables are added together and used in Network Analyst with the relative weights, a more realistic path for bicycle lanes is generally produced, balancing all seven variables while also getting a potential rider from Sandton to the Wits Main campus in 12.5 km, 1.4 km or 12.6% more than the simple shortest path. This added distance is potentially worth the benefits this route has based on avoiding and preferring certain routes as they relate to the seven variables (Ehrgott *et al.*, 2012).

Figure 9 shows routes solved using all the variables for each leg of the study area, creating a potential city-wide network of cycling lane routes. Alongside for comparison are the simple shortest path routes from node to node. In some cases, the solved route differs very slightly from the shortest path, such as the route from Florida to the Wits main campus, where only a few turns at the beginning of the route differ from the shortest path by distance. Similarly, the route from Sandton to Wits generally follows the shortest path, with a significant divergence near the Wits campus to avoid the busy intersections between Empire Rd., Jan Smuts Ave., and the M1 ramps (Figure 11A).

Figure 10 is pulled directly from the City of Johannesburg's *Framework for non-motorized transport* (2009) and indicates generalized routes for a citywide network of bicycle lanes and paths. Many portions of these routes in Figure 10, such as the Education Corridor (purple) and Soweto Network (lime green) have since been constructed. Compared to theoretical network in Figure 9, these routes cover similar areas and origin destination nodes. The routes that have been constructed are underused (Boye, 2015), and were not derived using any data from cyclists or residents. By making route and node adjustments based on survey and other data collected in this study, there is potential for greater use by cyclists in the future.

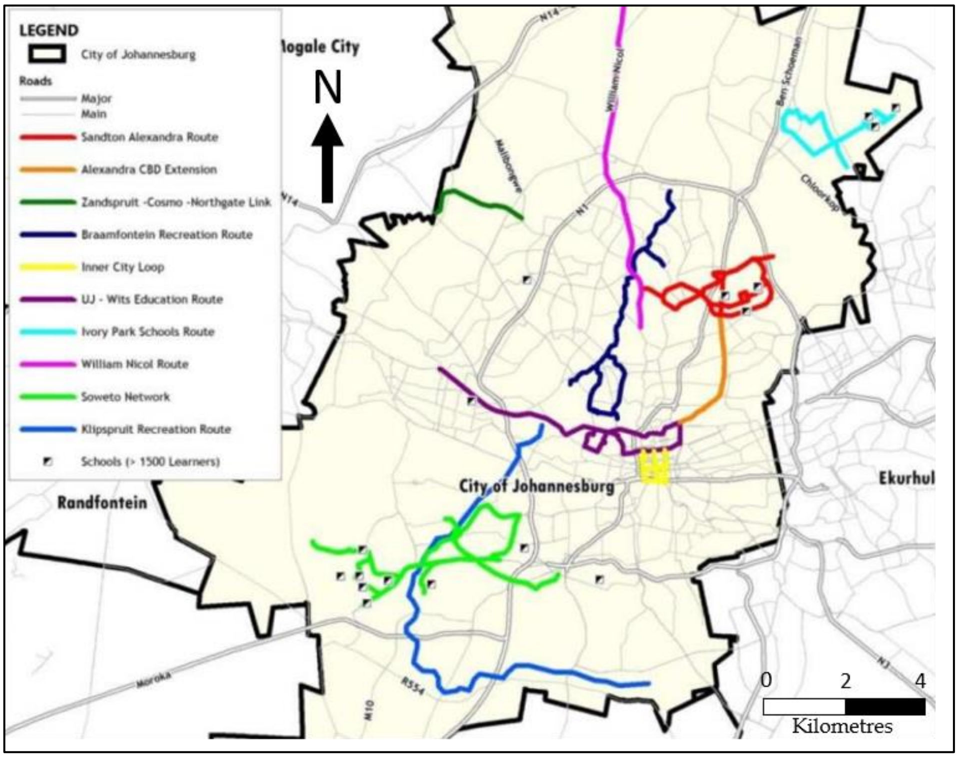


Figure 10 - City of Johannesburg and JRA proposed Non-motorised transport network (City of Johannesburg, 2009)

In some instances, the route that accounts for all selected variables is completely different from the shortest path, as seen in the route from Soweto to Florida (Figure 9). This highlights a potential shortcoming of the model, where a much longer, meandering route is solved at the expense of a short route. The main roads in Soweto are clogged with traffic during morning weekday commutes, rendering them mostly unusable in the model, which then defaults to the meandering side streets of the township neighbourhoods. The structure and patterns of these neighbourhoods differ greatly from those in the northern suburbs, a well-documented artefact of decades of Apartheid rule (Todes, 2012; Stones, 2013).

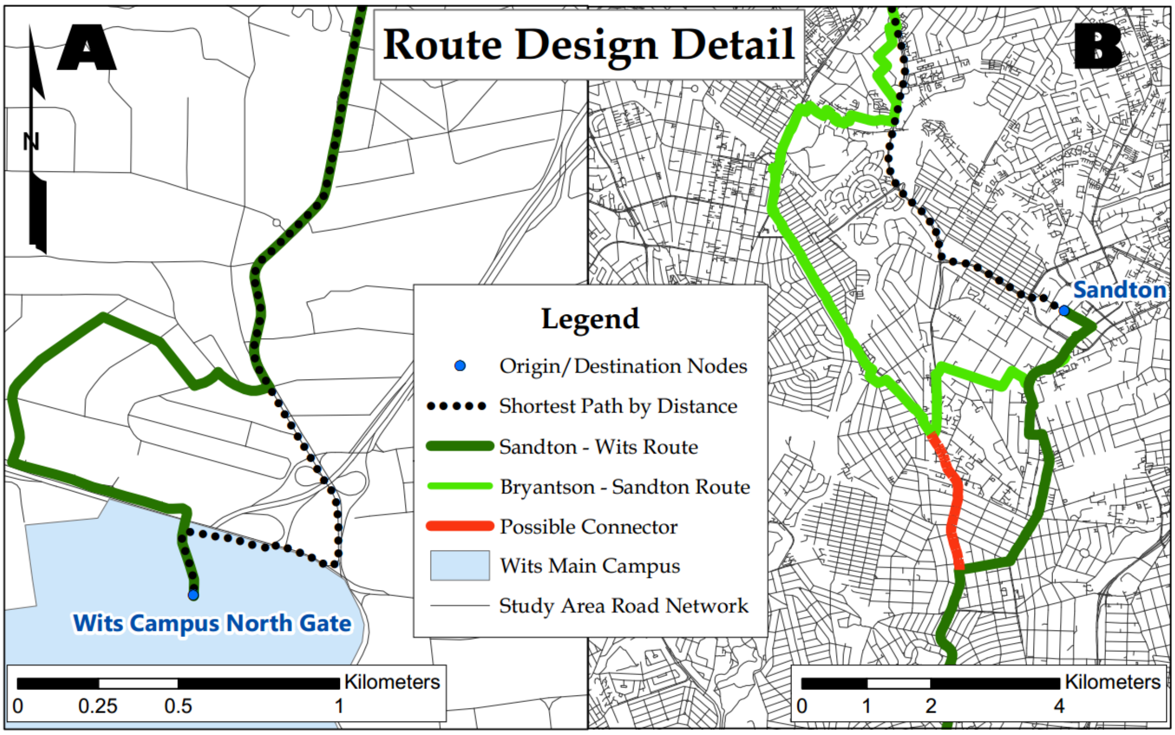


Figure 11 - Final solved route details. 11A represents detail at the northern main gate of the University of the Witwatersrand and the route used in dark green to avoid traffic along Jan Smuts Ave. 11B shows the large deviation the Bryanston-Sandton route takes to connect to more shopping areas and avoid traffic along William Nichol Drive.

Another major divergence from the shortest path is on the route from Bryanston to Sandton, where the variable-defined route cuts away from William Nichol Dr. for Jan Smuts Ave., mainly due to the preference for less auto traffic during morning commute hours (Figure 11B). This detour adds nearly eight extra kilometres to the route, which may not be worth it for many cyclists placing time and distance at the highest priority over the other spatial factors they felt were important on the survey.

5. Conclusions

The final output of this project represents spatial routing driven by the surveyed responses of local students and cyclists. As an alternative to the methods used by JRA, these routes are backed by data-driven research which can be coupled with urban planning and design to create user-friendly routes for cycle lanes. Likert scale data from the surveys were quantitatively ranked to identify the most important factors for locals. Principal component analysis, using a polychoric matrix of correlation coefficients, also identified which factors needed to be included based on the variability they explained. These ranking were combined, and the retained factors were quantified using an array of spatial data sets and dichotomized to allow for integration into the Network Analyst routing algorithm as restriction-based network attributes. Using this algorithm and the determined origin and destination nodes, a theoretical network of bicycle lanes was produced (Figure 9). These routes differed significantly from the simple shortest paths by distance between each node, validating the effectiveness of the route-finding method based on user-defined criteria. The solved routes also tended to be significantly longer in distance (12% to 109%) than the shortest path, and often made many unnecessary turns. Therefore, these solved routes represent a first step towards cycle route design that considers the views of current and potential users, rather than a rigid explanation for exactly where bicycle lanes should be installed. Local cyclists and potential users of cycling infrastructure can benefit the process in designing and planning bicycle routes by providing region-specific information on their spatial preferences as they relate to cycling infrastructure.

6. Limitations and Recommendations for Further Study

This study serves as a starting point in including the opinions of locals in the process of designing and planning future bicycle infrastructure in Johannesburg. Several aspects to this study could be improved upon in the future with access to higher-quality data, a larger target population for the survey, and a more task-specific, flexible algorithm.

The ideal target population for the survey would be the entire local population of Johannesburg, but distribution of the survey was limited to Wits students and staff as well as JUCA members. This smaller, younger, and more cycling-focused population no-doubt created biased results in the survey compared to a population of Johannesburg as a whole. The Wits and JUCA samples differed greatly in size, leading to potential statistical problems of comparing these two observational units (Kish, 1965).

In Figure 9, there are many instances of portions of the solved routes including an unnecessary amount of meandering turns. This could potentially be fixed by further fine-tuning of the network attributes or a more custom algorithm. Repeating the spatial analysis using different GIS software, such as the open-source QGIS, may be useful to compare results. Additional variables could be considered, such as bicycle lane construction cost, road width, or the availability of pavement space. The route-finding algorithm in Network Analyst allows for customization using scripts for each network attribute which could possibly be scaled to the percentages displayed in Table 11 as opposed to simple relative weighting. There is potential to also use cost-based network attributes instead of restriction-based attributes, which would eliminate the dichotomization of each variable, a practice that may often lead to statistically ambiguous results (Cohen, 1983).

This study was also limited by the data immediately available. Apart from the survey data, spatial data was entirely secondary, collected from a plethora of sources including the JRA, ESRI, JMPD, South Africa Census, and NGI. The data used in the route-finding model from these sources also ranged in collection date from 2011 (SA Census) to 2016 (HERE). Crime data, for instance, did not include the time of day for incidents, only crime types. These discrepancies in temporal accuracies must be acknowledged.

The modifiable area unit problem was encountered with measuring density crime locations using quadrats, as density map results change based when the quadrat size and shape is adjusted. Additionally, the mapping of survey respondent and crime point data are examples of an area-to-point change of support problem. The changing of these spatial scales to map point data comes with inherent unknowns and inaccuracies and must be acknowledged. Survey respondent locations were generally geocoded using a simple postcode, meaning this point data are only accurate to the centroid of each postal zone. The crime data, as discussed in Section 3.6.2, is also geocoded and grouped into precincts based on address data ranging from a detailed street address to simply a city name or postcode. For this reason, a large percentage of raw crime data was left out of the analysis because the exact point locations could not be considered accurate to the standards of this project.

7. Acknowledgements

Completion of this project was not possible without the excellent guidance of my supervisor, Dr. Stefania Merlo. Additional thanks go to Professor Jasper Knight for advice in implementing the survey. A vast amount of data from the HERE dataset was made available free of charge for this academic purpose thanks to Sanet Eksteen at ESRI South Africa. David du Preez provided great insight about current cycling culture in Johannesburg and a valuable platform to distribute the cycling survey to JUCA members. Thank you to Muhammed Suleman for taking the time to meet with me and discuss the findings of his research thesis regarding bicycle transportation in Johannesburg, and to Sulaiman Salau for providing me with the Johannesburg crime data which took him great effort to collect and curate. Finally, I'd like to thank my friends, new and old, family, and especially my wife for support in this endeavour.

8. References

Software

ESRI, 2015. ArcGIS Desktop: Release 10.3.1. Redlands, CA: Environmental Systems Research Institute.

R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/.

SAS Institute Inc. 2013. SAS/ACCESS® 9.4 Interface to SAS Enterprise Guide 7.1. Cary, NC: SAS Institute Inc.

Zaiontz, C. 2015. Real Statistics Using Excel – Principal Component Analysis. Accessed 18 September, 2011. http://www.real-statistics.com/multivariate-statistics/factor-analysis/principal-component-analysis/.

Literature

Akar, G. and Clifton, K. 2009. Influence of individual perceptions and bicycle infrastructure on decision to bike. *Transportation Research Record*, 2140, 165-172.

ArcGIS Telecom Team, 2015. *Live Traffic – ESRI*. Available from https://www.arcgis.com/home/item.html?id=8cd01a89cdc74c87893127f4438844b5, accessed 12 November, 2016.

Baddeley, A., Rubak, E. and Turner, R. 2015. *Spatial Point Patterns: Methodology and Applications with R*. London: Chapman and Hall/CRC Press, 2015.

Baglin, J. 2014. Improving your exploratory factor analysis for ordinal data: a demonstration using FACTOR. *Practical Assessment, Research & Evaluation*, *19*(5), 1-15.

Basto, M. and Pereira, J.M. 2012. An SPSS R-Menu for Ordinal Factor Analysis. *Journal of Statistical Software*, *46*(4), 1-29.

Bawa, H. 2015. Cycle Wits Strategy: Standards for integrated Planning. *Ludwig Hansen Architects and Urban Designers*, 1-36.

Bigazzi, A.Y. and Figliozzi, M.A. 2015. Roadway determinants of bicyclist exposure to volatile organic compounds and carbon monoxide. *Transportation Research Part D: Transport and Environment*, 41, 13-23.

Bil, M., Bilova, M. and Kubecek, J. 2012. Unified GIS database on cycle tourism infrastructure. *Tourism Management*, 33, 1554-1561.

Bivand, R. and Rundel, C. 2016. *rgeos: Interface to Geometry Engine - Open Source (GEOS)*. R package version 0.3-20. Available from https://CRAN.R-project.org/package=rgeos, accessed 18 September 2016.

Boye, A. 2015. City denies MMC parked in cycle lane during EcoMobility Festival. *Fourways Review,* 14 October. Available from http://fourwaysreview.co.za/210383/city-denies-mmc-parked-in-cycle-lane-during-ecomobility-festival/, accessed 20 April 2016.

Buehler, R. and Dill, J. 2016. Bikeway Networks: A Review of Effects on Cycling. *Transport Reviews*, 36, 9-27.

Caulfield, B., Brick, E. and McCarthy, O.T. 2012. Determining bicycle infrastructure preferences–A case study of Dublin. *Transportation research part D: transport and environment*, *17*(5), 413-417.

Cervero, R. and Duncan, M. 2003. Walking, bicycling, and urban landscapes: evidence from the San Francisco Bay Area. *American journal of public health*, 93(9), 1478-1483.

Chandra, S. 2014. Safety-based path finding in urban areas for older drivers and bicyclists. *Transportation Research Part C*, 48, 143-157.

City of Johannesburg, 2009. *Framework for non-motorized transport*. Internal Report, City of Joburg. Johannesburg, South Africa.

Cohen, J. 1968. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin*, *70*(4), 213.

Cohen, J. 1983. The cost of dichotomization. *Applied psychological measurement*, *7*(3), 249-253.

Dalton, A.M., Jones, A.P., Panter, J. and Ogilvie, D. 2015. Are GIS-modelled routes a useful proxy for the actual routes followed by commuters? *Journal of Transport & Health*, 2, 219-229.

Duthie, J. and Unnikrishnan, A. 2014. Optimization Framework for Bicycle Network Design. *Journal of Transportation Engineering*, *140*(7), 1-8.

Eck, J., Chainey, S., Cameron, J. and Wilson, R. 2005. *Mapping crime: Understanding hotspots.* Washington, DC. U.S. Department of Justice, National Institute of Justice.

Ehrgott, M., Wang, J., Raith, A. and van Houtte, C. 2012. A bi-objective cyclist route choice model. *Transportation Research Part A*, 46, 652-663.

Flora, D.B. and Curran, P.J. 2004. An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological methods*, *9*(4), 466-491.

GIBB, 2015. *Rosebank to Sandton Complete Streets Network: Feasibility and Design Report*. City of Johannesburg Transport Department.

Göb, R., McCollin, C. and Ramalhoto, M.F. 2007. Ordinal methodology in the analysis of Likert scales. *Quality & Quantity*, *41*(5), 601-626.

Goodman, A., Sahlqvist, S., and Ogilvie, D. 2013. Who uses walking and cycling infrastructure and how? Longitudinal results from the UK iConnect study. *Preventative Medicine*, 57, 518-524.

Green, A.M. 1997, March. Kappa statistics for multiple raters using categorical classifications. In *Proceedings of the 22nd annual SAS User Group International conference*. San Diego. *2*(3), 4-9.

Gwet, K.L., 2008. Intrarater reliability. *Wiley encyclopedia of clinical trials*. 1-13.

Harkey, D., Reinfurt, D. and Knuiman, M. 1998. Development of the bicycle compatibility index. *Transportation Research Record*, 1636, 13-20.

Heeren, T. and D'Agostino, R. 1987. Robustness of the two independent samples t-test when applied to ordinal scaled data. *Statistics in medicine*, *6*(1), 79-90.

Holgado–Tello, F.P., Chacón–Moscoso, S., Barbero–García, I. and Vila–Abad, E. 2010. Polychoric versus Pearson correlations in exploratory and confirmatory factor analysis of ordinal variables. *Quality & Quantity*, *44*(1), 153-166.

Hood, J., Sall, E. and Charlton, B. 2011. A GPS-based bicycle route choice model for San Francisco, California. *Transportation letters*, 3, 63-75.

Huang, Y. and Ye, G. 1995. Selecting bicycle commuting routes using GIS. *Berkeley Planning Journal*, *10*(1), 75-90.

Huchzermeyer, M. 2016. Cycling still an uphill battle. *Mail & Guardian*, 11 March. Available from http://mg.co.za/article/2016-03-10-cycling-is-still-an-uphill-battle/, accessed 5 May 2016.

Hudson, M. 1978. *The bicycle planning book*. Open Books Publishing Limited, London.

Kamargianni, M. 2015. Investigating next generation's cycling ridership to promote sustainable mobility in different types of cities. *Research in Transportation Economics*, 53, 45-55.

Kish, L., 1965. Sampling organizations and groups of unequal sizes. *American sociological review*. 30, 564-572.

Klobucar, M. and Fricker, J. 2007. Network evaluation tool to improve real and perceived bicycle safety. *Transportation Research Record*, 2031, 25-33.

Krenn, P.J., Oja, P. and Titze, S. 2014. Route choices of transport bicyclists: a comparison of actually used and shortest routes. *International Journal of Behavioral Nutrition and Physical Activity,* 11, 31-37.

Larsen, J., Patterson, Z. and El-Geneidy, A. 2013. Build it. But where? The use of geographic information systems in identifying locations for new cycling infrastructure. *International Journal of Sustainable Transportation*, *7*(4), 299-317.

Ledesma, R.D. and Valero-Mora, P. 2007. Determining the number of factors to retain in EFA: An easy-to-use computer program for carrying out parallel analysis. *Practical assessment, research & evaluation*, *12*(2), 1-11.

Li, R., Leung, Y., Lin, H. and Huang, B. 2013. An adaptive compromise programming method for multi-objective path optimization. *Journal of Geographic Systems*, 15, 211-228.

MacNaughton, P., Melly, S., Vallarino, J., Adamkiewicz, G. and Spengler, J.D. 2014. Impact of bicycle route type on exposure to traffic-related air pollution. *Science of the Total Environment*, 490, 37-43.

Mamoulis, N. 2012. Spatial Networks. In: Özsu, M. T. (ed), *Spatial Data Management*. Morgan & Claypool, London, 65-83.

Making Joburg a cycle friendly city. 2014. *Making Joburg a cycle friendly city*, Powerpoint Presentation, City of Joburg, Johannesburg, South Africa.

McClintock, H. 2002. *Planning for Cycling: Principles, Practice, and Solutions for Urban Planners.* Woodhead Publishing Ltd., Cambridge, United Kingdom.

Mohammed, S. 2009. *Pre-feasibility study on non-motorised transport (NMT) in the FIFA WC 2010 host cities.* Final Report, CSIR for Republic of South Africa Department of Environmental Affairs, October 2009.

Möhring, R.H., Schilling, H., Schütz, B., Wagner, D. and Willhalm, T. 2005. Partitioning graphs to speed up Dijkstra's algorithm. In *International Workshop on Experimental and Efficient Algorithms* (pp. 189-202). Springer Berlin Heidelberg.

Norman, G. 2010. Likert scales, levels of measurement and the "laws" of statistics. *Advances in health sciences education*. *15*(5), 625-632.

Perederieieva, O., Ehrgott, M., Wang, J.Y. and Raith, A. 2013. A computational study of traffic assignment algorithms. *Australasian transport research forum, C*(1), 1-18.

Pucher, J., Komanoff, C. and Schimek, P. 1999. Bicycling renaissance in North America? Recent trends and alternative policies to promote bicycling. *Transportation Research Part A*, 33, 625-654.

Rankin, G. and Stokes, M. 1998. Reliability of assessment tools in rehabilitation: an illustration of appropriate statistical analyses. *Clinical rehabilitation*, *12*(3), 187-199.

Raith, A., Wang, J.Y., Ehrgott, M. and Mitchell, S.A. 2014. Solving multi-objective traffic assignment. *Annals of Operations Research*, 222(1), 483-516.

Refaat, M. 2007. Data Preparation for Data Mining Using SAS (Morgan Kaufmann series in data management systems). Morgan Kaufmann Publishers.

Rissel, C., Greaves, S., Wen, L.M., Crane, M. and Standen, C. 2015. Use of and short-term impacts of new cycling infrastructure in inner-Sydney, Australia: a quasi-experimental design. *International Journal of Behavioral Nutrition and Physical Activity*, 12(1), 129.

Sahlqvist, S., Goodman, A., Jones, T., Powell, J., Song, Y. and Ogilvie, D. 2015. Mechanisms underpinning use of new walking and cycling infrastructure in different contexts: mixed-method analysis. *International Journal of Behavioral Nutrition and Physical Activity*, 12(1), 185.

Sainio, J., Westerholm, J. and Oksanen, J. 2015. Generating heat maps of popular routes online from massive mobile sports tracking application data in milliseconds while respecting privacy. *ISPRS International Journal of Geo-Information*, 4, 1813-1826.

Sampat, M.P., Wang, Z., Markey, M.K., Whitman, G.J., Stephens, T.W. and Bovik, A.C. 2006. October. Measuring intra-and inter-observer agreement in identifying and localizing structures in medical images. In *2006 International Conference on Image Processing*. IEEE, 81-84.

Scarf, P. and Grehan, P. 2005. An empirical basis for route choice cycling. *Journal of Sports Sciences*, 23:9, 919-925.

Stones, L. 2013. The Struggle for a cycle friendly city. *Mail & Guardian*, 4 October. Available from http://mg.co.za/article/2013-10-04-00-the-struggle-for-a-cycle-friendly-city, accessed May 11 2016.

Su, J.G., Winters, M., Nunes, M. and Brauer, M. 2010. Designing a route planner to facilitate and promote cycling in Metro Vancouver, Canada. *Transportation Research Part A*, 44, 495-505.

Todes, A. 2012. Urban growth and strategic spatial planning in Johannesburg, South Africa. *Urban Planning in Africa: Cities*, 29(3), 158-165.

Wahlgren, L. and Schantz, P. 2011. Bikeability and methodological issues using the active commuting route environment scale (ACRES) in a metropolitan setting. *BMC medical research methodology*, 11(1), 1-20.

Wang, J.Y.T., Mirza, L., Cheung, A.K.L. and Moradi, S. 2014. Understanding factors influencing choices of cyclists and potential cyclists: A case study at the University of Auckland. *Road and Transport Research: a journal of Australian and New Zealand research and practice*, *23*(4), 37-51.

Wigan, M., Richardson, A. and Brunton, P. 1998. Simplified estimation of demand for nonmotorized trails using geographic information systems. *Transportation Research Record*, 1636, 47-55.

Winters, M., Davidson, G., Kao, D. and Teschke, K. 2010. Motivators and deterrents of bicycling: comparing influences on decisions to ride. *Transportation*, 38, 153-168.

Wong, B.Y.M., Faulkner, G. and Buliung, R. 2011. GIS measured environmental correlates of active school transport: a systematic review of 14 studies. *International Journal of Behavioral Nutrition and Physical Activity*, *8*(39), 1479-5868.

Xiong, G. and Wang, Y. 2014. Best routes selection in multimodal networks using multi-objective genetic algorithm. *Journal of Combinatorial Optimization*, 28, 655-673.

Yiannakoulias, N., Bennet, S.A. and Scott, D.M. 2012. Mapping commuter cycling risk in urban areas. *Accident Analysis and Prevention*, 45, 164-172.

Zolnik, E. and Cromley, E. 2007. Poisson multilevel methodology of bicycle levels of service for road networks. *Transportation Research Record*, 2031, 1-8.

Appendix A – Wits and JUCA Survey Questions and Possible Answers

Demographic Questions

1. What is your gender
    a. Male
    b. Female
    c. Gender Neutral
    d. I'd rather not say
2. How old are you?
3. What is your ethnicity? (From 2011 RSA Census categories)
    a. Black African
    b. White
    c. Coloured
    d. Indian
    e. Asian
    f. Other
    g. I'd rather not say
4. Are you a South African citizen or a foreign national?
5. (If 4 = foreign national) What is your home country?
6. Are you a student or staff member at Wits? *What is your occupation? (JUCA survey version)*
    a. Student
    b. Staff
    c. Other
7. Where do you live? (You can provide a postal code, or just the neighbourhood you live in. Any addresses will be kept private and never be published.)
8. About how far away do live from your place of (work/study)?
    a. 0-1km b. 1-5 km c. 5-10km d. Greater than 10km

Bicycle Questions

1. Do you own or have regular access to a bicycle?
    a. Yes
    b. No
2. (From 1a) How often do you ride a bicycle?
    a) At least once per week
    b) 1-3 times per month
    c) 1-3 times per year
    d) I never ride a bicycle
3. (From 2abc) Why do you ride a bicycle? (Select all that apply)
    a) To commute to work, school, or the library
    b) For a fun leisure activity
    c) To increase my physical activity
    d) I cannot afford other modes of transportation
    e) To save money on fuel costs
    f) Other____________________
4. (From 1b, 2d, 3) What are the main constraints that influence you to **not ride** a bicycle? (Select all that apply)
    a. It is too strenuous due to hilly terrain.
    b. I live too far away from Wits/my place of work to cycle there.
    c. It is unsafe to ride among automobile traffic.
    d. It is unsafe to ride due to crime.
    e. There are no bicycle paths available.
    f. I have a car.
    g. I only ride for recreational purposes.
    h. I use public transportation.

i. The roads I would use aren't wide enough to accommodate bicycles.
j. I don't want to breathe in automobile pollution.
k. There are not enough bike paths and lanes to cycle on.
l. The roads I would use are not connected to public transportation stops.
m. The roads I would use are not connected to shopping areas.
n. The roads I would use are not connected to commercial hubs/office parks.
o. Other____________________

5. The City of Johannesburg is in the process of installing bicycle infrastructure in various neighbourhoods across the city. To encourage you to cycle more often, what factors do you think the City should consider when choosing which roads to install the paths on? (Select all that apply)
    a. Slope and steepness of roads
    b. Automobile traffic volume
    c. Road intersection size and type
    d. Intersection size and type
    e. Frequency of auto traffic Accidents
    f. Coordination with public transportation
    g. Connection to shopping areas
    h. Proximity to parks and green space
    i. Security (Avoiding areas of high crime)
    j. Connections to population centres
    k. Connections to commercial hubs and office parks
    l. Road width and type
    m. Other____________________
6. How important is it to you that bicycle paths avoid roads with steep slopes? (From 5)
    1 – Not Important at all
    2 – Slightly Important
    3 – Somewhat Important
    4 – Important
    5 – Very Important
7. How important is it to you that bicycle paths avoid roads with high automobile traffic volume? (1-5 Likert scale)
8. How important is it to you that bicycle paths avoid roads with high vehicle speeds relative to cycling speeds? (1-5 Likert scale)
9. How important is it to you that bicycle path routes are coordinated with public transportation stops? (1-5 Likert scale)
10. How important is it to you that bicycle path routes are connected to shopping areas? (1-5 Likert scale)
11. How important is it to you that bicycle paths follow parks and green space? (1-5 Likert scale)
12. How important is it to you that bicycle paths avoid areas with high crime rates? (1-5 Likert scale)
13. How important is it to you that bicycle path routes connect to areas with high population centres? (1-5 Likert scale)
14. How important is it to you that bicycle path routes connect to commercial hubs and office parks? (1-5 Likert scale)
15. Are there any other factors in determining bicycle path routes that you think are important? (1-5 Likert scale)

Appendix B – Intra-rater Agreement Tables for Calculating Weighted Kappa

B1 – Observations Matrix for First and Second Survey Responses

OBSERVATIONS

Survey 1	Survey 2: 1	2	3	4	5	Total
1	3	6	4	2	2	17
2	1	19	14	2	1	37
3	1	7	30	9	3	50
4	0	1	10	24	5	40
5	2	3	3	14	14	36
Total	7	36	61	51	25	180

B2 – Weights Matrix for closeness of agreement

WEIGHTS

	1	2	3	4	5
1	0	1	2	3	4
2	1	0	1	2	3
3	2	1	0	1	2
4	3	2	1	0	1
5	4	3	2	1	0

B3 – Expectations Matrix

EXPECTATIONS

Survey 1	Survey 2: 1	2	3	4	5	Total
1	0.66	3.40	5.76	4.82	2.36	17
2	1.44	7.40	12.54	10.48	5.14	37
3	1.94	10.00	16.94	14.17	6.94	50
4	1.56	8.00	13.56	11.33	5.56	40
5	1.40	7.20	12.20	10.20	5.00	36
Total	7	36	61	51	25	180

B4 – Calculated Weighted Cohen's Kappa Excel Code and Output

=1-SUMPRODUCT(Obs. Matrix,Weights Matrix)/SUMPRODUCT(Expectations Matrix,Weights Matrix)

=0.49936

Appendix C – Complete Response Percentages for Motivators/Deterrents for Cycling

C1 – Deterrents for Cycling in Johannesburg

What Constraints influence you to not ride a bicycle (or ride a bicycle less)?							
JUCA							
Too hilly	Too far	Unsafe traffic	Unsafe crime	Use car	Use PT	Road too narrow	
5	4	26	12	8	2	12	*Count*
14.71%	11.76%	76.47%	35.29%	23.53%	5.88%	35.29%	*Ratio*
No paths	pollution	NoPT	No Shopping	No Commercial	Only Rec	Not Enough Time	
19	2	1		1	4	2	*Count*
55.88%	5.88%	2.94%	0.00%	2.94%	11.76%	5.88%	*Ratio*
WITS							
Too hilly	Too far	Unsafe traffic	Unsafe crime	Use car	Use PT	Road too narrow	
99	242	355	262	221	119	141	*Count*
13.88%	33.94%	49.79%	36.75%	31.00%	16.69%	19.78%	*Ratio*
No paths	pollution	No PT	No Shopping	No Commercial	Only Rec	Don't have a bike	
264	76	38	23	39	139	28	*Count*
37.03%	10.66%	5.33%	3.23%	5.47%	19.50%	3.93%	*Ratio*

C2 – Motivators for Cycling in Johannesburg

Why do you ride a bicycle?							
JUCA					Other		
Commute	Physical	Leisure	Affordable	Savefuel	Competition		
15	28	24	0	8	1	*Count*	
45.45%	84.85%	72.73%	0.00%	24.24%	3.03%	*Ratio*	
WITS					Other		
Commute	Physical	Leisure	Affordable	Savefuel	Competition	Climatechange	
34	100	124	4	20	1	1	*Count*
22.22%	65.36%	81.05%	2.61%	13.07%	0.65%	0.65%	*Ratio*

Appendix D – Additional Comparison Test SAS Outputs

D1 – Comparison between Wits and JUCA respondents for the Question: How important do you think it is for bicycle lanes to be connected to public transportation stops?

Table of Group by Importance				
		Importance		
		0	1	Total
Group				
JUCA	Frequency	9	25	34
	Col Pct	11.54	3.76	
Wits	Frequency	69	640	709
	Col Pct	88.46	96.24	
Total	Frequency	78	665	743

Column 1 Risk Estimates						
	Risk	ASE	(Asymptotic) 95% Confidence Limits		(Exact) 95% Confidence Limits	
Row 1	0.2647	0.0757	0.1164	0.4130	0.1288	0.4436
Row 2	0.0973	0.0111	0.0755	0.1191	0.0765	0.1215
Total	0.1050	0.0112	0.0829	0.1270	0.0839	0.1293
Difference	0.1674	0.0765	0.0175	0.3173		
Difference is (Row 1 - Row 2)						

(SAS Institute Inc., 2013)

D2 – Comparison between Wits and JUCA respondents for the Question: How important do you think it is for bicycle lanes to be avoid areas of high crime?

Table of Group by Importance				
		Importance		
		0	1	Total
Group				
JUCA	Frequency	5	29	34
	Col Pct	21.74	4.02	
Wits	Frequency	18	692	710
	Col Pct	78.26	95.98	
Total	Frequency	23	721	744

Column 1 Risk Estimates						
	Risk	ASE	(Asymptotic) 95% Confidence Limits		(Exact) 95% Confidence Limits	
Row 1	0.1471	0.0607	0.0280	0.2661	0.0495	0.3106
Row 2	0.0254	0.0059	0.0138	0.0369	0.0151	0.0398
Total	0.0309	0.0063	0.0185	0.0434	0.0197	0.0460
Difference	0.1217	0.0610	0.0021	0.2413		
Difference is (Row 1 - Row 2)						

(SAS Institute Inc., 2013)

D3 – Comparison between Wits and JUCA respondents for the Question: How important do you think it is for bicycle lanes to be avoid roads with steep slopes?

Table of Group by Importance				
		Importance		
		0	1	Total
Group				
JUCA	Frequency	11	23	34
	Col Pct	9.73	3.65	
Wits	Frequency	102	608	710
	Col Pct	90.27	96.35	
Total	Frequency	113	631	744

Column 1 Risk Estimates						
	Risk	ASE	(Asymptotic) 95% Confidence Limits		(Exact) 95% Confidence Limits	
Row 1	0.3235	0.0802	0.1663	0.4808	0.1739	0.5053
Row 2	0.1437	0.0132	0.1179	0.1695	0.1187	0.1716
Total	0.1519	0.0132	0.1261	0.1777	0.1268	0.1797
Difference	0.1799	0.0813	0.0205	0.3392		
Difference is (Row 1 - Row 2)						

(SAS Institute Inc., 2013)

Appendix E – Principal Component Analysis SAS Outputs and Tables

E1 – Polychoric Coefficients for Wits responses generated by SAS PROC CORR .

Variable Coefficient	Correlated Variable	
Crime	Speed Differential	-0.02696
Crime	Automobile Traffic	-0.05948
Crime	Commercial Hubs	-0.01484
Crime	Population Centres	0.20595
Crime	Public Transportation	0.14218
Crime	Green Space	0.02262
Crime	Shopping Centres	0.05905
Crime	Steep Slopes	0.19114
Speed Differential	Automobile Traffic	0.18199
Speed Differential	Commercial Hubs	0.11557
Speed Differential	Population Centres	-0.01094
Speed Differential	Public Transportation	0.04526
Speed Differential	Green Space	0.01417
Speed Differential	Shopping Centres	0.02729
Speed Differential	Steep Slopes	0.02816
Automobile Traffic	Commercial Hubs	-0.00675
Automobile Traffic	Population Centres	-0.01392
Automobile Traffic	Public Transportation	0.04143
Automobile Traffic	Green Space	0.00681
Automobile Traffic	Shopping Centres	-0.10837
Automobile Traffic	Steep Slopes	0.0036
Commercial Hubs	Population Centres	0.14384
Commercial Hubs	Public Transportation	0.10819
Commercial Hubs	Green Space	0.02984
Commercial Hubs	Shopping Centres	0.15222
Commercial Hubs	Steep Slopes	0.13135
Population Centres	Public Transportation	0.20952
Population Centres	Green Space	-0.00809
Population Centres	Shopping Centres	0.0825
Population Centres	Steep Slopes	0.0689
Public Transportation	Green Space	0.07945
Public Transportation	Shopping Centres	0.19968
Public Transportation	Steep Slopes	0.15475
Green Space	Shopping Centres	0.06612
Green Space	Steep Slopes	-0.04944
Shopping Centres	Steep Slopes	0.10617

E2 – Polychoric Coefficients for JUCA responses generated by SAS PROC CORR.

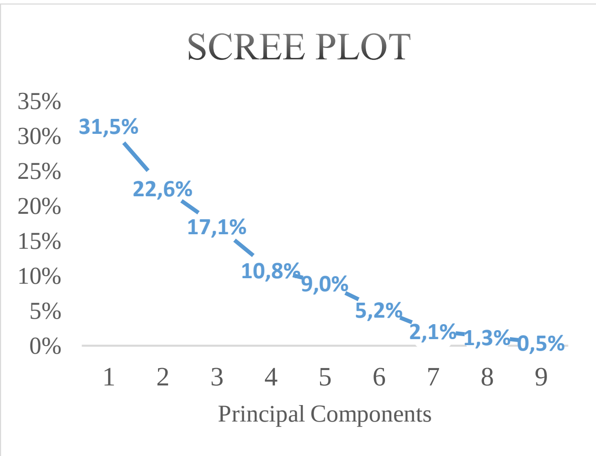
Polychoric Correlations		
Variable	**With Variable**	**Correlation**
Speed_diff	Crime	0.11695
Speed_diff	Population_centres	0.10204
Speed_diff	Auto_traffic	0.59412
Speed_diff	Commercial_hubs	-0.20969
Speed_diff	Green_space	0.15103
Speed_diff	Shopping_centres	0.10113
Speed_diff	Public_trans	-0.06755
Speed_diff	Slopes	0.14307
Crime	Population_centres	-0.03722
Crime	Auto_traffic	0.09397
Crime	Commercial_hubs	0.30547
Crime	Green_space	0.29143
Crime	Shopping_centres	-0.30656
Crime	Public_trans	-0.25125
Crime	Slopes	-0.0648
Population_centres	Auto_traffic	-0.0625
Population_centres	Commercial_hubs	0.64776
Population_centres	Green_space	0.23746
Population_centres	Shopping_centres	0.55816
Population_centres	Public_trans	0.45446
Population_centres	Slopes	-0.05468
Auto_traffic	Commercial_hubs	-0.15024
Auto_traffic	Green_space	0.37679
Auto_traffic	Shopping_centres	-0.03701
Auto_traffic	Public_trans	0.25071
Auto_traffic	Slopes	0.49335
Commercial_hubs	Green_space	-0.00856
Commercial_hubs	Shopping_centres	0.31972
Commercial_hubs	Public_trans	0.38491
Commercial_hubs	Slopes	0.05132
Green_space	Shopping_centres	0.21159
Green_space	Public_trans	0.04886
Green_space	Slopes	0.18767
Shopping_centres	Public_trans	0.66583
Shopping_centres	Slopes	0.45646
Public_trans	Slopes	0.63525

E3 – JUCA Responses Principal Component Analysis Outputs and Scree Plot in Excel

Polychoric Correlation Matrix

	Speed_df	Crime	Pop	Auto_traf	Comm	Green	Shopping	Pub_trans	Slopes
Speed_diff	1.0000	0.1170	0.1020	0.5941	-0.2097	0.1510	0.1011	-0.0676	0.1431
Crime	0.1170	1.0000	0.0372	0.0940	0.3055	0.2914	-0.3066	-0.2513	-0.0648
Population_centres	0.1020	-0.0372	1.0000	-0.0625	0.6478	0.2375	0.5582	0.4545	-0.0547
Auto_traffic	0.5941	0.0940	0.0625	1.0000	-0.1502	0.3768	-0.0370	0.2507	0.4934
Commercial_hubs	-0.2097	0.3055	0.6478	-0.1502	1.0000	-0.0086	0.3197	0.3849	0.0513
Green_space	0.1510	0.2914	0.2375	0.3768	-0.0086	1.0000	0.2116	0.0489	0.1877
Shopping_centres	0.1011	-0.3066	0.5582	-0.0370	0.3197	0.2116	1.0000	0.6658	0.4565
Public_trans	-0.0676	-0.2513	0.4545	0.2507	0.3849	0.0489	0.6658	1.0000	0.6353
Slopes	0.1431	-0.0648	0.0547	0.4934	0.0513	0.1877	0.4565	0.6353	1.0000

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9
Eigenvalues	**2.8314**	**2.0301**	**1.5405**	**0.9755**	**0.8081**	**0.4669**	**0.1914**	**0.1140**	**0.0420**
Speed_diff	0.0992	0.4794	0.1257	-0.5479	0.4046	0.3223	0.0056	0.0507	-0.4174
Crime	-0.0765	0.1250	0.6658	0.3971	0.1477	0.3696	0.4037	-0.0218	0.2360
Population_centres	0.4012	-0.2743	0.3046	-0.4279	0.0241	-0.2436	0.1261	0.5464	0.3370
Auto_traffic	0.1928	0.5933	0.0398	0.0381	0.1915	-0.5167	-0.1192	-0.3162	0.4358
Commercial_hubs	0.3061	-0.3714	0.4023	0.2016	0.3393	-0.1239	-0.5369	-0.2386	-0.3003
Green_space	0.1961	0.2867	0.3793	-0.0166	-0.7794	-0.1075	-0.0853	-0.0152	-0.3296
Shopping_centres	0.4946	-0.1282	-0.1513	-0.2063	-0.2009	0.5357	-0.0652	-0.4626	0.3576
Public_trans	0.5142	-0.0651	-0.2231	0.2160	0.1125	-0.2206	0.6444	-0.1389	-0.3728
Slopes	0.3807	0.2942	-0.2611	0.4769	0.0396	0.2656	-0.3030	0.5550	0.0305



Appendix F – Crime and POI Categories Table

Route Factors and Corresponding POI and Crime Types Used		
Connect to Commercial Hubs	**Connect to Shopping Areas**	**Avoid Areas of High Crime**
Business Facility	ATM	Assault
City Hall	Bank	Attempted Business Robbery
Civic/Community Centre	Bookstore	Attempted Common Robbery
Convention/Exhibition Centre	Clothing Store	Attempted House Robbery
County Council	Coffee Shop	Attempted Robbery: w/ Firearms
Court House	Consumer Electronics Store	Bank Robberies
Embassy	Convenience Store	Business Robbery
Government Office	Department Store	Car Jacking
	Grocery Store	Common Robbery
	Home Improvement & Hardware Store	House Robbery
	Home Specialty Store	Robbery Aggravating
	Office Supply & Services Store	Robbery: Cash in Transit
	Pharmacy	Robbery w/ Weapon (Not a Firearm)
	Post Office	Theft of Motor Vehicle/ Motor Cycle
	Shopping	Truck Hijacking
	Specialty Store	Theft off/from a Motor Vehicle
	Sporting Goods Store	Other